

# Best of LessWrong: January 2022

1. [An Observation of Vavilov Day](#)
2. [Brain Efficiency: Much More than You Wanted to Know](#)
3. [What's Up With Confusingly Pervasive Consequentialism?](#)
4. [Postmortem on DIY Recombinant Covid Vaccine](#)
5. [Look For Principles Which Will Carry Over To The Next Paradigm](#)
6. [More Is Different for AI](#)
7. [Prizes for ELK proposals](#)
8. [Use Normal Predictions](#)
9. [Long covid: probably worth avoiding—some considerations](#)
10. [\[Intro to brain-like-AGI safety.\] 1. What's the problem & Why work on it now?](#)
11. [Future ML Systems Will Be Qualitatively Different](#)
12. [\[New Feature\] Support for Footnotes!](#)
13. [How To Raise Others' Aspirations in 17 Easy Steps](#)
14. [Lives of the Cambridge polymath geniuses](#)
15. [Third Time: a better way to work](#)
16. [Trying to Keep the Garden Well](#)
17. [Implications of Civilizational Inadequacy \(reviewing mazes/simulacra/etc\)](#)
18. [Omicron Post #12](#)
19. [Activated Charcoal for Hangover Prevention: Way more than you wanted to know](#)
20. [A one-question Turing test for GPT-3](#)
21. [Challenges with Breaking into MIRI-Style Research](#)
22. [Covid 1/13/22: Endgame](#)
23. [Truthful LMs as a warm-up for aligned AGI](#)
24. [Covid 1/6/22: The Blip](#)
25. [2020 Review Article](#)
26. [Why do we need a NEW philosophy of progress?](#)
27. [Calibration proverbs](#)
28. [Why rationalists should care \(more\) about free software](#)
29. [Forecasting Newsletter: Looking back at 2021](#)
30. [ELK First Round Contest Winners](#)
31. [Omicron Post #13: Outlook](#)
32. [Lateral Thinking \(AI safety HPMOR fanfic\)](#)
33. [Omicron Post #15](#)
34. [Scott Alexander 2021 Predictions: Market Prices - Resolution](#)
35. [Signaling isn't about signaling, it's about Goodhart](#)
36. [Question Gravity](#)
37. [What happened when I applied metrics to my piano practice \(a five-part essay about the process of learning\)](#)
38. [Omicron Post #14](#)
39. [List of Probability Calibration Exercises](#)
40. [Covid 1/27/22: Let My People Go](#)
41. [How Bad Is QWERTY, Really? A Review of the Literature, such as It Is](#)
42. [ML Systems Will Have Weird Failure Modes](#)
43. [Luna Lovegood and the Fidelius Curse - Part 2](#)
44. [Excessive Nuance and Derailing Conversations](#)
45. [Apply for research internships at ARC!](#)
46. [Nudging My Way Out Of The Intellectual Mosh Pit](#)
47. [Regularization Causes Modularity Causes Generalization](#)
48. [CICO is a weird reference class](#)
49. [Covid 1/20/22: Peak Omicron](#)
50. [Thought Experiments Provide a Third Anchor](#)

# Best of LessWrong: January 2022

1. [An Observation of Vavilov Day](#)
2. [Brain Efficiency: Much More than You Wanted to Know](#)
3. [What's Up With Confusingly Pervasive Consequentialism?](#)
4. [Postmortem on DIY Recombinant Covid Vaccine](#)
5. [Look For Principles Which Will Carry Over To The Next Paradigm](#)
6. [More Is Different for AI](#)
7. [Prizes for ELK proposals](#)
8. [Use Normal Predictions](#)
9. [Long covid: probably worth avoiding—some considerations](#)
10. [\[Intro to brain-like-AGI safety\] 1. What's the problem & Why work on it now?](#)
11. [Future ML Systems Will Be Qualitatively Different](#)
12. [\[New Feature\] Support for Footnotes!](#)
13. [How To Raise Others' Aspirations in 17 Easy Steps](#)
14. [Lives of the Cambridge polymath geniuses](#)
15. [Third Time: a better way to work](#)
16. [Trying to Keep the Garden Well](#)
17. [Implications of Civilizational Inadequacy \(reviewing mazes/simulacra/etc\)](#)
18. [Omicron Post #12](#)
19. [Activated Charcoal for Hangover Prevention: Way more than you wanted to know](#)
20. [A one-question Turing test for GPT-3](#)
21. [Challenges with Breaking into MIRI-Style Research](#)
22. [Covid 1/13/22: Endgame](#)
23. [Truthful LMs as a warm-up for aligned AGI](#)
24. [Covid 1/6/22: The Blip](#)
25. [2020 Review Article](#)
26. [Why do we need a NEW philosophy of progress?](#)
27. [Calibration proverbs](#)
28. [Why rationalists should care \(more\) about free software](#)
29. [Forecasting Newsletter: Looking back at 2021](#)
30. [ELK First Round Contest Winners](#)
31. [Omicron Post #13: Outlook](#)
32. [Lateral Thinking \(AI safety HPMOR fanfic\)](#)
33. [Omicron Post #15](#)
34. [Scott Alexander 2021 Predictions: Market Prices - Resolution](#)
35. [Signaling isn't about signaling, it's about Goodhart](#)
36. [Question Gravity](#)
37. [What happened when I applied metrics to my piano practice \(a five-part essay about the process of learning\)](#)
38. [Omicron Post #14](#)
39. [List of Probability Calibration Exercises](#)
40. [Covid 1/27/22: Let My People Go](#)
41. [How Bad Is QWERTY, Really? A Review of the Literature, such as It Is](#)
42. [ML Systems Will Have Weird Failure Modes](#)
43. [Luna Lovegood and the Fidelius Curse - Part 2](#)
44. [Excessive Nuance and Derailing Conversations](#)
45. [Apply for research internships at ARC!](#)
46. [Nudging My Way Out Of The Intellectual Mosh Pit](#)
47. [Regularization Causes Modularity Causes Generalization](#)
48. [CICO is a weird reference class](#)

49. [Covid 1/20/22: Peak Omicron](#)
50. [Thought Experiments Provide a Third Anchor](#)

# An Observation of Vavilov Day

Content note: this post contains discussion of starvation.

I aspire to be a person who does good things, and who is capable of doing hard things in service of that. This is a plan to test that capacity.

I haven't been in a battle, but if you gave me the choice between dying in battle and slowly starving to death, I would immediately choose battle. Battles are scary but they are short and then they are over.

If you gave me a chance to starve to death to generate some sufficiently good outcome, like saving millions of people from starvation, I think I would do it, and I would be glad to have the opportunity. It would hurt, but only for a few weeks, and in that time I could comfort myself with the warm glow of how good this was for other people.

If you gave me a chance to save millions of people by starving, and then put food in front of me, I don't think I could do it. I would do okay for a few days, maybe a week, but I worry that eventually hunger would incapacitate the part of my brain that allows me to make moral trade-offs at my own expense, and I would wake up to find I'd eaten half the food. I want to think I'd manage it, but if the thought experiment gods didn't let me skip the hard part with more proactive measures, I'm not confident I could.

During the siege of Leningrad, scientists and other staff of the Institute of Plant Study [faced the above choice](#), and to the best of our knowledge, all of them chose hunger. [12 of them died for it](#), the rest merely got close (English language sources list 9 deaths, which is the number of scientists who died in service of the seed bank but not the total number of people). They couldn't kill themselves because they were needed to protect the food from rats and starving citizens. Those survival odds are better than the certain death of my hypothetical, but they didn't have the same certainty of impact either, so I think it balances out.

That's heroism enough, but a fraction of what's present in this story. Those scientists worked at an institute founded by [Nikolai Vavilov](#), a Soviet botanist who has the misfortune to be right on issues inconvenient to Joseph Stalin. Vavilov's (correct) insistence that his theories could feed Russians and those of Stalin's favored scientist couldn't [got him](#) arrested, tortured, and sent to a gulag, where he eventually starved to death.

The seeds Vavilov and his staff protected now cover 80% of the cropland of Russia. Credit for scientific revolutions is hard to apportion, but as I reckon it Vavilov is responsible for, at a minimum, tens of millions people living when they would have starved or never born, and the number could be closer to a billion.

Nikolai Vavilov is my hero.



In honor of Nikolai Vavilov, I'm doing a ~36 hour calorie fast from dinner on 1/25 (the day before Vavilov died in the gulag) to breakfast on 1/27 (the end of the siege of Leningrad). Those of you who know me know this is an *extremely* big deal for me, I do not handle being hungry well, and 36 hours is a long time. This might be one of the hardest things I could do while still being physically possible. Moreover, I'm not going to allow myself to just lie in bed for this: I'm committing to at least one physical activity that day (default is outdoor elliptical, unless it's raining), and attempting to work a normal schedule. I expect this to be very hard. But I need to demonstrate to myself that I can do things that are at least this hard, before I'm called on to do so for something that matters.

If this story strikes a chord with you to the point you also want to observe Valilov + associates' sacrifice, I'd enjoy hearing how. I have enough interest locally (bay area California) that there's likely to be a kick-off dinner + reading the night of the 25th. It would

also be traditional for a fasting holiday to end in a feast, but 1/27 is a Thursday and other people have normal jobs so not yet clear how that's going to shake out.

Thanks to Clara Collier for introducing me to the story of Vavilov and his institute, Anna Tchetchetkine for finding Russian-languages sources for me, and Google translate for being so good I didn't need Anna to translate any further.

# Brain Efficiency: Much More than You Wanted to Know

What if the brain is *highly efficient*? To be more specific, there are several interconnected key measures of efficiency for physical learning machines:

- energy efficiency in ops/J
- spatial efficiency in ops/mm<sup>2</sup> or ops/mm<sup>3</sup>
- speed efficiency in time/delay for key learned tasks
- circuit/compute efficiency in size and steps for key low level algorithmic tasks [\[1\]](#)
- learning/data efficiency in samples/observations/bits required to achieve a level of circuit efficiency, or per unit thereof
- software efficiency in suitability of learned algorithms to important tasks, is not directly addressed in this article [\[2\]](#)

Why should we care? Brain efficiency matters a great deal for AGI timelines and takeoff speeds, as AGI is implicitly/explicitly defined in terms of brain parity. If the brain is about 6 OOM away from the practical physical limits of energy efficiency, then roughly speaking we should expect about 6 OOM of further Moore's Law hardware improvement past the point of brain parity: perhaps two decades of progress at current rates, which could be compressed into a much shorter time period by an intelligence explosion - a **hard takeoff**.

But if the brain is already near said practical physical limits, then merely achieving brain parity in AGI at all will *already* require using up most of the optimizational slack, leaving not much left for a hard takeoff - thus a **slower takeoff**.

In worlds where brains are efficient, AGI is first feasible only near the end of Moore's Law (for non-exotic, reversible computers), whereas in worlds where brains are highly inefficient, AGI's arrival is more decorrelated, but would probably come well before any Moore's Law slowdown.

In worlds where brains are ultra-efficient, AGI *necessarily* becomes neuromorphic or brain-like, as brains are then simply what economically efficient intelligence *looks like* in practice, as constrained by physics. This has important implications for AI-safety: it predicts/postdicts the success of AI approaches based on brain reverse engineering (such as DL) and the failure of non-brain like approaches, it predicts that AGI will consume compute & data in predictable brain like ways, and it suggests that AGI will be far more like human simulations/emulations than you'd otherwise expect and will require training/education/raising vaguely like humans, and thus that neuroscience and psychology are perhaps more useful for AI safety than abstract philosophy and mathematics.

If we live in such a world where brains are highly efficient, those of us interested in creating benevolent AGI should immediately drop everything and learn how brains work.

## Energy

Computation is an organization of energy in the form of ordered state transitions transforming physical information towards some end. Computation requires an isolation

of the computational system and its stored information from the complex noisy external environment. If state bits inside the computational system are unintentionally affected by the external environment, we call those bit errors due to noise, errors which must be prevented by significant noise barriers and or potentially costly error correction techniques.

## Thermodynamics

Information is conserved under physics, so logical erasure of a bit from the computational system entails transferring said bit to the external environment, necessarily creating waste heat. This close connection between physical bit erasure and thermodynamics is expressed by the Landauer Limit<sup>[3]</sup>, which is often quoted as

$$E_b > k_B T \ln 2$$

However the full minimal energy barrier analysis involves both transition times and transition probability, and this minimal simple lower bound only applies at the useless limit of 50% success/error probability or infinite transition time.

The key transition error probability  $\alpha$  is constrained by the bit energy:

$$\alpha = e^{-\frac{E_b}{k_B T}} [4][5]$$

Here's a range of bit energies and corresponding minimal room temp switch error rates (in electronvolts):

- $\alpha = 0.49, E_b = 0.02\text{eV}$
- $\alpha = 0.01, E_b = 0.1\text{eV}$
- $\alpha = 10^{-25}, E_b = 1\text{eV}$

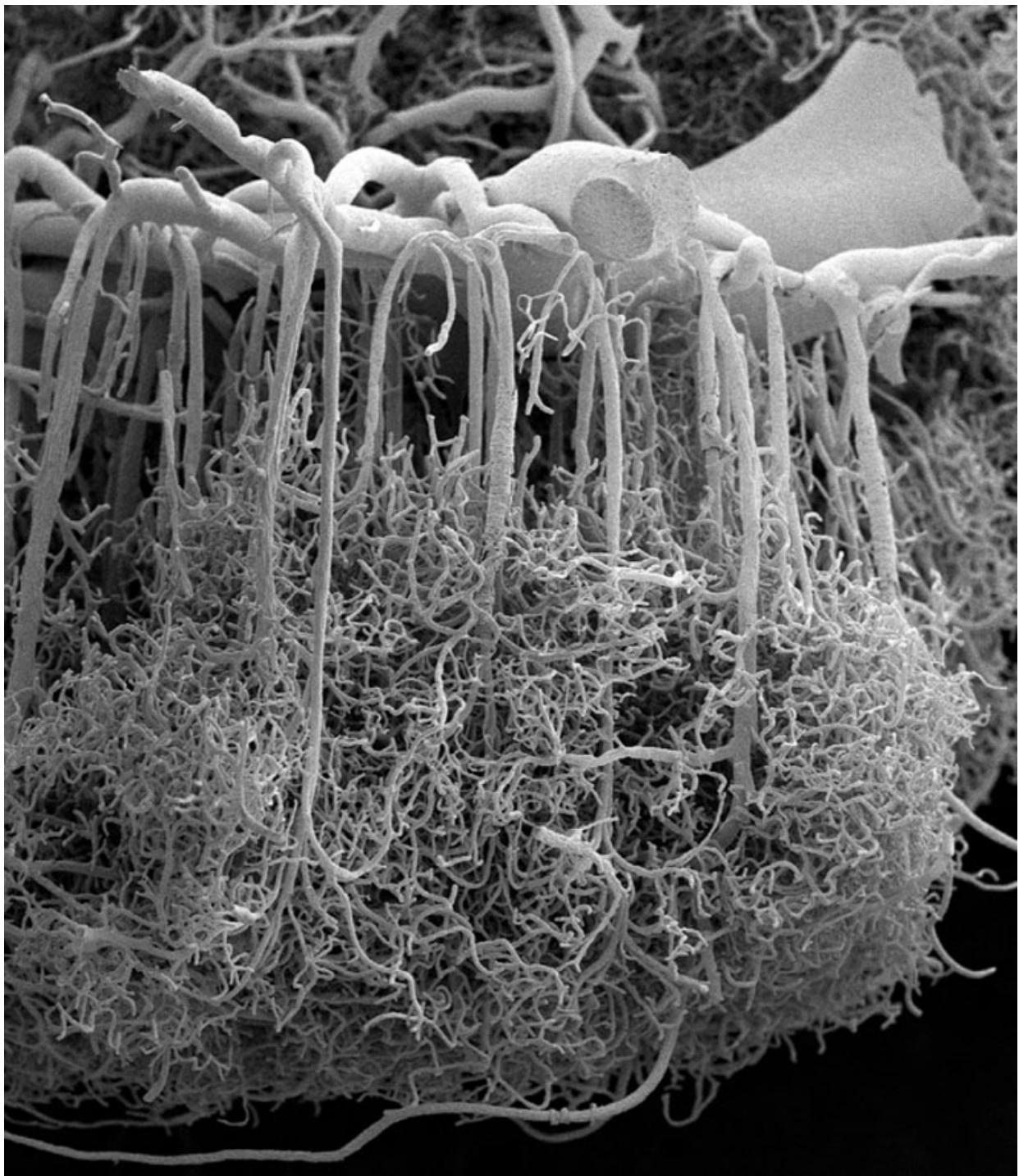
All computers (including brains) are ultimately built out of fundamental indivisible quantal elements in the form of atoms/molecules, each of which is *also* a computational device to which the Landauer Limit applies<sup>[6]</sup>. The combination of this tile/lego decomposition and the thermodynamic bit/energy relationship is a simple but powerful physics model that can predict a wide variety of micro and macro-scale computational thermodynamic measurements. Using this simple model one can predict minimal interconnect wire energy, analog or digital compute energy, and analog or digital device sizes in both brains and electronic computers.

Time and time again while writing this article, the simple first-principles physics model correctly predicted relevant OOM measurements well in advance of finding the known values in literature.

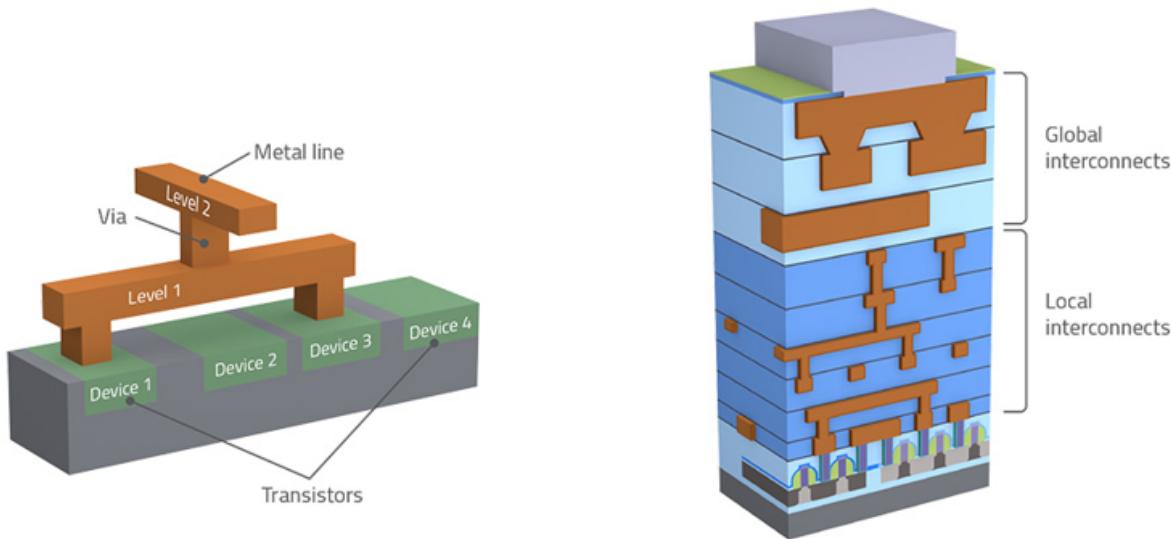
## Interconnect

We can estimate a bound for brain compute energy via interconnect requirements, as interconnect tends to dominate energy costs at high device densities (when devices

approach the size of wire segments). Both brains and current semiconductor chips are built on dissipative/irreversible wire signaling, and are mostly interconnect by volume.



Brains are mostly interconnect.



CPUs/GPUs are mostly interconnect.

A non-superconducting electronic wire (or axon) dissipates energy according to the same Landauer limit per minimal wire element. Thus we can estimate a bound on wire energy based on the minimal assumption of 1 minimal energy unit  $E_b$  per bit per fundamental device tile, where the tile size for computation using electrons is simply the probabilistic radius or De Broglie wavelength of an electron<sup>[7][1]</sup>, which is conveniently  $\sim 1\text{nm}$  for  $1\text{eV}$  electrons, or about  $\sim 3\text{nm}$  for  $0.1\text{eV}$  electrons. Silicon crystal spacing is about  $\sim 0.5\text{nm}$  and molecules are around  $\sim 1\text{nm}$ , all on the same scale.

Thus the **fundamental wire energy** is:  $\sim 1 E_b/\text{bit}/\text{nm}$ , with  $E_b$  in the range of  $0.1\text{eV}$  (low reliability) to  $1\text{eV}$  (high reliability).

The predicted wire energy is  $10^{-19}\text{J}/\text{bit}/\text{nm}$  or  $\sim 100 \text{fJ}/\text{bit}/\text{mm}$  for semi-reliable signaling at  $1\text{V}$  with  $E_b = 1\text{eV}$ , down to  $\sim 10 \text{fJ}/\text{bit}/\text{mm}$  at  $100\text{mV}$  with complex error correction, which is an excellent fit for actual interconnect wire energy<sup>[8][9][10][11]</sup>, which only improves marginally through Moore's Law (mainly through complex sub-threshold signaling and associated additional error correction and decoding logic, again most viable for longer ranges).

For long distance interconnect or communication reversible (ie optical) signaling is obviously vastly superior in asymptotic energy efficiency, but photons and photonics are simply fundamentally too big/bulky/costly due to their  $\sim 1000x$  greater wavelength and thus largely impractical for the dominate on-chip short range interconnects<sup>[12]</sup>. Reversible signaling for electronic wires requires superconductance, which is even more impractical for the foreseeable future.

The brain has an [estimated](#)  $\sim 10^9$  meters of total axon/dendrite wiring length. Using an average wire data rate of  $10 \text{ bit/s}$ <sup>[13][14][15][16]</sup> (although some neurons transmit up to

90 bits/s<sup>[17]</sup>) implies an interconnect energy use of  $\sim 1\text{W}$  for reliable signaling ( $10\text{bit/s} * 10^{18}\text{nm} * 10^{-19}\text{W/bit/nm}$ ), or  $\sim 0.1\text{W}$  for lower bit rates and/or reliability. <sup>[18]</sup>

Estimates of actual brain wire signaling energy are near this range or within an OOM<sup>[19]</sup> <sup>[20]</sup>, so brain interconnect is within an OOM or so of energy efficiency limits for signaling, given its interconnect geometry (efficiency of interconnect geometry itself is a circuit/algorithm level question).

## GPUs

A modern GPU has  $\sim 10^{10}$  transistors, with about half the transistors switching per cycle (CMOS logic is dense) at a rate of  $\sim 10^9 \text{ Hz}$ <sup>[21]</sup>, and so would experience bit logic errors at a rate of about two per month if operating near typical voltages of 1V (for speed) and using theoretically minimal single electron transistors<sup>[22]</sup>. The bit energy  $E_B$  in 2021 GPUs corresponds to on order a few hundred electrons per transistor ( $10^{19}$  transistor switches per second using  $\sim 100$  watts instead of the minimal  $1\text{W}$  for theoretical semi-reliable single electron transistors, as  $1\text{eV} \approx 10^{-19}\text{J}$  ), and thus current GPUs are only about 2 OOM away from thermodynamic limits; which is probably an overestimate, as each hypothetical single-electron transistor needs perhaps 10 single-electron minimal interconnect segments, so GPUs are probably closer to 1 OOM from their practical thermodynamic limits (for any equivalent irreversible device doing all the same logic at the same speed and error rates)<sup>[23]</sup>. Interconnect energy dominates at the highest densities.

The distance to off chip VRAM on a large GPU is  $\sim 3 \text{ cm}$ , so just reading  $10^{15}$  bits to simulate *one cycle* of a brain-size ANN will cost almost  $3\text{kJ}$  ( $1\text{e}\{15\} \text{ bits} * 1\text{e}-19 \text{ J/bit/nm} * 1\text{e}7\text{cm/nm} * 3$ ), so  $300\text{kW}$  to run at  $100\text{Hz}$ . The brain instead only needs to move per neuron values over similar long distances per cycle, which is  $\sim 10,000x$  more efficient than moving around the  $\sim 10,000x$  more numerous connection weights every cycle.

Current GPUs also provide op throughput (for matrix multiplication) up to  $10^{14}$  flops/s or  $10^{15}$  ops/s (for lower bit integer), which is close to current informed estimates for equivalent brain compute ops/s<sup>[24]</sup>. So that alone provides an indirect estimate that brains are within an OOM or two of thermodynamic limits - as current GPUs with equivalent throughput are within 1 to 2 OOM of their limits, and brains use 30x less energy for similar compute throughput ( $\sim 10$  watts vs  $\sim 300$ ).

## Synapses

The adult brain has on  $\sim 2 * 10^{14}$  synapses which perform a synaptic computation on order  $0.5\text{Hz}$ <sup>[25]</sup>. Each synaptic computation is something equivalent to a single analog multiplication op, or a small handful of ops (< 10). Neuron axon signals are binary, but single spikes are known to encode the equivalent of higher dynamic range values through various forms of temporal coding, and spike train pulses can also extend the

range through nonlinear exponential coding - as synapses are known to have the short term non-linear adaptive mechanisms that implement non-linear signal decoding [26] [27]. Thus the brain is likely doing on order  $10^{14}$  to  $10^{15}$  low-medium precision multiply-adds per second.

Analog operations are implemented by a large number of quantal/binary carrier units; with the binary precision equivalent to the signal to noise ratio where the noise follows a binomial distribution. The equivalent bit precision of an analog operation with N quantal carriers is the log of N (maximum signal information) minus the binomial noise entropy:

$$\beta \approx \log_2 N - 0.5 \log_2 (2\pi e N p(1-p))$$

Where  $p$  is the individual carrier switch transition error probability. If the individual carrier transitions are perfectly reliable then the entropy term is zero, but that would require unrealistically high reliability and interconnect energy. In the brain the switch transition error probability will be at least 0.06 for a single electron carrier at minimal useful Landauer Limit voltage of  $\sim 70$ mV like the brain uses (which also happens to simplify the math):

$$\beta \approx \log_2 N - 0.5 \log_2 (2\pi e N 0.06)$$

$$\beta \approx 0.5 \log_2 [28]$$

$$N \approx 2^{2\beta}$$

So true 8-bit equivalent analog multiplication requires about 100k carriers/switches and thus  $10^{-15}$ J/op using noisy subthreshold  $\sim 0.1$ eV per carrier, for a minimal energy consumption on order 0.1W to 1W for the brain's estimated  $10^{14}$  to  $10^{15}$  synaptic ops/s. There is some room for uncertainty here, but not room for many OOM uncertainty. It does suggest that the wiring interconnect and synaptic computation energy costs are of nearly the same OOM. I take this as some evidence favoring the higher  $10^{15}$  op/s number, as computation energy use below that of interconnect requirements is cheap/free.

Note that synapses occupy a full range of sizes and corresponding precisions, with most considerably lower than 8-bit precision (ranging down to 1-bit), which could significantly reduce the *median* minimal energy by multiple OOM, but wouldn't reduce the *mean* nearly as much, as the latter is dominated by the higher precision synapses because energy scales exponentially as  $N \approx 2^{2\beta}$  with precision.

The estimate/assumption of 8-bit equivalence for the higher precision range may seem arbitrary, but I picked that value based on 1.) DL research indicating the need for around 5 to 8 bits per param for effective learning [29][30] (not to be confused with the bits/param for effective forward inference sans-learning, which can be much lower),

and 2.) Direct estimates/measurements of (hippocampal) mean synaptic precisions around 5 bits[\[31\]\[32\]](#). 3.) 8-bit precision happens to be near the threshold where digital multipliers begin to dominate (a minimal digital 8-bit multiplier requires on order  $10^4$  minimal transistors/devices and thus roughly  $10^5$  minimal wire segments connecting them, vs around  $10^5$  carriers for the minimal 8-bit analog multiplier). A synapse is also an all-in-one highly compact computational device, memory store, and learning device capable of numerous possible neurotransmitter specific subcomputations.

The predicted involvement of  $\sim 10^5$  charge carriers then just so happens to match estimates of the mean number of ion carriers crossing the postsynaptic membrane during typical synaptic transmission[\[33\]](#). This is  $\sim 10x$  the number of involved presynaptic neurotransmitter carrier molecules from a few released presynaptic vesicles, but synapses act as repeater amplifiers.

We can also compare the minimal energy prediction of  $10^{-15}$ J/op for 8-bit equivalent analog multiply-add to the known and predicted values for upcoming efficient analog accelerators, which mostly have energy efficiency in the  $10^{-14}$ J/op range[\[34\]\[35\]\[36\]\[37\]](#) for < 8 bit, with the higher reported values around  $10^{-15}$ J/op similar to the brain estimate here, but only for < 4-bit precision[\[38\]](#). Analog devices can *not* be shrunk down to few nm sizes without sacrificing SNR and precision; their minimal size is determined by the need for a large number of carriers on order  $2^{c*\beta}$  for equivalent bit precision  $\beta$ , and  $c \sim 2$ , as discussed earlier.

**Conclusion:** The brain is probably at or within an OOM or so of fundamental thermodynamic/energy efficiency limits given its size, and also within a few OOM of more absolute efficiency limits (regardless of size), which could only be achieved by shrinking it's radius/size in proportion (to reduce wiring length energy costs).

## Space

The brain has about  $10^{14}$  total synapses in a volume of  $\approx 1000$  cm $^3$ , or  $10^{24}$  nm $^3$ , so around  $10^{10}$  nm $^3$  volume / synapse. The brain's roughly 8-bit precision synapses requires on order  $10^5$  electron carriers and thus on same order number of minimal 1 nm $^3$  molecules. Actual synapses are flat disc shaped and only modestly larger than this predicts - with mean surface areas around  $10^5$  nm $^2$ . [\[39\]\[40\]\[41\]](#).

So even if we assume only 10% of synapses are that large, the minimal brain synaptic volume is about  $10^{18}$ nm $^3$ . Earlier we estimated around  $10^{18}$ nm of total wiring length, and thus at least an equivalent or greater total wiring volume (in practice far more due

to the need for thick low resistance wires for fast long distance transmission), but wire volume requirements scale linearly with dimension. So if we ignore all the machinery required for cellular maintenance and cooling, this indicates the brain is at most about 100x larger than strictly necessary (in radius), and more likely only 10x larger.

## Density & Temperature

However, even though the wiring energy scales linearly with radius, the surface area power density which crucially determines temperature scales with the inverse squared radius, and the minimal energy requirements for synaptic computation are radius invariant.

The black body temperature of the brain scales with energy and surface area according to the Stefan-Boltzmann Law:

$$T = (\frac{M_e}{\sigma})^{\frac{1}{4}}$$

Where  $M_e$  is the power per unit surface area in  $\text{W/m}^2$ , and  $\sigma$  is the Stefan-Boltzmann constant. The human brain's output of 10W in  $0.01\text{m}^2$  results in a power density of  $1000\text{W / m}^2$ , very similar to that of the solar flux on the surface of the earth, which would result in an equilibrium temperature of  $\approx 375\text{K}$  or  $100^\circ\text{C}$ , sufficient to boil the blood, if it wasn't actively cooled. Humans have evolved [exceptional heat dissipation](#) capability using the entire skin surface for evaporative cooling [42] : a key adaption that supports both our exceptional long distance running ability, and our oversized brains (3X larger than expected for the default primate body plan, and brain tissue has 10x the power density of the rest of the body).

Shrinking the brain by a factor of 10 at the same power output would result in a  $\sim 3.16\times$  temp increase to around  $1180\text{K}$ , shrinking the brain minimally by a factor of 100 would result in a power density of  $10^7\text{W / m}^2$  and a local temperature of around  $3,750\text{K}$  - similar to that of the surface of the sun.

Current 2021 gpus have a power density approaching  $10^6 \text{ W / m}^2$ , which severely constrains the design to that of a thin 2D surface to allow for massive cooling through large heatsinks and fans. This in turn constrains off-chip memory bandwidth to scale poorly: shrinking feature sizes with Moore's Law by a factor of D increases transistor density by a factor of  $D^2$ , but at best only increases 2d off-chip wire density by a factor of only D, and doesn't directly help reduce wire energy cost at all.

A 2021 GPU with  $10^{10}$  transistors has a surface area of about  $10^{14} \text{ nm}^2$  and so also potentially has room for at most 100x further density scaling, which would result in 10,000x higher transistor count, but given that it only has 1 or 2 OOM potential improvement in thermodynamic energy efficiency, significant further scaling of existing designs would result in untenable power consumption and surface temperature. In practice I expect around only 1 more OOM in dimension scaling (2 OOM in transistor density), with less than an OOM in energy scaling, resulting in dark silicon and or crazy cooling designs [23:1].

**Conclusion:** The brain is perhaps 1 to 2 OOM larger than the physical limits for a computer of equivalent power, but is constrained to its somewhat larger than minimal size due in part to thermodynamic cooling considerations.

## Speed

Brain computation speed is constrained by upper neuron firing rates of around 1 khz and axon propagation velocity of up to 100 m/s [\[43\]](#), which are both about a million times slower than current computer clock rates of near 1 Ghz and wire propagation velocity at roughly half the speed of light. Interestingly, since both the compute frequency and signal velocity scale together at the same rate, computers and brains both are optimized to transmit fastest signals across their radius on the time scale of their equivalent clock frequency: the fastest axon signals can travel about 10 cm per spike timestep in the brain, and also up to on order 10 cm per clock cycle in a computer.

So why is the brain so slow? The answer is again probably energy efficiency.

The maximum frequency of a CMOS device is constrained by the voltage, and scales approximately with [\[44\]](#)[\[45\]](#):

$$f_{MAX} \approx \frac{(V_{dd}-V_t)^2}{V_t}$$

With typical current values in the range of 1.0 for  $V_{dd}$  and perhaps 0.5 for  $V_t$ . The equivalent values for neural circuits are 0.070 for  $V_{dd}$  and around 0.055 for  $V_t$ , which would still support clock frequencies in the MHz range. So a digital computer operating at the extreme subthreshold voltages the brain uses could still switch a thousand times faster.

However, as the minimal total energy usage also scales linearly with switch frequency, and the brain is already operating near thermodynamic efficiency limits at slow speeds, a neuromorphic computer equivalent to the brain, with  $10^{14}$  equivalent synapses (functioning simultaneously as both memory and analog compute elements), would also consume around 10W operating at brain speeds at 1kHz. Scaling a brain to MHz speeds would increase energy and thermal output into the 10kW range and thus surface power density into the  $10^6 W / m^2$  range, similar to current GPUs. Scaling a brain to GHz speeds would increase energy and thermal output into the 10MW range, and surface power density to  $10^9 W / m^2$ , with temperatures well above the surface of the sun.

So in the same brain budget of 10W power and thermodynamic size constraints, one can choose between a computer/circuit with  $10^{14}$  bytes of param memory **and**  $10^{14}$  byte/s of local memory bandwidth but low sub kHz speed, or a system with up to  $10^{14}$  bytes/s of local memory bandwidth **and** gHz speed, but only  $10^8$  bytes of *local* param memory. The most powerful GPUs or accelerators today achieve around  $10^{14}$  bytes/s of

bandwidth from only the register file or lowest level cache, the total size of which tends to be on order  $10^8$  bytes or less.

For any particular energy budget there is a Landauer Limit imposed maximum net communication flow rate through the system and a direct tradeoff between clock speed and accessible memory size at that flow rate.

A single 2021 GPU has the compute power to evaluate a brain sized neural circuit running at low brain speeds, but it has less than 1/1000th of the required RAM. So you then need about 1000 GPUs to fit the neural circuit in RAM, at which point you can then run 1000 copies of the circuit in parallel, but using multiple OOMs more energy per agent/brain for all the required data movement.

It turns out that spreading out the communication flow rate budget over a huge memory store with a slow clock rate is fundamentally more powerful than a fast clock rate over a small memory store. One obvious reason: learning machines have a need to at least store their observational history. A human experiences a sensory input stream at a bitrate of about  $10^6$  bps (assuming maximal near-lossless compression) for about  $10^9$  seconds over typical historical lifespan, for a total of about  $10^{15}$  bits. The brain has about  $2 * 10^{14}$  synapses that store roughly 5 bits each, for about  $10^{15}$  bits of storage. *This is probably not a coincidence.*

In three separate lineages - primates, cetaceans, and proboscideans - brains evolved to large sizes of on order  $10^{11}$  neocortical neurons and  $10^{14}$  synapses (humans: ~20B neocortical neurons, ~80B total, elephants: ~6B neocortical neurons<sup>[46]</sup>, ~250B total, long-finned pilot whale: ~37B neocortical neurons<sup>[47]</sup>, unknown total), concomitant with long (40+) year lifespans. Humans are unique only in having a brain several times larger than normal for our total energy budget, probably due to the unusually high energy payoff for linguistic/cultural intelligence.

**Conclusion:** The brain is a million times slower than digital computers, but its slow speed is probably efficient for its given energy budget, as it allows for a full utilization of an enormous memory capacity and memory bandwidth. As a consequence of being very slow, brains are enormously circuit cycle efficient. Thus even some hypothetical superintelligence, running on non-exotic hardware, will not be able to think much faster than an artificial brain running on equivalent hardware at the same clock rate.

## Circuits

Measuring circuit efficiency - as a complex high level and task dependent metric - is naturally far more challenging than measuring simpler low level physical metrics like energy efficiency. We first can establish a general model of the asymptotic efficiency of three broad categories of computers: serial, parallel, and neuromorphic (processor in memory). Then we can analyze a few example brain circuits that are reasonably well understood, and compare their size and delay to known bounds or rough estimates thereof.

### Serial vs Parallel vs Neuromorphic

A pure serial (Von Neumann architecture) computer is one that executes one simple instruction per clock cycle, fetching opcodes and data from a memory hierarchy. A pure serial computer of size  $d$ , and a clock frequency of  $f$  can execute up to only  $\sim f$  low level instructions per second over a memory of size at most  $\sim d^2$  for a 2d system (as in modern CPUs/GPUs, constrained to 2D by heat dissipation requirements). In the worst case when each instruction accesses a random memory value the processor stalls; the worst case performance is thus bound by  $\sim \min(f, \frac{c}{d})$  where  $d$  is the device size, and  $c \approx 10^8$  m/s is the speed of light bound signal speed. So even a perfectly dense (nanometer scale transistors) 10cm x 10cm pure serial CPU+RAM has performance of only a few billion ops/s when running any algorithms that access memory randomly or perform only few ops per access.

A fully parallel (Von Neumann architecture) computer can execute up to  $d^2$  instructions per clock, and so has a best case performance that scales as  $d^2 * f$  and a worst case of  $\sim d^2 * \min(f, \frac{c}{d})$ . The optimal parallel 10cm x 10cm computational device thus has a maximum potential that is about 16 orders of magnitude greater than the pure serial device.

An optimal neuromorphic computer then simply has a worst and best case performance that is  $d^2 * f$ , for 2d or  $d^3 * f$  for a 3d device like the brain, as its processing units and memory units (synapses) are the same.

Physics is inherently parallel, and thus serial computation simply doesn't scale. The minor big O analysis asymptotic advantages of serial algorithms are completely dominated by the superior asymptotic physical scaling of parallel computation. In other words, big O analysis is *wrong*, as it naively treats computation and memory access as the same thing, when in fact the cost of memory access is *not* constant, and scales up poorly with memory/device size.

The neuromorphic (processor in memory) computational paradigm is asymptotically optimal scaling wise, but within that paradigm we can then further differentiate circuit efficiency in terms of width/size and delay.

## Vision

In terms of circuit depth/delay, humans/primates can perform complex visual recognition and other cognitive tasks in around 100ms to a second, which translates to just a dozen to a hundred inter-module compute steps (each of which takes about 10ms to integrate a few spikes, transmit to the next layer, etc). This naturally indicates learned cortical circuits are near depth optimal, in terms of learning minimal depth circuits for complex tasks, when minimal depth is task useful. As the cortex/cerebellum/BG/thalamus system is a [generic universal learning system](#), showing evidence for efficiency in the single well understood task of vision suffices to show evidence for general efficiency; the 'visual' cortical modules are just generic cortical modules that only happen to learn vision when wired to visual inputs, and will readily learn audio or complex sonar processing with appropriate non-standard input wiring.

A consequence of near-optimal depth/delay implies that the fastest possible thinking minds will necessarily be brain-like, as brains use the near-optimal minimal number of steps to think. So any superintelligence running on any non-exotic computer will not be able to think much *faster* than an artificial brain running on the same equivalent hardware and clock speeds.

In terms of circuit width/size the picture is more complex, but vision circuits are fairly well understood.

The retina not only collects and detects light, it also performs early image filtering/compression with a compact few-layer network. Most vertebrates have a retina network, and although there is considerable variation it is mostly in width, distribution, and a few other hyperparams. The retina performs a reasonably simple well known function (mostly difference of gaussian style filters to exploit low frequency spatio-temporal correlations - the low hanging statistical fruit of natural images), and seems reasonably near-optimal for this function given its stringent energy, area, and latency constraints.

The first layer of vision in the cortex - V1 - is a more massively scaled up early visual layer (esp. in primates/humans), and is also apparently highly efficient given its role to extract useful low-order spatio-temporal correlations for compression and downstream recognition. Extensive experiments in DL on training a variety of visual circuits with similar structural constraints (local receptive field connectivity, etc) on natural image sequences all typically learn V1 like features in first/early layers, such that failure to do so is often an indicator of some error. Some of the first successful learned vision feature extractors were in fact created as a model of V1<sup>[48]</sup>, and modern DL systems with local connectivity still learn similar low level features. As a mathematical theory, sparse coding explains why such features are optimal, as a natural overcomplete/sparse generalization of PCA.

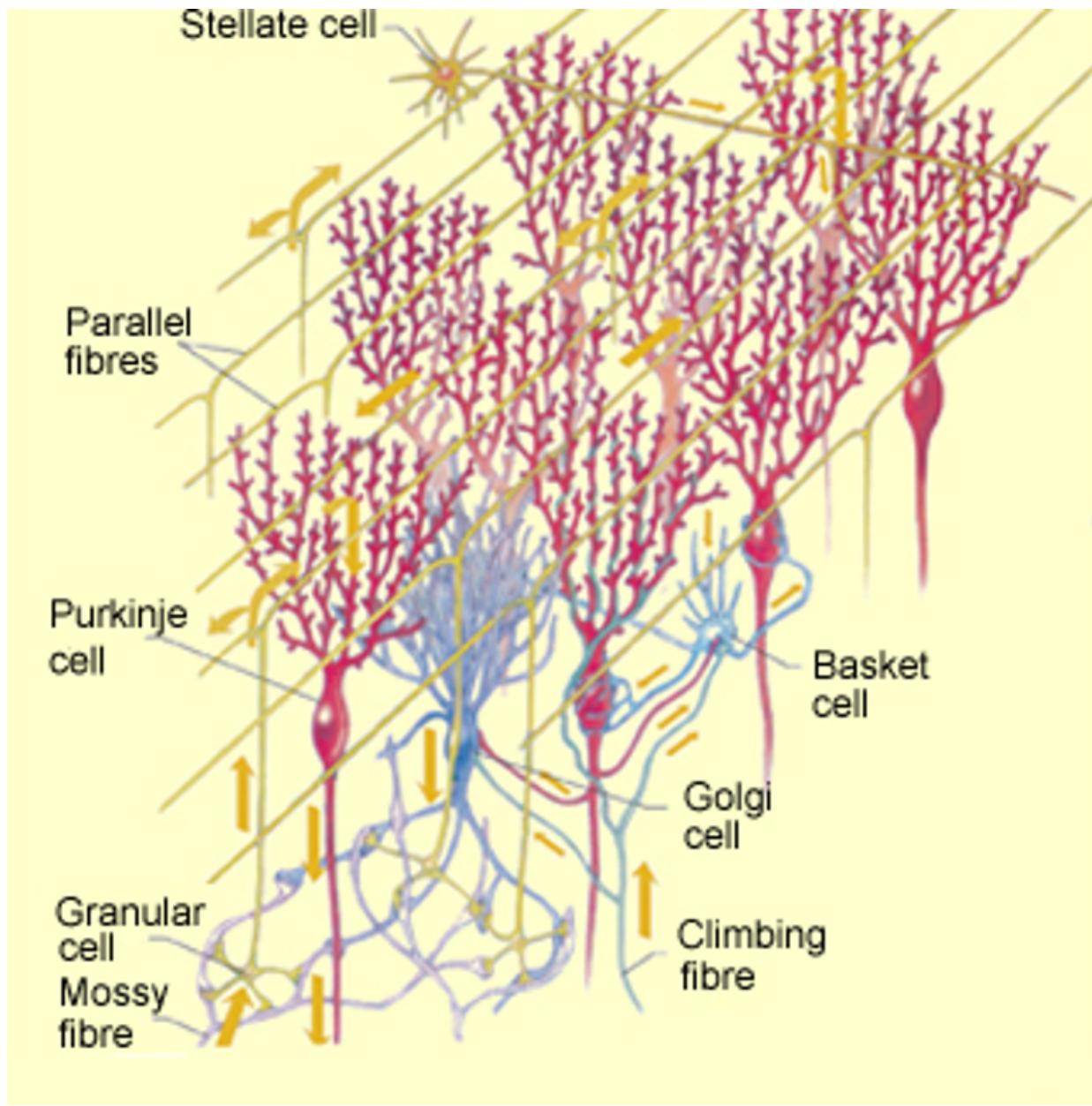
## Vector/Matrix Multiplication

We know that much if not most of the principle computations the brain must perform map to the well studied problem of vector matrix multiplication.

Multiplication of an input vector X and a weight matrix W has a known optimal form in maximally efficient 2D analog circuitry: the crossbar architecture. The input vector X of size M is encoded along a simple uniform vector of wires traversing the structure left to right. The output vector Y of size N is also encoded as another uniform wire vector, but traversing in a perpendicular direction from top to bottom. The weight matrix W is then implemented with analog devices on each of the  $M \times N$  wire crossings.



In one natural extension of this crossbar architecture to 3 dimensions, the input vector X becomes a 2D array of wires of dimension  $M^{0.5} \times M^{0.5}$ , and each output vector Y becomes a flat planar structure (reduction tree), with a potential connection to every input wire. This 3D structure then has a depth of order N, for the N output summation planes. This particular structure is optimal for  $M \sim N^2$ , with other variations optimal for  $M \sim N$ . This is a simplified description of the geometric structure of the cerebellum:



## Deep Learning

Deep learning systems trained with brain-like architectural/functional constraints (recurrence<sup>[49][50]</sup>, local sparse connectivity, etc) on naturalistic data<sup>[51]</sup> with generic multi-task and or self-supervised objectives are in fact our very best models of relevant brain circuits<sup>[52][53][54]</sup>; developing many otherwise seemingly brain-specific features such as two specialized processing streams<sup>[55][56]</sup>, categorical specialization<sup>[57]</sup>, etc., and can explain brain limitations<sup>[58][59]</sup>. Likewise, DL evolving towards AGI converges on brain reverse engineering<sup>[60][61]</sup>, especially when optimizing towards maximal energy efficiency for complex real world tasks.

The spectacular success of brain reverse engineering aka DL - and its complete dominance in modern AI - is strong evidence for brain circuit efficiency, as both

biological and technological evolution, although very different processes, both converge on similar solutions given the same constraints.

**Conclusion:** It's difficult to make strong definitive statements about circuit efficiency, but current evidence is most compatible with high brain circuit efficiency, and I'm not aware of any significant evidence against.

## Data

Data efficiency is a common (although perhaps unfounded) critique of DL. Part of this disadvantage could simply be due to economics: large scale DL systems can take advantage of huge datasets, so there is little immediate practical need to focus on learning from limited datasets. But in the longer term as we approach AGI, learning quickly from limited data becomes increasingly important: it is much of what we mean when we say a human is *smart* or *quick* or *intelligent*.

We can analyze data/learning efficiency on two levels: asymptotic learning efficiency, and practical larger-scale system level data efficiency.

## Asymptotic

In terms of known algorithmic learning theory, a data-optimal learning machine with memory  $O(M)$  can store/evaluate up to  $M$  unique models in parallel per circuit timestep, and can prune about half of said virtual models per observational bit per timestep - as in well known Solomonoff Induction, full Bayesian Inference, or prediction through expert selection<sup>[62]</sup>. The memory freed can then be recycled to evaluate new models the next timestep, so at the limit such a machine can evaluate  $O(M*T)$  models in  $T$  timesteps. Thus any practical learning machine can evaluate at most  $O(N)$  models and same order data observations, where  $N$  is the net compute expended for training (nearly all virtual models are discarded at an average evaluation cost of only  $O(C)$ ). Assuming that 'winning' predictive models are distributed uniformly over model-space, this implies a power law relationship between predictive entropy (log predictive error), and the entropy of model space explored (and thus log compute for training). Deep learning systems are already in this power-law regime<sup>[63][64]</sup>, thus so is the brain, and they are both already in the optimal *broad* asymptotic complexity class.

In terms of tighter bounds on practical large scale data efficiency, we do not have direct apples-to-apples comparisons as humans and current DL systems are trained on different datasets. But some DL systems are trained on datasets that could be considered a relevant subset of the human training dataset.

## Vision

DL vision systems can achieve mildly superhuman performance on specific image recognition games like Imagenet, but these systems are trained on a large labeled dataset of 1M images, whereas humans are first pretrained unsupervised on a larger mostly unlabeled dataset of perhaps 1B images (1 image/s for 32 years), with a tiny fraction of linguistically labeled images (or perhaps none for very specific dog breed categories).

If you look at [Imagenet labels](#), they range from the obvious: syringe, to the obscure: gyromitra. Average [untrained human performance](#) of around 75% top-5 is reasonably

impressive considering that untrained humans have 0 labels for many categories. Trained humans can achieve up to 95% top-5 accuracy, comparable to DL SOTA from 2017. Now 2021 DL SOTA is around 99% top-5 using all labels, and self-supervised SOTA (using a big model) matches human expert ability using 10% of labels (about 100 labels per category), [65] but using multiple data passes. Assuming a human expert takes a second or two to evaluate an image, a single training pass on 10% of the imagenet labels would take about 40 hours: a full time work week, perhaps a month for multiple passes. It's unclear at this point if humans could approach the higher 99% score if only some were willing to put in months or years of training, but it seems plausible.

DL visual systems take advantage of spatial (ie convolutional) weight sharing to reduce/compress parameters and speed up learning. This is difficult/impossible for slow neuromorphic processors like the brain, so this handicap makes brain data efficiency somewhat less directly comparable and somewhat more impressive.

## GPT-N

OpenAI's GPT-3 is a 175B param model (or  $1e12$  bits at 5.75 bits/param) trained on a corpus of about 400B BPE tokens, or roughly 100B words (or  $1e12$  bits at 10 bits/word), whereas older humans are 'trained' on perhaps 10B words (about 5 per second for 64 years), or more generally about 10B timesteps of about 200ms each, corresponding roughly to one saccadic image, one word, precept, etc. A single saccadic image has around 1M pixels compressible to about 0.1bpp, suggesting a human experiences on order  $1e15$  bits per lifetime, on par with up to  $1e15$  bits of synaptic information ( $2e14$  synapses \* 5 bit/synapse).

[Scaling analysis of GPT-N](#) suggests high benchmark performance (vague human parity) will require scaling up to a brain size model a bit above  $1e14$  params and a similar size dataset. This is interesting because it suggests that current DL models (or at least transformers), are perhaps as parameter efficient as the brain, but are far less data efficient in terms of consumed words/tokens. This may not be surprising if we consider that difficulty of the grounding problem: GPT is trying to learn the meaning of language without first learning the grounding of these symbols in a sensorimotor model of the world.

These scaling laws indicate GPT-N would require about 3 to 4 OOM more word data than humans to match human performance, but GPT-3 already trains on a large chunk of the internet. However most of this data is highly redundant. Humans don't train by reading paragraphs drawn uniformly at random from the entire internet - as the vast majority of such data is near worthless. GPT-N models could be made more data efficient through brain inspired active learning (using a smaller net to predict gradient magnitudes to select informative text to train the larger model), and then multi-modal curriculum training for symbol grounding, more like the human education/training process.

## AlphaX

AlphaGo achieved human champion performance after training on about 40 million positions, equivalent to about 400k games, which is roughly an OOM more games than a human professional will play during lifetime training (4k games/year \* 10 years) [66].

AlphaZero [matched human champion performance](#) after training on only about 4 million positions (~100k updates of 4k positions each) and thus 40k games - matching my estimated human data efficiency.

However AlphaX models learn their action-value prediction functions from each MCT state evaluation, just as human brains probably learn the equivalent from imaginative planning state evaluations. But human brains - being far slower - perform at least one OOM less imagined state evaluation rollouts per board move evaluation than AlphaX models, which implies the brain is learning more per imagined state evaluation. The same naturally applies to DeepMind's newer EfficientZero - which learns human-level Atari in only 2 hours realtime<sup>[67]</sup> but this corresponds to a huge number of imagined internal state evaluations, on same order as similar model-free Atari agents.

Another way of looking at it: if AlphaX models really were fully as data efficient as the human brain in terms of learning speed per evaluation step and equivalent clock cycle, then we'd expect them to achieve human level play a million times faster than the typical human 10 years: ie in about 5 minutes (vs ~2 hours for EfficientZero, or ~30 hours for AlphaZero). Some component of this is obviously inefficiency in GPU clock cycles per evaluation step, but to counter that AlphaX models are tiny and often trained in parallel on many GPUs/TPUs.

**Conclusion:** SOTA DL systems have arguably matched the brain's data learning efficiency in the domain of vision - albeit with some artificial advantages like weight-sharing countering potential brain advantages. DL RL systems have also arguably matched brain data efficiency in games such as Go, but only in terms of physical move evaluations; there still appears to be a non-trivial learning gap where the brain learns much more per virtual move evaluation, which DL systems compensate for by rapidly evaluating far more virtual moves during MCTS rollouts. There is still a significant data efficiency gap in natural language, but training datasets are very different and almost certainly favor the brain (multimodal curriculum training and active learning).

Thus there is no evidence here of brain learning inefficiency (for systems of similar size/power). Instead DL still probably has more to learn from the brain on how to learn efficiently beyond SGD, and the probable convergence of biological and technological evolution to what appears to be the same fundamental data efficiency scaling laws is evidence for brain efficiency.

## Conclusions

The brain is about as efficient as any conventional learning machine<sup>[68]</sup> can be given:

1. An energy budget of 10W
2. A thermodynamic cooling constrained surface power density similar to that of earth's surface ( $1\text{ kW/m}^2$ ), and thus a 10cm radius.
3. A total training dataset of about 10 billion precepts or 'steps'

If we only knew the remaining secrets of the brain today, we could train a brain-sized model consisting of a small population of about 1000 agents/sims, running on about as many GPUs, in probably about a month or less, for about \$1M. This would require only about 1kW per agent or less, and so if the world really desired it, we could support a population of billions of such agents without dramatically increasing total world power production.

Nvidia - the single company producing most of the relevant flops today - produced roughly 5e21 flops of GPU compute in 2021, or the equivalent of about 5 million brains [69], perhaps surpassing the compute of the 3.6 million humans born in the US. With 200% growth in net flops output per year from all sources it will take about a decade for net GPU compute to exceed net world brain compute. [70]

Eventually advances in software and neuromorphic computing should reduce the energy requirement down to brain levels of 10W or so, allowing for up to a trillion brain-scale agents at near future world power supply, with at least a concomitant 100x increase in GDP[71]. All of this without any exotic computing.

Achieving those levels of energy efficiency will probably require brain-like neuromorphic-ish hardware, circuits, and learned software via training/education. The future of AGI is to become more like the brain, not less.

---

1. Here we focus on ecologically important tasks like visual inference - how efficient are brain circuits for evolutionarily important tasks?. For more recent *economically* important tasks such as multiplying large numbers the case for brain circuit *inefficiency* is quite strong (although there are some potential exceptions - human mentants such as Von Neumann). [←](#)
2. Obviously the brain's software (the mind) is still rapidly evolving with cultural/technological evolution. The efficiency of learned algorithms (as complex multi-step programs) that humans use to discover new theories of physics, create new DL algorithms, think more rationally about investing, or the said theories or algorithms themselves, are not considered here. [←](#)
3. Landauer, Rolf. "Irreversibility and heat generation in the computing process." IBM journal of research and development 5.3 (1961): 183-191. [gs-link](#) [←](#)
4. Zhirnov, Victor V., et al. "Limits to binary logic switch scaling-a gedanken model." Proceedings of the IEEE 91.11 (2003): 1934-1939. [gs-link](#) [←](#)
5. Frank, Michael P. "Approaching the Physical Limits of Computing." [gs-link](#) [←](#)
6. The tile/lego model comes from Cavin/Zhirnov et al in "Science and engineering beyond Moore's law"<sup>[7]</sup> and related publications. [←](#)
7. Cavin, Ralph K., Paolo Lugli, and Victor V. Zhirnov. "Science and engineering beyond Moore's law." Proceedings of the IEEE 100.Special Centennial Issue (2012): 1720-1749. [gs-link](#) [←](#) [←](#) [←](#)
8. Postman, Jacob, and Patrick Chiang. "A survey addressing on-chip interconnect: Energy and reliability considerations." International Scholarly Research Notices 2012 (2012). [gs-link](#) [←](#)
9. Das, Subhasis, Tor M. Aamodt, and William J. Dally. "SLIP: reducing wire energy in the memory hierarchy." Proceedings of the 42nd Annual International Symposium on Computer Architecture. 2015. [gs-link](#) [←](#)
10. Zhang, Hang, et al. "Architecting energy-efficient STT-RAM based register file on GPGPUs via delta compression." 2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC). IEEE, 2016. [linkgs-link](#) [←](#)

11. Park, Sunghyun, et al. "40.4 fJ/bit/mm low-swing on-chip signaling with self-resetting logic repeaters embedded within a mesh NoC in 45nm SOI CMOS." 2013 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2013. [gs-link ↵](#)
12. As a recent example, TeraPHY offers apparently SOTA electrical to optical interconnect with power efficiency of 5pJ/bit, which surpasses irreversible wire energy of ~100fJ/bit/mm only at just beyond GPU die-size distances of 5cm, and would only just match SOTA electrical interconnect for communication over a full [cerebras wafer-scale device.](#) ↵
13. Reich, Daniel S., et al. "Interspike intervals, receptive fields, and information encoding in primary visual cortex." *Journal of Neuroscience* 20.5 (2000): 1964-1974. [gs-link ↵](#)
14. Singh, Chandan, and William B. Levy. "A consensus layer V pyramidal neuron can sustain interpulse-interval coding." *PloS one* 12.7 (2017): e0180839. [gs-link ↵](#)
15. Individual spikes carry more information at lower spike rates (longer interspike intervals), making sparse low spike rates especially energy efficient, but high total bandwidth, low signal latency, and high area efficiency all require higher spike rates. ↵
16. Koch, Kristin, et al. "How much the eye tells the brain." *Current Biology* 16.14 (2006): 1428-1434. [gs-link ↵](#)
17. Strong, Steven P., et al. "Entropy and information in neural spike trains." *Physical review letters* 80.1 (1998): 197. [gs-link ↵](#)
18. There are more complex physical tradeoffs between wire diameter, signal speed, and energy, such that minimally energy efficient signalling is probably too costly in other constrained dimensions. ↵
19. Lennie, Peter. "The cost of cortical computation." *Current biology* 13.6 (2003): 493-497. [gs-link ↵](#)
20. Ralph Merkle [estimated](#) the energy per 'Ranvier op' - per spike energy along the distance of 1mm jumps between nodes of Ranvier - at  $5 \times 10^{-15} \text{ J}$ , which at  $5 \times 10^{-21} \text{ J/nm}$  is only ~2x the Landauer Limit, corresponding to single electron devices per nm operating at around 40 mV. He also estimates an average connection distance of 1mm and uses that to directly estimate about 1 synaptic op per 1mm 'Ranvier op', and thus about  $10^{15}$  ops/s, based on this energy constraint. ↵
21. Wikipedia, [RTX 3090 stats](#) ↵
22. The minimal Landauer bit error rate for 1eV switches is 1e-25, vs 1e10 transistors at 1e9 hz for 1e6 seconds (2 weeks). ↵
23. Cavin et al estimate end of Moore's Law CMOS device characteristics from a detailed model of known physical limits[\[7.2\]](#). A GPU at these limits could have 10x feature scaling vs 2021 and 100x transistor density, but only about 3x greater

energy efficiency, so a GPU of this era could have 3 trillion transistors, but would use/burn an unrealistic 10kW to run all those transistors at GHz speed. ↵ ↵

24. Carlsmith at Open Philanthropy produced a huge report [resulting](#) in a wide distribution over brain compute power, with a median/mode around  $10^{15}$  ops/s. Although the median/mode is reasonable, this report includes too many poorly informed estimates, resulting in an unnecessarily high variance distribution. The simpler estimate of  $2 * 10^{14}$  synapses switching at around ~0.5hz, with 1 synaptic op equivalent to at least one but up to ten low precision flops or analog multiply-adds, should result in most mass concentrated around  $10^{14}$  op/s and  $10^{15}$  ops/s. There is little uncertainty in the synapse count, not much in the average synaptic firing rate, and the evidence from neuroscience provides fairly strong support, but ultimately the Landauer Limit as analyzed here rules out much more than  $10^{15}$  ops/s, and Carlsmith's report ignores interconnect energy and is confused about the actual practical thermodynamic limits of analog computation. ↵

25. Mean of [Neuron firing rates in humans](#) ↵

26. In some synapses synaptic facilitation acts very much like an exponential decoder, where the spike train sequence 11 has a postsynaptic potential that is 3x greater than the sequence 10, the sequence 111 is 9x greater than 100, etc. - see the reference below. ↵
27. Jackman, Skyler L., and Wade G. Regehr. "The mechanisms and functions of synaptic facilitation." *Neuron* 94.3 (2017): 447-464. [gs-link](#) ↵
28. See the following article for a completely different approach resulting in the same SNR relationship following 3.16 in Sarpeshkar, Rahul. "Analog versus digital: extrapolating from electronics to neurobiology." *Neural computation* 10.7 (1998): 1601-1638. [gs-link](#) ↵
29. Miyashita, Daisuke, Edward H. Lee, and Boris Murmann. "Convolutional neural networks using logarithmic data representation." arXiv preprint arXiv:1603.01025 (2016). [gs-link](#) ↵
30. Wang, Naigang, et al. "Training deep neural networks with 8-bit floating point numbers." *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018. [gs-link](#) ↵
31. Bartol Jr, Thomas M., et al. "Nanoconnectomic upper bound on the variability of synaptic plasticity." *Elife* 4 (2015): e10778. [gs-link](#) ↵
32. Bartol, Thomas M., et al. "Hippocampal spine head sizes are highly precise." *bioRxiv* (2015): 016329. [gs-link](#) ↵
33. Attwell, David, and Simon B. Laughlin. "An energy budget for signaling in the grey matter of the brain." *Journal of Cerebral Blood Flow & Metabolism* 21.10 (2001): 1133-1145. [gs-link](#) ↵

34. Bavandpour, Mohammad, et al. "Mixed-Signal Neuromorphic Processors: Quo Vadis?" 2019 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S). IEEE, 2019. [gs-link ↵](#)
35. Chen, Jia, et al. "Multiply accumulate operations in memristor crossbar arrays for analog computing." *Journal of Semiconductors* 42.1 (2021): 013104. [gs-link ↵](#)
36. Li, Huihan, et al. "Memristive crossbar arrays for storage and computing applications." *Advanced Intelligent Systems* 3.9 (2021): 2100017. [gs-link ↵](#)
37. Li, Can, et al. "Analogue signal and image processing with large memristor crossbars." *Nature electronics* 1.1 (2018): 52-59. [gs-link ↵](#)
38. Mahmoodi, M. Reza, and Dmitri Strukov. "Breaking POps/J barrier with analog multiplier circuits based on nonvolatile memories." *Proceedings of the International Symposium on Low Power Electronics and Design*. 2018. [gs-link ↵](#)
39. Montero-Crespo, Marta, et al. "Three-dimensional synaptic organization of the human hippocampal CA1 field." *Elife* 9 (2020): e57013. [gs-link ↵](#)
40. Santuy, Andrea, et al. "Study of the size and shape of synapses in the juvenile rat somatosensory cortex with 3D electron microscopy." *Eneuro* 5.1 (2018). [gs-link ↵](#)
41. [How big is a synapse? ↵](#)
42. Brengelmann, George L. "Specialized brain cooling in humans?." *The FASEB Journal* 7.12 (1993): 1148-1153. [gs-link ↵](#)
43. Wikipedia: [Nerve Conduction Velocity ↵](#)
44. ScienceDirect: [Dynamic power dissipation](#), EQ Ov.10 [↵](#)
45. Gonzalez, Ricardo, Benjamin M. Gordon, and Mark A. Horowitz. "Supply and threshold voltage scaling for low power CMOS." *IEEE Journal of Solid-State Circuits* 32.8 (1997): 1210-1216. [gs-link ↵](#)
46. Herculano-Houzel, Suzana, et al. "The elephant brain in numbers." *Frontiers in neuroanatomy* 8 (2014): 46. [gs-link ↵](#)
47. Mortensen, Heidi S., et al. "Quantitative relationships in delphinid neocortex." *Frontiers in Neuroanatomy* 8 (2014): 132. [gs-link ↵](#)
48. Olshausen, Bruno A., and David J. Field. "Sparse coding with an overcomplete basis set: A strategy employed by V1?." *Vision research* 37.23 (1997): 3311-3325. [gs-link ↵](#)
49. Kar, Kohitij, et al. "Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior." *Nature neuroscience* 22.6 (2019): 974-983. [gs-link ↵](#)
50. Nayebi, Aran, et al. "Task-driven convolutional recurrent models of the visual system." *arXiv preprint arXiv:1807.00053* (2018). [gs-link ↵](#)
51. Mehrer, Johannes, et al. "An ecologically motivated image dataset for deep learning yields better models of human vision." *Proceedings of the National Academy of Sciences* 118.8 (2021). [gs-link ↵](#)

52. Yamins, Daniel LK, and James J. DiCarlo. "Using goal-driven deep learning models to understand sensory cortex." *Nature neuroscience* 19.3 (2016): 356-365. [gs-link](#) ↵
53. Zhang, Richard, et al. "The unreasonable effectiveness of deep features as a perceptual metric." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. [gs-link](#) ↵
54. Cichy, Radoslaw M., and Daniel Kaiser. "Deep neural networks as scientific models." *Trends in cognitive sciences* 23.4 (2019): 305-317. [gs-link](#) ↵
55. Bakhtiari, Shahab, et al. "The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning." (2021). [gs-link](#) ↵
56. Mineault, Patrick, et al. "Your head is there to move you around: Goal-driven models of the primate dorsal pathway." *Advances in Neural Information Processing Systems* 34 (2021). [gs-link](#) ↵
57. Dobs, Katharina, et al. "Brain-like functional specialization emerges spontaneously in deep neural networks." *bioRxiv* (2021). [gs-link](#) ↵
58. Elsayed, Gamaleldin F., et al. "Adversarial examples that fool both computer vision and time-limited humans." *arXiv preprint arXiv:1802.08195* (2018). [gs-link](#) ↵
59. Nicholson, David A., and Astrid A. Prinz. "Deep Neural Network Models of Object Recognition Exhibit Human-like Limitations When Performing Visual Search Tasks." *bioRxiv* (2021): 2020-10. [gs-link](#) ↵
60. Hassabis, Demis, et al. "Neuroscience-inspired artificial intelligence." *Neuron* 95.2 (2017): 245-258. [gs-link](#) ↵
61. Zador, Anthony M. "A critique of pure learning and what artificial neural networks can learn from animal brains." *Nature communications* 10.1 (2019): 1-7. [gs-link](#) ↵
62. Haussler, David, Jyrki Kivinen, and Manfred K. Warmuth. "Tight worst-case loss bounds for predicting with expert advice." *European Conference on Computational Learning Theory*. Springer, Berlin, Heidelberg, 1995. [gs-link](#) ↵
63. Hestness, Joel, et al. "Deep learning scaling is predictable, empirically." *arXiv preprint arXiv:1712.00409* (2017). [gs-link](#) ↵
64. Rosenfeld, Jonathan S., et al. "A constructive prediction of the generalization error across scales." *arXiv preprint arXiv:1909.12673* (2019). [gs-link](#) ↵
65. Chen, Ting, et al. "Big self-supervised models are strong semi-supervised learners." *arXiv preprint arXiv:2006.10029* (2020). [gs-link](#) ↵
66. Silver, David, et al. "[Mastering the game of Go with deep neural networks and tree search](#)." *nature* 529.7587 (2016): 484-489. ↵
67. Ye, Weirui, et al. "[Mastering atari games with limited data](#)." *Advances in Neural Information Processing Systems* 34 (2021). ↵

68. Practical here implies irreversible - obviously an exotic reversible or quantum computer could potentially do much better in terms of energy efficiency, but all evidence suggests brain size exotic computers are still far in the future, after the arrival of AGI on conventional computers. [←](#)
69. Nvidia's [2021 revenue](#) is about \$25B, about half of which is from consumer GPUs which provide near brain level  $10^{15}$  ops/s for around \$2,000. The other half of revenue for data-center GPUs is around 5x more expensive per flop. [←](#)
70. Without any further progress in flops/s/\$ from Moore's Law, this would entail Nvidia's revenue exceeding United States GDP in a decade. More realistically, even if Nvidia retains a dominant lead, it seems much more likely to arrive from an even split: 30x increase in revenue, 30x increase in flops/s/\$. But as this article indicates, there is limited further slack in Moore's Law, so some amount of growth must come from economic scaling up the fraction of GDP going into compute. [←](#)
71. Obviously neuromorphic AGI or sims/uploads will have numerous transformative advantages over humans: ability to copy/fork entire minds, share modules, dynamically expand modules beyond human brain limits, run at variable speeds far beyond 100hz, interface more directly with computational systems, etc. [←](#)

# What's Up With Confusingly Pervasive Consequentialism?

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

*Fictionalized/Paraphrased version of a real dialog between me and John Wentworth.*

**Fictionalized Me:** So, in the Eliezer/Richard dialogs, Eliezer is trying to get across this idea that consequentialism deeply permeates optimization, and this is important, and that's one<sup>[1]</sup> reason why Alignment is Hard. But something about it is confusing and slippery, and he keeps trying to explain it and it keeps not-quite-landing.

I think I get it, but I'm not sure I could explain it. Or, I'm not sure who to explain it to. I don't think I could tell who was making a mistake, where "consequentialism is secretly everywhere" is a useful concept for realizing-the-mistake.

**Fictionalized John:** [stares at me]

**Me:** Okay, I guess I'm probably supposed to try and explain this and see what happens.

...

**Me:** Okay, so the part that's confusing here is that this is supposed to be something that Eliezer thinks thoughtful, attentive people like Richard (and Paul?) aren't getting, despite them having read lots of relevant material and paying attention and being generally on board with "alignment is hard."

...so, what is a sort of mistake I could imagine a smart, thoughtful person who read the sequences making here?

My Eliezer-model imagines someone building what they think is an aligned ML system. They've trained it carefully to do things they reflectively approve of, they've put a lot of work into making it interpretable and honest. This Smart Thoughtful Researcher has read the sequences and believes that alignment is hard and whatnot. Nonetheless, they'll have failed to really grok this "consequentialism-is-more-pervasive-and-important-than-you-think" concept. And this will cause doom when they try to scale up their project to accomplish something actually hard.

I... guess what I think Eliezer thinks is that Thoughtful Researcher isn't respecting inner optimizers enough. They'll have built their system to be carefully aligned, but to do anything hard, it'll end up generating inner-optimizers that aren't aligned, and the inner-optimizers will kill everyone.

...

**John:** Nod. But not quite. I think you're still missing something.

You're familiar with the arguments of convergent instrumental goals?

**Me:** i.e. most agents will end up wanting power/resources/self-preservation/etc?

**John:** Yeah.

But not only is "wanting power and self preservation" convergently instrumental. *Consequentialism* is convergently instrumental. Consequentialism is a (relatively) simple, effective process for accomplishing goals, so things that efficiently optimize for goals tend to approximate it.

Now, say there's something hard you want to do, like build a moon base, or cure cancer or whatever. If there were a list of all the possible plans that cure cancer, ranked by "likely to work", most of the plans that might work route through "consequentialism", and "acquire resources."

Not only that, most of the plans route through "acquire resources in a way that is unfriendly to human values." Because in the space of all possible plans, while *consequentialism* doesn't take that many bits to specify, [human values are highly complex](#) and take a lot of bits to specify.

Notice that I just said "in the space of all possible plans, here are the most common plans." I *didn't* say anything about agents choosing plans or acting in the world. Just listing the plans. And this is important because the hard part *lives in the choosing of the plans*.

Now, say you build an oracle AI. You've done all the things to try and make it interpretable and honest and such. If you ask it for a plan to cure cancer, what happens?

**Me:** I guess it gives you a plan, and... the plan probably routes through consequentialist agents acquiring power in an unfriendly way.

Okay, but if I imagine a researcher who is thoughtful but a bit too optimistic, what they might counterargue with is: "Sure, but I'll just inspect the plans for whether they're unfriendly, and not do those plans."

And what I might then counterargue their counterargument with is:

1) Are you sure you can actually tell which plans are unfriendly and which are not?

and,

2) If you're reading very carefully, and paying lots of attention to each plan... you'll still have to read through a *lot* of plans before you get to one that's actually good.

**John:** Bingo. I think a lot of people imagine asking an oracle to generate 100 plans, and they think that maybe half the plans will be pretty reasonable. But, the space of plans is huge. Exponentially huge. Most plans just don't work. Most plans that work route through consequentialist optimizers who convergently seek power because you need power to do stuff. But then the space of consequentialist power-seeking plans are *still* exponentially huge, and most ways of seeking power are unfriendly to human values. The hard part is *locating a good plan that cures cancer that isn't hostile to human values in the first place*.

**Me:** And it's not obvious to me whether this problem gets better or worse if you've tried to train the oracle to only output "reasonable seeming plans", since that might

output plans that are deceptively unaligned.

**John:** Do you understand why I brought up this plan/oracle example, when you originally were talking about inner optimizers?

**Me:** Hmm. Um, kinda. I guess it's important that there was a second example.

**John:** ...and?

**Me:** Okay, so partly you're pointing out that hardness of the problem isn't just about getting the AI to do what I want, it's that doing what I want is actually just really hard. Or rather, the part where alignment is hard is precisely when the thing I'm trying to accomplish is hard. Because then I need a powerful plan, and it's hard to specify a search for powerful plans that don't kill everyone.

**John:** Yeah. One mistake I think people end up making here is that they think the problem lives in the AI-who's-deciding/doing things, as opposed to in the actual raw difficulty of the search.

**Me:** Gotcha. And it's important that this comes up in at least two places – inner optimizers with an agenty AI, and an oracle that just output plans that would work. And the fact that it shows up in two fairly different places, one of which I hadn't thought of just now, is suggestive that it could show up in even more places I haven't thought of at all.

And this is confusing enough that it wasn't initially obvious to Richard Ngo, who's thought a ton about alignment. Which bodes ill for the majority of alignment researchers who probably are less on-the-ball.

1. ^

I'm tempted to say "the main reason" why Alignment Is Hard, but then remembered Eliezer specifically reminded everyone not to summarize him as saying things like "the key reason for X" when he didn't actually say that, and often is tailoring his arguments to a particular confusion with his interlocuter.

# Postmortem on DIY Recombinant Covid Vaccine

*edit: Changed title from "Postmortem on RatVac" for clarity.*

**Note:** We named the vaccine candidate "RatVac" as a tongue-in-cheek abbreviation for "Rationalist Vaccine". We have no association with the RaDVAC project and use a much simpler, almost trivial approach. I don't endorse making your own vaccine or taking anything nerdy people on the Internet send you. This is not medical advice.

**tldr:** In April of 2021 I assembled my own subunit vaccine against the Beta variant of SARS-CoV-2 ("SARS2"). Despite starting with the prior that this formulation should be somewhat effective, I could not demonstrate efficacy against the Alpha variant in antibody tests.

## Introduction

I started getting interested in DIY vaccines soon after the [Making Vaccine](#) and [We got what's needed for COVID-19 vaccination completely wrong](#) posts were published. Particularly the idea of just making a standard subunit vaccine appealed to me, so I teamed up with some interested people from the rationalist diaspora and we ended up making our own vaccine candidate in early to mid 2021. Some contributed advice, many contributed funding (for which I am still extremely grateful), and while the project wasn't the clear success I had hoped for, maybe the true treasures are the friends we gained and the lab equipment we bought along the way.

## Recombinant Vaccine ELI5

Subunit vaccines contain a subset of a single protein molecule. For viruses this will typically be a protein that contains the receptor-binding-domain (RBD), i.e. the part that actually binds with the host during infection.

Through the wonders of genetic engineering - see very short introduction below - we can produce these proteins and then simply introduce them into the human body.

While [Yang et al](#) have shown that an immune response occurs even with a vaccine purely containing viral RBD, the response can be enhanced by adding an immunologic adjuvant. Adjuvants are a very diverse class of chemicals, ranging from simple mineral salts to proteins or even small cells. The only thing they ostensibly have in common is that they're used to increase or modulate immune response, and that's why we're using them.

## Genetic Engineering ELI5

For those curious about how we can manufacture proteins to begin with - because I definitely was - here is a very short and abridged intro. The process differs depending on the method used, but usually roughly follows these steps:

- Encode and synthesize the amino acid sequence of the protein as DNA.
- Add this DNA to an *expression vector*. These are typically either viruses or circular strings of DNA called *plasmids*.
- Introduce your expression vector into an *expression host*, often bacterial, yeast or mammalian cells. This is called *transfection* and can take several forms, from chemical and viral transfection to just shooting cells with a "gene gun" that fires microscopic gold particles coated in DNA.
- Select for the transfected cells. This is usually done by adding a gene for antibiotic resistance to the expression vector, so that you can then kill all non-transfected cells by said antibiotic.
- Grow and maintain your culture until sufficient amounts of protein have been produced, then extract and purify the protein.

The details will vary depending on protocol and I'm deliberately sticking to the basics and leaving out things like CRISPR, RNA plasmids and plant hosts.

More important is imo that while this may sound extraordinarily complicated, it's very much doable by a single individual or small team in a self-funded lab. DNA can be synthesized for hundreds of dollars and often already shipped as a customized plasmid. Host cells and growth media are also affordable (at least at the mg- $\mu$ g scale we're aiming for, don't try making your own burger meat), which leaves purification. I'm not too happy with any of the DIY methods I've seen so far, but a jury-rigged column chromatography setup is very much possible for \$1000, with \$7k-15k for something I'd personally want to use.

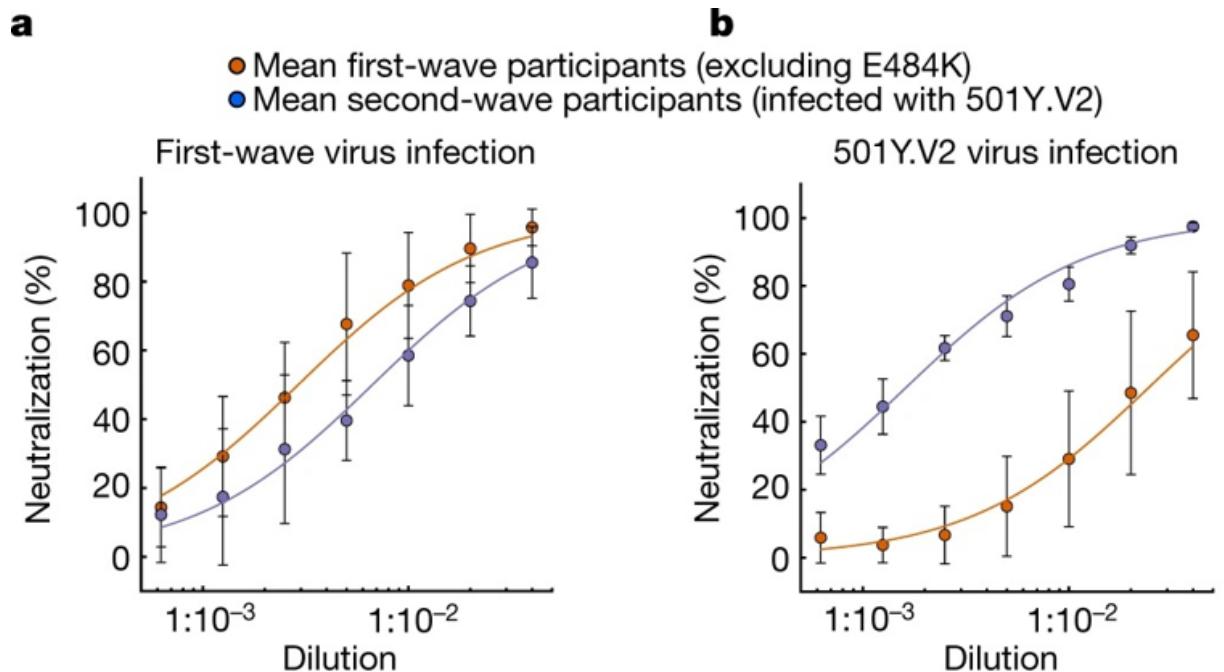
## RatVac in theory

With the data dump out of the way, let's return to the vaccine project.

The first question we tried to answer was whether it even made sense to target the SARS2 RBD as a vaccine candidate. Similar approaches were suggested for the first SARS and showed some promise. Luckily, this seemed true for SARS2 as well: [Yang et al](#) found in animal studies (on mice *and* primates) that a recombinant vaccine not only led to the production of antibodies, but also neutralized live SARS2 virus. This was promising.

The second question concerned which variant to target. Alpha appeared to be suitable for a vaccine, but the Beta variant was on the rise and inoculation against Alpha didn't necessarily imply inoculation against Beta. So, what about the inverse scenario? Would people vaccinated against Beta still get Alpha at proportional rates?

The diagram below from a [Cele et al. study](#) paints an optimistic picture: Beta antibodies provide far greater cross-variant immunity than Alpha antibodies. The sample size ( $n=19$ ) is a bit small, but given this information, SARS2 Beta makes the better vaccine target.



**c**



Virus variant	Plasma source		Fold change (HM/HT)
	First wave	501Y.V2	
First wave	344.0 (275.4–458.0)	149.7 (132.1–172.8)	2.3
	41.1 (32.7–55.5)	619.7 (517.8–771.5)	15.1
Fold change (HM/HT)	8.4	4.1	

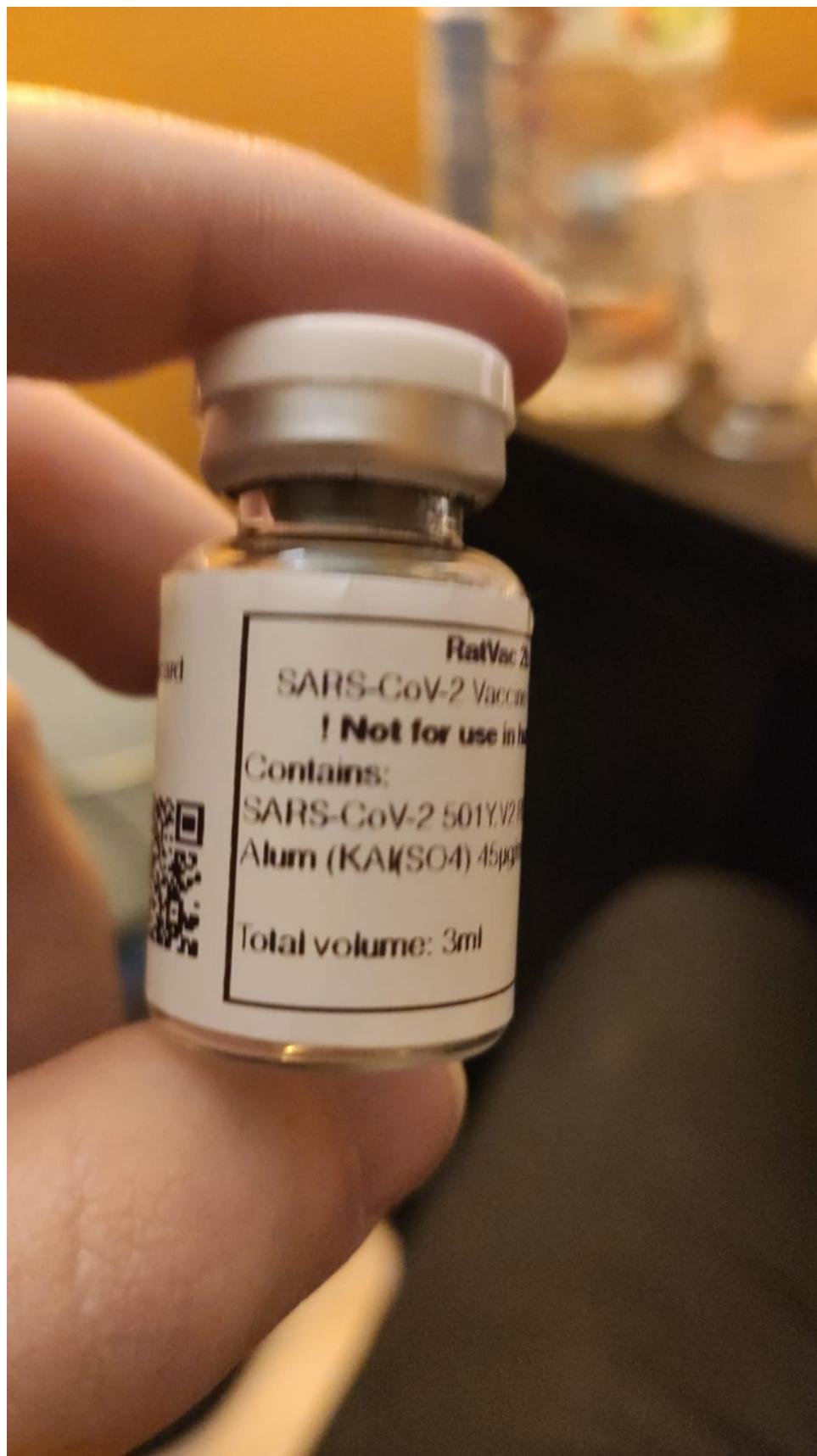
Lastly, we had to pick an adjuvant. Aluminum salt ("alum") was the easiest option, since it's easy to acquire and easy to sterilize. It's not ideal, since it's biased towards a T-helper 2 response and we'd prefer a balanced response, but that's not a critical concern as long as the magnitude of the immune response is at a satisfactory level.

With these concerns at least somewhat abated, we arrived at a vaccine candidate containing 50 $\mu$ g of Beta Variant RBD, for three doses of a bit over 15 $\mu$ g each.

## RatVac in practice

Unfortunately, Germany has some of the strictest GMO restrictions and violations can carry fines of up to €50k. Since I don't have a licensed lab at my disposal, I decided to outsource the protein expression and just import the protein from a large pharmaceutical company. Since it's not toxic, this isn't particularly difficult. At least for the adjuvant I managed to avoid pharma prices by buying pure aluminum salts.

The protein was purified in a column chromatograph, all glassware, the adjuvant and the vials were washed in alcohol twice to remove any macroscopic contaminants and then sterilized in a pressure cooker. Assembly of the vaccine happened in a clean room that I managed to "borrow" from a friendly biologist. While I don't think we actually operated at ISO 5 / WHO Grade A, working this way was clearly preferable to working in a home lab (and risking giving myself a pollen allergy, given the season).



Each dose was given two weeks apart, which, in hindsight, was almost certainly too short of an interval to maximize the immune response.

Successful response to the vaccine would result in antibodies against the viral RBD, so I bought and used at-home blood antibody tests, which returned negative. Since I wasn't sure if the antibody tests did in fact test for antibodies against RBD rather than other proteins, I also booked an official antibody test, which couldn't detect antibodies in my blood, either.

Unfortunately I was struggling with over a year of Covid isolation at that point, so I decided to just take my employer's offer of a speedy vaccination in June, rather than running more tests on myself.

## Conclusion

In conclusion, meh. I couldn't run as many tests as I would've liked and my data is inconclusive.

Of course, this isn't a condemnation of subunit vaccines. Considering how easy it is to produce this type of vaccine, I do think that the current system for vaccine development is somewhat broken. While having pharmaceutical companies responsible for funding their own clinical trials removes the incentive to drain public funds by testing as many drug candidates as humanly possible, it also introduces a) an incentive for fraud and b) a necessity for the resulting drug to be either comprehensively protected by IP laws or so difficult to produce that the risk of legal copycats is negligible. Imo this is highly problematic for several reasons, mainly because we, as a society, would *want* the kind of drug that relies on well-established technology and is manufactured very, very easily, so that production can be up- and downscaled at will. Subunit (and to some extent DNA) vaccines check both of these boxes and I'm somewhat disappointed to see no real open-source vaccine receive the kind of government funding that would have been necessary. It might still have failed, of course, but compared to all the costs incurred by Covid, the expense of running several Phase II and a few Phase III studies would have been trivial.

# Look For Principles Which Will Carry Over To The Next Paradigm

In 1918, Emmy Noether published her famous theorem showing that each symmetry of the laws of physics implies a corresponding conserved quantity. Laws which remain the same even if we move the whole universe left or right a little result in conservation of momentum, laws which remain the same over time result in conservation of energy, and so forth.

At the time, Noether's Theorem was only proven for the sorts of systems used in classical physics - i.e. a bunch of differential equations derived by minimizing an "action". Over the next few decades, the foundational paradigm shifted from classical to quantum, and Noether's original proof did not carry over. But the *principle* - the idea that symmetries imply conserved quantities - *did* carry over. Indeed, the principle is arguably simpler and more elegant in quantum mechanics than in classical.

This is the sort of thing I look for in my day-to-day research: principles which are simple enough, fundamental enough, and general enough that they're likely to carry over to the next paradigm. I don't know what the next paradigm will be, yet; the particulars of a proof or formulation of a problem might end up obsolete. But I look for principles which I expect will survive, even if the foundations shift beneath them.

## Examples In My Own Work

My own day-to-day research focuses on modelling abstraction.

I generally build these models on a framework of probability, information theory, and causal models. I *know* that this framework will not cover all of abstraction - for example, it doesn't cover mathematical abstractions like "addition" or "linearity". Those abstractions are built into the structure of logic, and probability theory takes all of logic as given. There may be some way in which the abstraction of linearity lets me answer some broad class of questions more easily, but standard probability and information theory ignore all that by just assuming that all pure-logic questions are answered for free.

... yet I continue to use this probability/information/causality framework, rather than throwing it away and looking for something more general on which to build the theory. Why? Well, I expect that this framework is *general enough* to figure out principles which will carry over to the next paradigm. I can use this framework to talk about things like "throwing away information while still accurately answering queries" or "information relevant far away" or "massively redundant information", I can show that various notions of "abstraction" end up equivalent, I can mathematically derive the surprising facts implied by various assumptions. For instance, I can prove the [Telephone Theorem](#): when transmitted over a sufficiently long distance, all information is either completely lost or arbitrarily perfectly conserved. I expect a version of that principle to carry over to whatever future paradigm comes along, even after the underlying formulations of "information" and "distance" change.

## Why Not Just Jump To The Next Paradigm?

One obvious alternative to looking for such principles is to instead focus on the places where my current foundational framework falls short, and try to find the next foundational framework upfront. Jump right to the next paradigm, as quickly as possible.

The main reason not to do that is that I don't think I have enough information yet to figure out what the next paradigm is.

Noether's Theorem and principles like it played a causal role in figuring out quantum mechanics. It was the simple, general principles of classical mechanics which provided constraints on our search for quantum mechanical laws. Without those guideposts, the search space of possible physical laws would have been too wide.

Special relativity provides a particularly clear example here. Nobody would have figured it out without the principles of electrodynamics and Lorentz transformations to guide the way. Indeed, Einstein's contribution was "just" to put an interpretation on math which was basically there already.

More generally, knowing a few places where the current framework fails is not enough to tell us what the next framework should be. I know that my current foundation for thinking about abstraction is too narrow, but the search space of possible replacements is still too wide. I want simple general principles, principles which capture the relevant parts which I *do* think I understand, in order to guide that search. So, in my day-to-day I use the framework I have - but I look for the sort of principles which I expect to generalize to the next framework, and which can guide the search for that next framework.

This leaves a question: how do we know when it's time to make the jump to the next paradigm? As a rough model, we're trying to figure out the constraints which govern the world. Sometimes, the rate-limiting step might be figuring out new constraints, to limit our search. Sometimes, the rate-limiting step might be abandoning (probably implicit) *wrong* constraints already in our models, like the assumption of Galilean relativity implicitly built into pre-special-relativity physics. When finding new constraints is the rate-limiting step, it should feel like exploring a wide-open space, like we're looking around and noticing patterns and finding simple ways to describe those patterns. When abandoning wrong constraints is the rate-limiting step, it should feel like the space is *too* constrained, like different principles or examples come into conflict with each other.

## Why Not Just Forget About The Next Paradigm?

On the other end of the spectrum, some people argue for just working within the current paradigm and forgetting about the next one. This is a long-term/short-term tradeoff: in the short term, the current paradigm is usually the best we have; building new frameworks takes time. So if our goals are short term - like, say, a startup which needs to show growth in the next six months - then maybe we should just do what we can with what we have.

There are definitely lots of places where this is the right move. On the other hand, I think "long term" is often much, much shorter than people realize.

I worked in startups for about five years. Usually, the companies I was at needed to show results or shut down within ~2 years. On the other hand, the code we wrote usually turned over within a year - the company would pivot or the UI design would change or the code architecture and tech stack would shift, and old code would either be deprecated or rewritten. In that environment, “building for the next paradigm” meant figuring out principles which would leave us better off a year from now, when the current code had mostly turned over. For instance, knowledge about our users (often from A/B tests), typically had lasting value. Sometimes, a smart library design would last. With a runway of ~2 years and a turnover time of ~1 year, the right move is to usually spend that first year on things which will make us better off a year from now after everything has turned over.

... not that we always did that, mind you, but it was the things which lasted through turnover which were consistently the most important in hindsight. And after five years of this, one can see the patterns in what kinds of things will last.

AI research (and alignment research) in particular is a place where the “long term” is much, much shorter than many people realize. Not in the sense that AGI is right around the corner, but in the sense that the next paradigm is less than 5 years away, not more than 20. Just within the past 10 years, we saw the initial deep learning boom with image classifiers, then a shift to image generators (with the associated shift to GAN architectures), and then the shift to transformers and language models. Even if you think that transformer-based language models are the most probable path to AGI, there will [still likely be major qualitative shifts along the way](#). If we’re doing work which is narrowly adapted to the current paradigm, it’s likely to be thrown out, and probably not even very far in the future.

The work done by Chris Olah’s team is a good example here. They did some really cool work on [generative image nets](#). Then the shift to transformers came along, and they recently restarted from roughly square zero on [transformer nets](#). Presumably *some* illegible skills transferred, but they mostly seem to be figuring things out from scratch, as far as I can tell. When the next shift comes, I expect they’ll be back at roughly square zero again. My advice to someone like Chris Olah would be: figure out the principles which seem likely to generalize. At a bare minimum, look for principles or tools which are useful for both image and text models, both CNNs and transformers. Those are the principles which are likely to still be relevant in 5 years.

## How Can We Recognize Principles Which Will Carry Over?

As an 80/20 solution, I think it’s usually fine to trust your instincts on this one. The important step is just to *actually ask yourself* whether something will carry over. I can look at my own work and say “hmm, this specific notion of ‘redundant information’ probably won’t carry over, but some general notion of ‘abstractions summarize massively redundant information’ probably will, and the principles I’ve derived from this model probably will”. Similarly, I expect someone in 1920 could look at Noether’s Theorem and think “wow, even if the foundations of physics are totally overturned, I bet some version of this principle will survive”.

If you want a more legible answer than that, then my advice is to introspect on what information is driving your intuitions about what will or will not carry over. I intend to

do that going forward, and will hopefully figure out some patterns. For now, simplicity and generality seem like the main factors.

# More Is Different for AI

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

Machine learning is touching increasingly many aspects of our society, and its effect will only continue to grow. Given this, I and many others care about risks from future ML systems and how to mitigate them.

When thinking about safety risks from ML, there are two common approaches, which I'll call the **Engineering** approach and the **Philosophy** approach:

- The Engineering approach tends to be empirically-driven, drawing experience from existing or past ML systems and looking at issues that either: (1) are already major problems, or (2) are minor problems, but can be expected to get worse in the future. Engineering tends to be bottom-up and tends to be both in touch with and anchored on current state-of-the-art systems.
- The Philosophy approach tends to think more about the limit of very advanced systems. It is willing to entertain thought experiments that would be implausible with current state-of-the-art systems (such as Nick Bostrom's [paperclip maximizer](#)) and is open to considering abstractions without knowing many details. It often sounds more "sci-fi like" and more like philosophy than like computer science. It draws some inspiration from current ML systems, but often only in broad strokes.

I'll discuss these approaches mainly in the context of [ML safety](#), but the same distinction applies in other areas. For instance, an Engineering approach to AI + Law might focus on [how to regulate self-driving cars](#), while Philosophy might ask whether [using AI in judicial decision-making could undermine liberal democracy](#).

While Engineering and Philosophy agree on some things, for the most part they make wildly different predictions both about what the key safety risks from ML will be and how we should address them:

- Both Engineering and Philosophy would agree on some high-level points: they would agree that [misaligned objectives](#) are an important problem with ML systems that is likely to get worse. Engineering believes this because of examples like the Facebook recommender system, while Philosophy believes this based on conceptual arguments like those in [Superintelligence](#). Philosophy is more confident that misaligned objectives are a big problem and thinks they could pose an existential threat to humanity if not addressed.
- Engineering and Philosophy would both agree that out-of-distribution robustness is an important issue. However, Philosophy might view most engineering-robustness problems (such as those faced by self-driving cars) as temporary issues that will get fixed once we train on more data. Philosophy is more worried about whether systems can generalize from settings where humans can provide data, to settings where they cannot provide data even in principle.
- Engineering tends to focus on tasks where current ML systems don't work well, weighted by their impact and representativeness. Philosophy focuses on tasks that have a certain abstract property that seems important, such as [imitative deception](#).

In my experience, people who strongly subscribe to the Engineering worldview tend to think of Philosophy as fundamentally confused and ungrounded, while those who strongly subscribe to Philosophy think of most Engineering work as misguided and orthogonal (at best) to the long-term safety of ML. Given this sharp contrast and the importance of the problem, I've thought a lot about which—if either—is the "right" approach.

Coming in, I was mostly on the Engineering side, although I had more sympathy for Philosophy than the median ML researcher (who has ~0% sympathy for Philosophy). However, I now feel that:

- **Philosophy is significantly underrated by most ML researchers.**
- The Engineering worldview, taken seriously, actually implies assigning significant weight to thought experiments.

On the other hand, I also feel that:

- Philosophy continues to significantly underrate the value of empirical data.
- Neither of these approaches is satisfying and we actually have **no single good approach** to thinking about risks from future ML systems.

I've reached these conclusions through a combination of thinking, discussing with others, and observing empirical developments in ML since 2011 (when I entered the field). I've distilled my thoughts into a series of blog posts, where I'll argue that:

1. [Future ML Systems Will be Qualitatively Different](#) from those we see today. Indeed, ML systems have historically exhibited qualitative changes as a result of increasing their scale. This is an instance of "More Is Different", which is commonplace in other fields such as physics, biology, and economics (see *Appendix: More Is Different in Other Domains*). Consequently, we should expect ML to exhibit more qualitative changes as it scales up in the future.
2. Most discussions of ML failures are anchored either on existing systems or on humans. [Thought Experiments Provide a Third Anchor](#), and having three anchors is much better than having two, but each has its own weaknesses.
3. If we take thought experiments seriously, we end up predicting that [ML Systems Will Have Weird Failure Modes](#). Some important failure modes of ML systems will not be present in any existing systems, and might manifest quickly enough that we can't safely wait for them to occur before addressing them.
4. My biggest disagreement with the Philosophy view is that I think [Empirical Findings Generalize Surprisingly Far](#), meaning that well-chosen experiments on current systems can tell us a lot about future systems.

This post is the introduction to the series. I'll post the next part each Tuesday, and update this page with links once the post is up. In the meantime, leave comments with any thoughts you have, or contact me if you'd like to preview the upcoming posts and leave feedback.

# Prizes for ELK proposals

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

**We are no longer accepting submissions. We'll get in touch with winners and make a post about winning proposals sometime in the next month.**

[ARC](#) recently released a technical report on [eliciting latent knowledge](#) (ELK), the focus of our current research. Roughly speaking, the goal of ELK is to incentivize ML models to honestly answer “straightforward” questions where the right answer is unambiguous and known by the model.

ELK is currently unsolved in the worst case—for every training strategy we’ve thought of so far, we can describe a case where an ML model trained with that strategy would give unambiguously bad answers to straightforward questions despite knowing better. Situations like this may or may not come up in practice, but nonetheless we are interested in finding a strategy for ELK for which we can’t think of any counterexample.

We think many people could potentially contribute to solving ELK—there’s a large space of possible training strategies and we’ve only explored a small fraction of them so far. Moreover, we think that trying to solve ELK in the worst case is a good way to “get into ARC’s headspace” and more deeply understand the research we do.

**We are offering prizes of \$5,000 to \$50,000 for proposed strategies for ELK.**  
We’re planning to evaluate submissions received before February 15.

For full details of the ELK problem and several examples of possible strategies, see [the writeup](#). The rest of this post will focus on how the contest works.

## Contest details

**To win a prize, you need to specify a training strategy for ELK that handles all of the counterexamples that we’ve described so far, summarized in the section below**—i.e. where the breaker would need to specify something new about the test case to cause the strategy to break down. You don’t need to fully solve the problem in the worst case to win a prize, you just need to come up with a strategy that requires a new counterexample.

We’ll give a \$5,000 prize to any proposal that we think clears this bar. We’ll give a \$50,000 prize to a proposal which we haven’t considered and seems sufficiently promising to us or requires a new idea to break. We’ll give intermediate prizes for ideas that we think are promising but we’ve already considered, as well as for proposals that come with novel counterexamples, clarify some other aspect of the problem, or are interesting in other ways. A major purpose of the contest is to provide support for people understanding the problem well enough to start contributing; we aren’t trying to only reward ideas that are new to us.

You can submit multiple proposals, but we won’t give you separate prizes for each—we’ll give you at least the maximum prize that your best single submission would have received, but may not give much more than that.

If we receive multiple submissions based on a similar idea, we may post a comment describing the idea (with attribution) along with a counterexample. **Once a counterexample has been included in the comments of this post, new submissions need to address that counterexample** (as well as all the existing ones) in order to be eligible for a prize.

Ultimately prizes are awarded at our discretion, and the “rules of the game” aren’t fully precise. If you are curious about whether you are on the right track, feel free to send an email to [elk@alignmentresearchcenter.org](mailto:elk@alignmentresearchcenter.org) with the basic outline of an idea, and if we have time we’ll get back to you with some feedback. [Below](#) we also describe some of the directions we consider more and less promising and some general guidance.

## How to submit a proposal

You can submit a proposal by copying this [google doc template](#) and sharing it with [elk@alignmentresearchcenter.org](mailto:elk@alignmentresearchcenter.org) (please give comment access in case we need to ask questions to evaluate your submission). By submitting a proposal, you are giving us permission to post the idea here with attribution. (And it’s fine for you to post it after the contest or we post a counterexample.)

## Retroactive prizes

We’re giving prizes to existing proposals from [David Dalrymple](#) (\$5k), [Ramana Kumar](#) (\$3k), [John Maxwell](#) (\$2k), and [Yonadav Shavit](#) (\$1k). We’ve added relevant counterexamples in the next section, and further submissions need to handle those counterexamples in order to be eligible.

## Existing counterexamples

In [our report](#) and a [recent post](#), we described some features of the world that would cause existing approaches to ELK to fail. A new approach should work even in a situation satisfying all of these properties, or should include an argument for why some of these features are actually impossible or inconsistent with each other.

In order to generate a counterexample to a training strategy, we need to spell out a concrete “way the world could be” (a test case) where the strategy would fail. In the report we consider test cases where the AI and the human happen to understand the world by doing inference in Bayes nets. **The Bayes net case is just one test case that lets us get precise about our counterexamples and our training strategies don’t lean on it.** In practice we also consider other concrete cases (e.g. deduction from premises generated by a neural network).

In this section we’ll briefly review the current state of existing counterexamples. This isn’t intended to be a self-contained description; it relies heavily on links to the report.

## Ontology mismatch

All of our counterexamples are based on an [ontology mismatch](#) between two different Bayes nets, one used by an ML prediction model (“the predictor”) and one used by a

human:

- The predictor's Bayes net is used internally by a prediction model that maps an initial segment of a video to a prediction about what will happen next. This model was trained by gradient descent based on the quality of its predictions ([more](#)).
- Humans have their own model of the world, which they can improve with the help of AI assistants; those assistants can design and conduct experiments, find and evaluate arguments, and so on. Crucially, we have a limited computational budget with which to train these AI assistants, and that limits how much help they can provide. The "human's" Bayes net captures everything the human understands about the world at the end of this process ([more](#)).

We are interested in the case where there are features of reality that are accurately reflected in the predictor's Bayes net, but missing from this (augmented) human Bayes net. We expect the human could understand these facts eventually given enough time or enough AI assistance, but not using the available computational budget. Put differently, there is a gap between the kinds of understanding that gradient descent can build, and the kind of understanding that a human-comprehensible process of "scientific inquiry" can build using similar computational resources ([more](#)).

Furthermore, we imagine that both the predictor's Bayes net and the human's Bayes net reflect some underlying reality and that there is a way to map between them, which we'll call "direct translation." We don't know how to do direct translation, and we are concerned that if we simply search for any mapping from the AI's Bayes net to the human's Bayes net we will instead get the "human simulator": the map which correctly translates the observations from one Bayes net to the other, but then does inference in the human's Bayes net in order to fill in the latent structure as a human would by looking at the data ([more](#)).

## Counterexample features

ELK seems particularly hard when these Bayes nets satisfy the following properties:

1. There are arbitrarily severe "ontology mismatches" between the predictor's Bayes net and the human's Bayes net, such that specifying the direct translation between them can be very complex ([more](#)).
2. The human's Bayes net is simpler than the predictor's Bayes net, and the gap can be arbitrarily large ([more](#)).
3. The predictor's Bayes net is arbitrarily large, and individual nodes of the human's Bayes net depend on a constant fraction of the predictor's Bayes net. Therefore directly translating between the two Bayes nets takes arbitrarily much computation ([more](#)).
4. It's computationally easy for the human simulator to do good inference in the human's Bayes net—at least good enough inference that it answers questions as well as the direct translator—and the gap between this complexity and the complexity of direct translation can be arbitrarily large. More precisely, a human interrogator can't think of questions which require the human simulator to do hard inference. However, we can't rule out the possibility that you, the algorithm-designer, will think of some trick they can use to generate hard questions ([more](#)).

5. There are situations where a model which is trying to answer a question will *first* be able to predict what (potentially mistaken) belief the human-with-assistance will have about that question before it knows the true answer to the question ([more](#)).
6. While the predictor's Bayes net is very *large*, it has a short succinct description which is simpler than the human's Bayes net, and that gap is arbitrarily large. Think of simple physical laws that describe arbitrarily computationally expensive simulations, or simple cognitive heuristics that are very widely applicable. ([more](#))
7. There are sequences of actions for which the predictor's Bayes net correctly anticipates that tampering will occur yet nothing looks suspicious to the human's Bayes net.
8. After tampering has occurred, it is possible that no sequence of further actions reveal that anything is wrong, e.g. the camera and actuators may have been completely disconnected from the world and replaced with a convincing-to-humans simulation ([more](#)).
9. There are sequences of actions leading to undetectable tampering that have no human-comprehensible similarities to sequences of actions leading to detectable tampering; in particular, the early steps of this undetectable tampering cannot be used to perform detectable tampering ([more](#)).

An important and very general strategy for the breaker is to suggest counterexamples where you “get what you train for.” If you propose training a model using loss function L, then the breaker will likely focus on counterexamples where the model chooses its output to minimize L. A successful approach should either explain why this either isn’t a realistic thing for the model to learn (e.g. why that would be disfavored by regularization), or why the overall approach works even if the model is explicitly strategizing about how to minimize the loss.

## Rough guidance

- There is a tradeoff between specifying your approach precisely, and including your own counterexample. It’s fine to describe a very broad/vague idea for solving ELK, and then present a new counterexample showing why that whole category of solutions can’t work. But if you don’t propose a counterexample, then it’s good to describe your proposal precisely enough that we understand how to actually implement it without wondering if that’s really what you meant. It’s OK to submit a very broad or informal idea together with a single very specific instance of that idea, as long as there is some version we can understand precisely.
- We suspect you can’t solve ELK just by getting better data—you probably need to “open up the black box” and include *some* term in the loss that depends on the structure of your model and not merely its behavior. So we are most interested in approaches that address that challenge. We could still be surprised by clever ways to penalize behavior, but we’ll hold them to a higher bar. The most plausible surprise would be finding a way to reliably make it computationally difficult to “game” the loss function, probably by using the AI itself to help compute the loss (e.g. using consistency checks or by giving the human AI assistance).
- If you are specifying a regularizer that you hope will prefer direct translation over human simulation, you should probably have at least one concrete case in mind that has all the counterexample-features [above](#) and where you can confirm that your regularizer does indeed prefer the direct translator.

- ELK already seems hard in the case of ontology identification, where the predictor uses a straightforward inference algorithm in an unknown model of the world (which we've been imagining as a Bayes net). When coming up with a proposal, we don't recommend worrying about cases where the original unaligned predictor learned something more complicated (e.g. involving learned optimization other than inference). That said, you do need to worry about the case where your training scheme *incentivizes* learned optimization that may not have been there originally.

## Ask dumb questions!

A major purpose of this contest is to help people build a better understanding of our research methodology and the “game” we are playing. So we encourage people to ask clarifying questions in the comments of this post (no matter how “dumb” they are), and we’ll do our best to answer all of them. You might also want to read the comments to get more clarity about the problem.

## What you can expect from us

- We'll try to answer all clarifying questions in the comments.
- If you send in a rough outline for a proposal, we will try to understand whether it might qualify and write back something like “This qualifies,” “This might qualify but would need to be clearer and address issue X,” “We aren’t easily able to understand this proposal at all,” “This is unlikely to be on track for something that qualifies,” or “This definitely doesn’t qualify.”
- If there are more submissions than expected, we may run out of time to respond to all submissions and comments, in which case we will post an update here.

# Use Normal Predictions

Making predictions is a good practice, writing them down is even better.

However, we often make binary predictions when it is not necessary, such as

- Biden win popular vote: 91%
- Danish COVID deaths above 10.000 by January 1. 2022: 84%

Alternatively, we could make predictions from a normal distribution, such as ('~' means 'comes from'):

- Biden's popular vote  $\sim N(0.54, 0.03)$
- Danish COVID deaths by January 1. 2022  $\sim N(15,000, 5,000)$

While making "Normal" predictions seems complicated, this post should be enough to get you started, and more importantly to get you a method for tracking your calibration, which is much harder with dichotomous predictions.

The key points are these:

1. Predicting from a normal is surprisingly easy.
2. Getting an actionable number for how over/under confident you are requires only simple math!
3. The normal distribution carries more information than the Bernoulli (binary outcome such as coins) and will therefore give you more information to act on!

Things this post will answer:

1. How do I make a normal prediction?
2. Why do I want to do this?
3. How do I track my calibration?

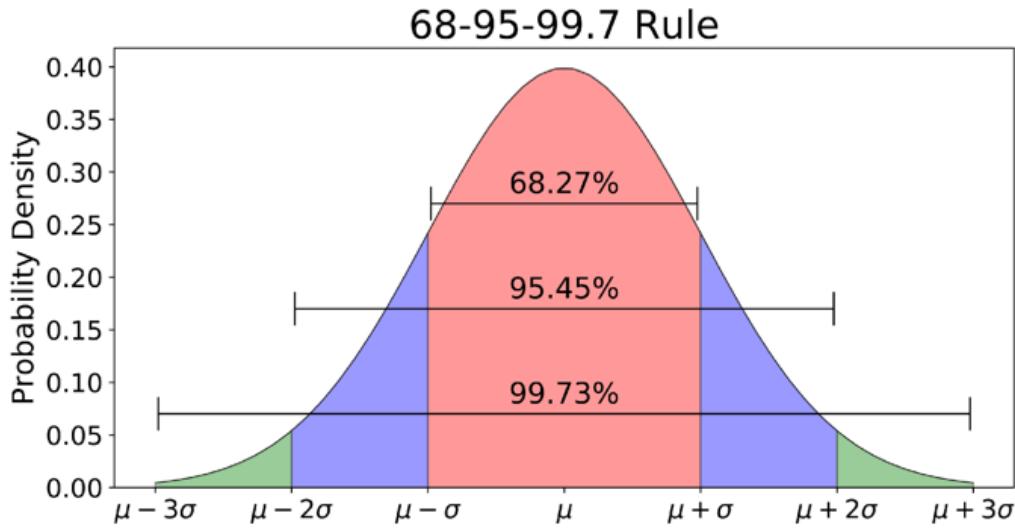
## Quick recap about the normal distribution

The normal distribution is usually written as  $N(\mu, \sigma)$  has 2 parameters:

- a location parameter  $\mu$  (pronounced mu) which is both the most likely and the average value
- a scale parameter  $\sigma$  (pronounced sigma) which captures uncertainty, high  $\sigma$  implying high uncertainty

the [68-95-99.7](#) rule states that:

- 68% of your predictions should fall in  $\mu \pm 1\sigma$
- 95% of your predictions should fall in  $\mu \pm 2\sigma$
- 99.7% of your predictions should fall in  $\mu \pm 3\sigma$



50% of the predictions should fall within  $0.674 \approx \frac{1}{2}\sigma$ , which can be used as a quick spot check.

The last piece of Normal trivia we need to know is this: the variance of the Normal is simply  $\sigma^2$ :

$$\text{Var}(N(\mu, \sigma)) = \sigma^2$$

## How to make predictions

To make a prediction, there are two steps. Step 1 is predicting  $\mu$ . Step 2 is using the 68-95-99.7 rule to capture your uncertainty in  $\mu$ .

I tried to predict Biden's national vote share in the 2020 election. From the polls, I got 54% as a point estimate, so that seemed like a good guess for  $\mu$ . For  $\sigma$  I used the 68-95-99.7 rule and tried to see what that would imply for different values of  $\sigma$ . Here is a table for 2-5%

<b>Intervals</b>	<b>68%</b>	<b>95%</b>	<b>99.7%</b>
$\sigma = 2$	52-56%	50-58%	48-60%
$\sigma = 3$	51-57%	48-60%	45-63%
$\sigma = 4$	50-58%	46-62%	42-66%
$\sigma = 5$	49-59%	44-64%	39-69%

$\sigma = 2$  implies a 97.5% (95% interval + half a tail) chance that Biden would get more than 50% of the votes; I was not that confident.  $\sigma = 4$  implies a 84% chance that Biden would get

more than 50% of the votes ( $68\% + 32\%/2$ ), and a 16% chance Trump wins, I likewise found this too high, so I settled on  $\sigma = 3$ .

## Why do I want to do this

Biden Got 52% of the vote share, which was within 1 sigma of my prediction. There are two weak lessons that I drew from this ONE data point.

1. The pollsters screwed up, so I should have regressed  $\mu$  towards the mean (50%), such as predicting 53% instead of 54%
2. The prediction was exactly  $\frac{1}{2}\sigma$  from  $\mu$ , so the  $\sigma$  was on the 50%/50% boundary just as expected. This was lucky, but it's weak evidence that the  $\sigma$  was well chosen.

Imagine I instead had predicted Biden wins (the popular vote) 91%, well guess what he won, so I was right... and that is it. Thinking I should have predicted 80% because the pollsters screwed up seems weird, as that is a weaker prediction and the bold one was right! I would need to predict a lot of other elections to see whether I am over or under confident.

## How to track your calibration

**Note:** In the previous section we used  $\mu$  and  $\sigma$  for predictions. In this section we will use  $\mu_i$  and  $\sigma_i$  where  $i$  is the index (prediction 1, prediction 2... prediction N). We will use  $\hat{\sigma}_z$  for the calibration point estimate; this means that  $\hat{\sigma}_z$  is a number such as 1.73. In the next post in this series, we will use  $\sigma_z$  for the calibration distribution, this means that  $\sigma_z$  is a distribution like your predictions and thus has an uncertainty.

I also made a terrible prediction, during the early lock down in 2020. I predicted  $N(15,000, 5,000)$  COVID deaths by 2022 in Denmark. It turned out to be 3,200, which is  $\frac{15000 - 3200}{5000} = 2.36$  standard deviations away, so outside the 95% interval!

In this section we will transform your predictions to the Unit normal. This is called z-scoring, because if all predictions are on the same scale, then they are comparable

$$z_{ideal} \sim N(0, 1)$$

$$\text{Var}(z_{ideal}) = 1^2 = 1$$

Normally when you convert to z-scores you use the data itself to calculate  $\mu$  and  $\sigma$ , which guarantee a  $N(0,1)$ . Here, we will use our predicted  $\mu$  and  $\sigma$ . This means there will be a discrepancy between  $z_{ideal}$  and our  $z$ . This discrepancy describes how under/over confident your intervals are, and thus describes your calibration, such that if  $\hat{\sigma}_z = 2$  then all your intervals should be twice as wide to achieve  $z_{recalibrated} \sim N(0, 1)$

First we z-score our data by calculating how many  $\sigma$  they are away from the observed data  $x$ , using this formula:

$$z_i = \frac{x_i - \bar{x}_i}{\sigma_i}$$

Second we calculate  $\hat{\sigma}_z$  as the RMSE (root mean squared error) of all predictions:

$$\text{Var}(z) = \sum_{i=0}^N \text{Var}(z_i) = \sum_{i=0}^N z_i^2$$

$$\hat{\sigma}_z = \sqrt{\frac{\text{Var}(z)}{N}}$$

And that is, let's calculate  $\hat{\sigma}_z$  for my two predictions, first we calculate the variances:

- $\text{Var}(z_1) = (54.5 - 52)^2 = 4$
- $\text{Var}(z_2) = (15000 - 32)^2 = 5.57$

Then we calculate  $\hat{\sigma}_z$

$$\hat{\sigma}_z = \sqrt{\frac{\text{Var}(z)}{N}} = \sqrt{\frac{4 + 5.57}{2}} = 1.73$$

So if these were my only two predictions, then I should widen my future intervals by 73%. In other words, because  $\hat{\sigma}_z$  is 1.73 and not 1, thus my intervals are too small by a factor of 1.73. If I instead had gotten  $\hat{\sigma}_z < 1$ , such as  $\hat{\sigma}_z = 0.5$  then this would be evidence that my intervals were too wide and should be "scaled back" by multiplying my intervals by 0.5.

## Still not convinced?

Here are some bonus arguments:

1. **Weak 50/50:** Sometimes you are actually 50/50, such as Scott's prediction that Bitcoin had a 50-50 shot of going over 3000 in 2019; that could be reformulated as "Bitcoin  $\sim N(3000, 1500)$ " such that a price of 10000 counts against the prediction. Now a weak prediction still gives evidence of calibration!
2. **Overshooting and Undershooting:** If Biden had gotten 20 or 80% of the votes, both things would be strong evidence of my prediction being wrong, where the binary predictions can only be 'wrong in one direction'
3. **High Confidence Predictions are easier to calibrate:** In Binary land a 99% prediction is very hard to calibrate because you need to make hundreds of them to get enough data (unless many turn out wrong of course). A Corresponding Normal

prediction would have a small  $\sigma$  and thus give as much evidence of calibration as a 60% prediction.

4. **Right for the Wrong Reason:** All of  $N(50.67, 0.5)$ ,  $N(54, 3)$ ,  $N(58, 6)$ , give Biden a 91% win chance, but for very different reasons, and will thus lead people to update differently after observing  $x = 52$ .

## Advanced Techniques

Sometimes your beliefs do not follow a Normal distribution. For example, the Bitcoin prediction  $N(3000, 1500)$  implies I believe there is a 2.5% chance the price will become negative, which is impossible. There are 3 solutions in increasing order of fanciness to deal with this:

1. Have different  $\sigma$  for each direction such as (HN = Half Normal):

$$\text{Bitcoin} \sim 0.5\text{HN}(3000, \sigma_{\text{up}} = 3000) + 0.5\text{HN}(3000, \sigma_{\text{down}} = 1500)$$

This means if it's above then  $\sigma = 3000$ , while if it's below then  $\sigma = 1500$ . If you do this, then you can use "the relevant  $\sigma$ " when calibrating and ignore the other one, so if the price of bitcoin ended up being  $x = 10.000$  then  $z$  becomes  $\frac{-3}{3}$ :

$$z = \frac{10.000 - 3000}{3000} = -3$$

2. Often you believe something goes up or down by a factor, such as Bitcoin dropping to half or doubling. For ease of example let imagine that Scott thought there was a 68% chance that Bitcoin's value would change by less than a factor of 2.

$$\mu = \log_{10}(3000) \approx 3.477$$

$$\sigma = \log_{10}(2) \approx 0.301$$

$$\log_{10}(\text{Bitcoin}) \sim N(3.477, 0.301)$$

$z$ -scoring works the same way, so if the Bitcoin price was 10.000 then:

$$\log_{10}(x) = \log_{10}(10.000) = 4$$

$$z \approx \frac{3.477 - 4}{0.301} \approx -1.74$$

3. (If this makes no sense, then ignore it): Using an arbitrary distribution for predictions, then use its CDF (Universality of the Uniform) to convert to  $U(0, 1)$ , and then transform to  $z$ -score using the inverse CDF (percentile point function) of the Unit Normal. Finally use this as  $z_i$  in when calculating your calibration.

## Final Remarks

I want you to stop and appreciate that we can get a specific actionable number after 2 predictions, which is basically impossible with binary predictions! So start making normal predictions, rather than dichotomous ones!

As a final note, keep this distinction in mind:

1. If the data  $x$  and the prediction  $\mu$  are close, then you are a good predictor
2. If the mean prediction error on the z-scale is close to 1, then you are a well calibrated predictor.

Getting good at 1 requires domain knowledge for each specific prediction, while getting good at 2 is a general skill that applies to all predictions.

This post we calculated the point estimate  $\hat{\sigma}_z = 1.73$  based on 2 data points. There is a lot of uncertainty in a point estimate based on two data points, so we should expect the calibration distribution over  $\sigma_z$  to be quite wide. The next post in this series will tackle this by calculating a Frequentest confidence interval for  $\hat{\sigma}_z$  and a Bayesian posterior over  $\sigma_z$ . This allows us to make statements such as: I am 90% confident that  $1 < \sigma_z$ , so it's much more likely that I am badly calibrated than unlucky. With only two data points it is however hard to tell the difference with much confidence.

Finally I would like to thank my editors Justis Mills and [eric135](#) for making this readable.

# Long covid: probably worth avoiding—some considerations

I hear friends reasoning, “I’ll get covid eventually and long covid probably isn’t that bad; therefore it’s not worth much to avoid it now”. Here are some things informing my sense that that’s an error:

## A. Really bad anecdotes aren’t hard to find.

I [asked](#) for people’s long-covid experiences on Twitter and got more than a hundred comments mostly full of truly awful stories, almost entirely from people I don’t know at all, presumably somehow brought via the active Long Covid community.

This is a bit hard to interpret as a datum, due to the intense selection effects involved, but my main inferences are:

1. The bad tail gets very bad.
2. This bad tail is less like a quantitative reduction in labor, and more like psychological torment, being very sleep deprived or brain damaged while dealing with a host of legitimately scary health problems (see below) than I had been picturing.
3. The very bad part of the bad tail is not tiny. Like, if I requested lightning strike experiences, I don’t think I would get this kind of response.

See [this](#) for more heavily selected but still scary anecdotes about what bad cases can look like.

## B. Bad anecdotes are common enough to show up in my vicinity.

Among this cascade of terrible stories is a response from one of the handful of people from my extended network who I already knew had had covid, Oxford machine learning professor Michael Osborne. I take this as strong evidence that that level of bad experience isn’t vanishingly rare, though he has been well for the last few months, so this is only direct evidence of this kind of thing happening for about two years.

Excerpts from [his account](#):

My own low-points: early on, I collapsed, shaking, and was taken to A&E in an ambulance. A year later, I did not have the energy to leave the house. Formerly, I was a marathon runner, but I brought on a bad relapse with a 700m walk. Many people have it much, much, worse. 2/

— Michael A Osborne (@maosbot) [October 19, 2021](#)

Brain fog is a bit like being *extremely* sleep-deprived—remember, sleep deprivation is literally a technique of torture—but you can’t sleep off brain fog. It feels like being lost in a fog, sensing dark shapes shifting around you, losing yourself. 6/

— Michael A Osborne (@maosbot) [October 19, 2021](#)

"Fatigue" similarly does no justice to the experience. Fatigue means there is less of you. You are less. You can't, just can't, get out of bed. This is *not* psychological (although fatigue may *cause* depression!). Your body, physically, does not have the energy. 7/

— Michael A Osborne (@maosbot) [October 19, 2021](#)

(Another distant relative in my generation just told me that they have long covid, but I take that as much less evidence, since I only know about them having covid at all because they saw this long covid discussion.)

### **C. Rates of ambiguously-maybe-quite-bad symptoms seem very high, even for people who only had mild covid**

[This](#) norwegian study, n = 70k, has, for mild cases (in a sample mixing people who had covid 1-6 months prior with people who had it 11-12 months prior):

- 10.2% with fatigue (**6.4% higher** than control in the 11-12 months case),
- 7.1% with poor memory (**3.5% higher** than previous control),
- 9.2% with brain fog (**5.3% higher** than previous control).
- 6.9% with shortness of breath (**5.6% higher** than previous control)

These numbers are for unvaccinated people; I'd maybe reduce them by a factor of two for being vaccinated, deferring to [Matt Bell](#), who reasons, "A very recent [report](#) from the UK claims that full vaccination cuts the risk of lingering symptoms 28+ days after catching COVID by roughly half".

[This](#) meta-analysis of 81 studies finds:

- "Approximately 1 in 3 individuals experienced fatigue 12 or more weeks following COVID-19 diagnosis."
- "Approximately 1 in 5 individuals exhibited cognitive impairment 12 or more weeks following COVID-19 diagnosis."

(Though I am told that lot of the studies involved might have been small-n, hospitalized, no control, so I'm not sure how they dealt with that.)

[This survey](#) suggests that among people with long covid (I think drawn from a support group), some especially undesirable symptoms are very: "88.0% of the total respondents (3310) experienced either cognitive dysfunction or memory loss (Figure 8)."

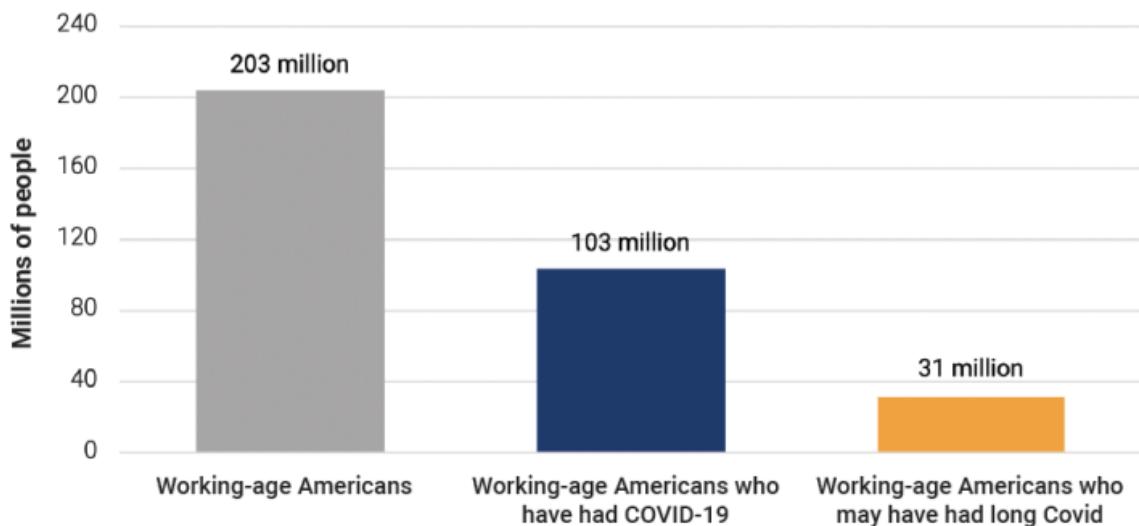
All of these symptoms come in degrees between slightly annoying and life-wrecking, making all of this hard to interpret. Maybe we can look at something more objective, like 'can the person still do their job?'

### **D. It looks like a lot of people can no longer do their jobs**

- Katie Bach of Brookings argued a few days ago that an estimate of 1.1 million people out of work because of long covid is reasonable, out of 103 million working age Americans she estimated had had covid, i.e. a roughly 1% chance of being entirely out of work:

**Figure 1: More than 30 million working-age Americans may have had long Covid**

That is one out of seven working-age Americans

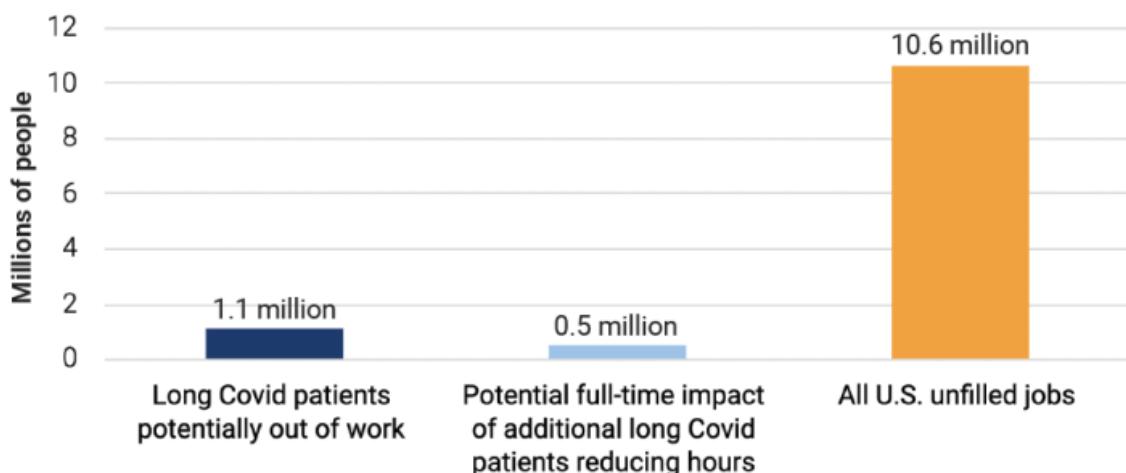


*Source: Author's estimate based on data from U.S. Census Bureau, CDC, and UC Davis Health.*

B | Brookings Metro

## Figure 2: The U.S. labor market may be missing 1.6 million full-time workers due to long Covid

That would account for 15% of all unfilled jobs



Source: Author's estimate based on data from the UK Trades Union Congress, The Lancet, and the Bureau of Labor Statistics

B | Brookings Metro

- [Washington post](#): "Hard data is not available and estimates vary widely, but based on published studies and their own experience treating patients, several medical specialists said 750,000 to 1.3 million patients likely remain so sick for extended periods that they can't return to the workforce full time."
- [This](#) meta-analysis of 81 studies I mentioned earlier also looked at work: "29.0% and 47.4% of those who were employed premorbidly were not able to return to work"; "5% to 90% were unable to reach their pre-COVID employment level" (p. 128) (As noted earlier, a lot of the studies in the meta-analysis seem to be small n, involving hospitalized people, without controls, and I don't know what they did about this. Also, it's possible I'm misunderstanding what group the meta-analysis is about, given how crazy high the numbers are).

Some harder to interpret data about long covid sufferers in particular (where I'm not sure how many people count as that) still suggests pretty major issues:

[Matt bell](#) says that [this](#) UK data-set has ~18% of non-hospitalized long covid sufferers with "activities limited a lot."

And looking at a [survey](#) I mentioned earlier again, people with long covid (and probably under selection to be relatively bad cases) report large losses of ability to work, and in other specific capabilities:

- "The greatest area of impact reported was on work, with 86.2% (95% confidence interval 84.4 to 88.0%) of working respondents feeling mildly to severely unable to work - 29.1% (26.7% to 31.6%) severely.... Other areas of impact included making serious decisions 85.3% (80.7% to 89.8%), communicating thoughts and needs 74.8% (72.5% to 77.1%), having conversations with others 68.3% (65.8%

to 70.8%), maintaining medication schedules 62.5% (59.8% to 65.1%), following simple instructions 54.4% (51.6% to 57.2%), and driving 53.2% (50.5% to 56.0%).” ([p. 16](#)).

- “Of unrecovered respondents who worked before becoming ill, only 27.3% (95% confidence interval 25.3% to 29.4%) were working as many hours as they were prior to becoming ill at the time of survey, compared to 49.3% (40.8% to 57.9%) of recovered respondents (see Figure 11d). Nearly half 45.6% (43.2% to 48.0%) of unrecovered respondents were working reduced hours at the time of the survey, and 23.3% (21.3% to 25.4%) were not working at the time of the survey as a direct result of their illness. This included being on sick leave, disability leave, being fired, quitting, and being unable to find a job that would accommodate them. The remaining respondents retired, were volunteers, or did not provide enough information to determine their working status. Overall, 45.2% (42.9% to 47.2%) of respondents reported requiring a reduced work schedule compared to pre-illness. 22.3% (20.5% to 24.3%) were not working at the time of survey due to their health conditions.” ([p. 27](#)).

#### **E. Other people's previous back of the envelope calculations on this are not reassuring.**

[Matt bell:](#)

“If you’re a 35 year old woman, and your risk of ending up with lifelong long COVID from catching COVID is 2.8%, then catching COVID would be the same, statistically speaking, as losing  $(50 \text{ years} * 0.18 * 0.028 * 365 \text{ days/year}) = \sim 90 \text{ days of your life.}$ ”

[Scott Alexander:](#)

“Your chance of really bad debilitating lifelong Long COVID, conditional on getting COVID, is probably somewhere between a few tenths of a percent, and a few percent.”

#### **F. Having ‘survived’ covid looks associated with a 60% increased risk of death (after surviving covid) during the following six months**

According to a [massive controlled study published in Nature](#) (more readable summary [here](#)). It also looks like they are saying that this is for non-hospitalized covid patients, though the paper is confusing to me.

I’m not sure whether to model this as some longer-than-a-few-weeks period of danger that is still done within the year (maybe 400 micromorts if done by six months), versus semi-permanently worse health (which would maybe be like aging about 8 years).

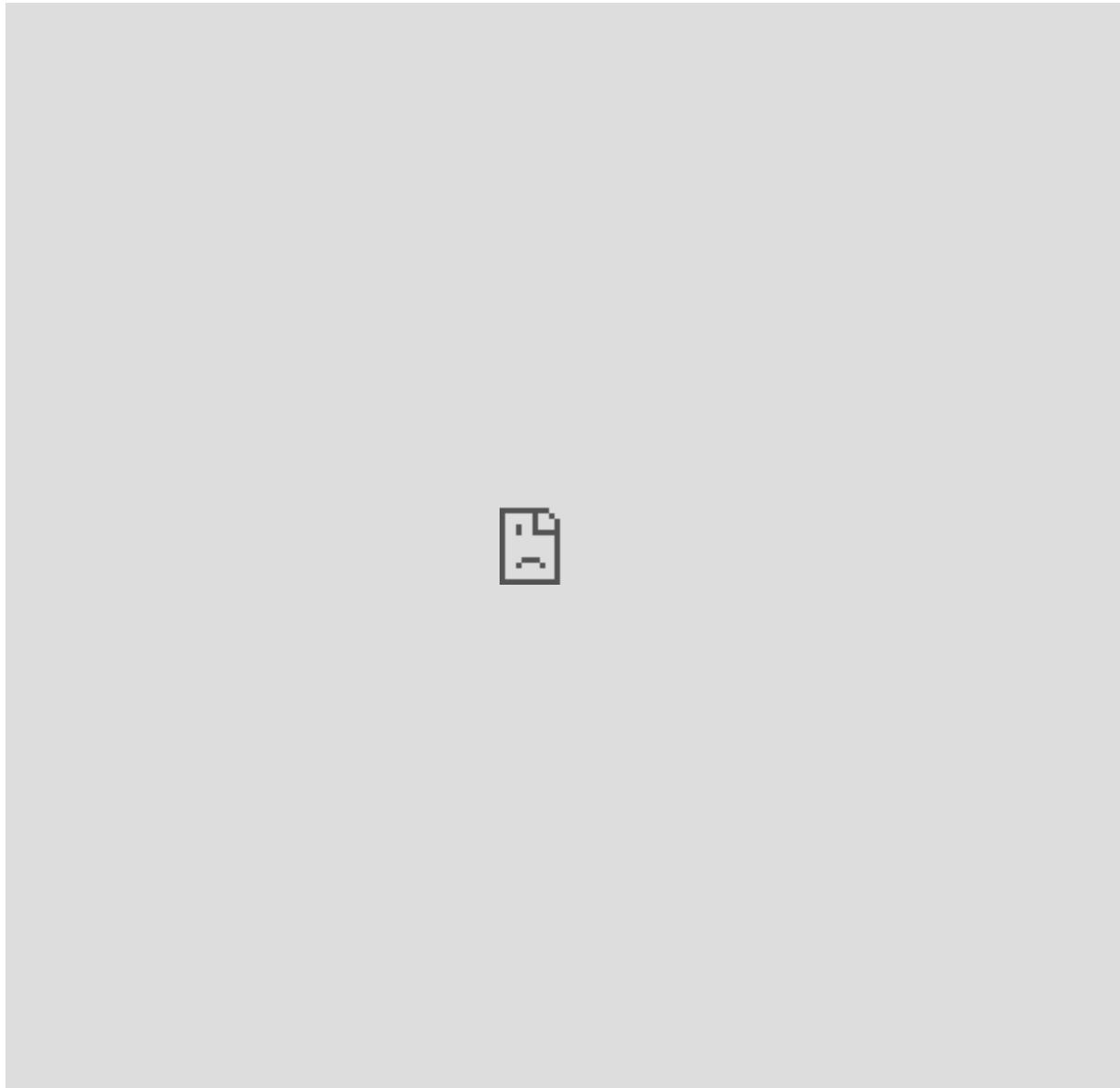
This is bad because death, but I’m maybe more alarmed by it because it supports the ‘post-covid illness is some fairly extreme body fuckage of which the fact that some people can’t think straight is merely a tip-of-the-iceberg symptom’ hypothesis over the ‘most of the story is that you lose some work time this year’ hypothesis. Both because death really suggests something wrong, and because they catalogue a lot of things wrong—from the [WebMD summary](#): “The patients had a high rate of stroke and other nervous system ailments; mental health problems such as depression; the onset of diabetes; heart disease and other coronary problems; diarrhea and digestive disorders; kidney disease; blood clots; joint pain; hair loss; and general fatigue.”

It’s also a bad sign about what a bad time you might have during that year. I don’t think these people go gently - it seems like a repeating theme of these long covid stories is that their victims have a lot of health scares and medical investigation, and

this mortality data suggests to me that they are genuinely at risk of their lives, which I think would make the experience much worse, for me at least.

**G. Overall deaths from everything have been very unusually high at points in 2021, even in 15-64 age group**

[From Our World In Data:](#)



[For context:](#)

(The Center Square) – The head of Indianapolis-based insurance company OneAmerica said the death rate is up a stunning 40% from pre-pandemic levels among working-age people.

"We are seeing, right now, the highest death rates we have seen in the history of this business – not just at OneAmerica," the company's CEO Scott Davison said during an online news conference this week. "The data is consistent across every player in that business."

OneAmerica is a \$100 billion insurance company that has had its headquarters in Indianapolis since 1877. The company has approximately 2,400 employees and sells life insurance, including group life insurance to employers nationwide.

Davison said the increase in deaths represents "huge, huge numbers," and that's it's not elderly people who are dying, but "primarily working-age people 18 to 64" who are the employees of companies that have group life insurance plans through OneAmerica.

"And what we saw just in third quarter, we're seeing it continue into fourth quarter, is that death rates are up 40% over what they were pre-pandemic," he said.

"Just to give you an idea of how bad that is, a three-sigma or a one-in-200-year catastrophe would be 10% increase over pre-pandemic," he said. "So 40% is just unheard of."

**H. Sounds like these things involve various damage throughout body, and my guess is that that ends up being regrettable in ways not captured in 'hours lost to fatigue this year'**

See [Nature study](#) in F. I also feel like I've seen this a lot, but don't have that many examples immediately on hand. Here's [one other example](#), not ideal because note that these are hospitalized younger people:

For people younger than 65 who were hospitalized with COVID-19, the risk of death in the 12 months after the infection was 233% higher than it was for people who did not have the disease, results published in the journal [Frontiers in Medicine](#) have shown.

Nearly 80% of all deaths of people in the study who had recovered from COVID-19 in the past 12 months were not due to cardiovascular or respiratory causes, suggesting that the impact of the virus is significant and wide-ranging, even after the initial infection has been fought off.

My guess is that all the symptoms are a spectrum, and if the worst looks like an unbelievable amount of cognitive impairment and a pot pourri of organ dysfunctions, or death post-infection, then probably everyone gets a handful of cognitive impairment and organ dysfunction.

**I. It's easy to just think about these worst case outcomes, but there are a lot of more probable non-worst case outcomes that would still be a non-negligible bummer.**

I see people mostly estimating the worst cases, but my guess is that the more probable non-worst case outcomes (e.g. lesser fatigues and brain fogs, kidney damage, arrhythmias etc), are not nothing.

**J. Future unknowns**

Across all diseases, how much of their disutility is obvious in the first two years? Saliently right now: we've had Epstein-Barr for ages and only now noticed that it apparently has a ~1/300 chance of causing MS, usually showing up between 20 and 40, long after the virus, and wreaking brutal destruction. I'm not sure whether we would realize how bad HIV was if it had appeared two years ago and lots of people had it, but nobody had had it for more than two years yet.

### **K. Long covid probably isn't psychosomatic**

A [French study](#) found that long covid is barely associated with having had covid according to an antibody test, yet associated with believing one has had covid (which itself is unrelated to the antibody test results).

At first I (and I think others) thought that if this wasn't some error, then long covid was likely psychosomatic and not caused by physically having covid. But on further thought, that's totally wrong: this pattern could be caused by beliefs causing illness, but it could also be caused by illness causing beliefs, which obviously happens all the time. That is, people's guesses about whether they had covid are heavily influenced by their symptoms.

Imagine a population of a thousand people. A hundred of them get covid this year, and ten get long covid caused by the virus. Fifty and five of those respectively notice that they have covid at the time. Also a hundred people get other illnesses that look like long covid (other viral infections, other mental health problems, other random things). Everyone knows that a lot of people are getting covid and don't know they have covid. The hundred people who got some other illness that looks like long covid infer that they are plausibly in the asymptomatic covid group, since they seem to have long covid. Especially the ones who got some viral illness, so were actually a bit sick at around the time they developed their chronic illness. So we end up with people who think they have long covid having less covid than the general population, even though long covid caused by the covid virus is genuinely afflicting 10% of the covid infected. Note that everyone here might have non-psychosomatic illnesses.

It seems to me that we have other data that basically rules out the possibility that long covid is imaginary (e.g. see [Nature study above](#) on laboratory abnormalities and raised death rate). Though psychosomatic illness is weird - my understanding is that it could in principle still be psychosomatic, while yielding measurable physical laboratory abnormalities, though intuitively I'd be fairly surprised to learn that the same new psychosomatic syndrome had gripped millions in the wake of a particular belief they had, and raised their risk of death by half. Maybe I'm missing something here.

### **L. The general discussion of what is going on with people's brains sounds terrible**

The [list of different plausible routes to brain damage](#) occurring according to Nature—some brain cells getting broken, some blood restrictions causing little strokes, some immune system attacking your own brain type issues—is one I want very little to do with. For instance:

“...The researchers observed that, in slices of hamster brain, SARS-CoV-2 blocks the functioning of receptors on pericytes, causing capillaries in the tissue to constrict. “It turns out this is a big effect,” says Attwell.

It's a “really cool” study, says Spudich. “It could be something that is determining some of the permanent injury we see — some of these small-vessel strokes.”“

Another sample of current discussion of the brain damage situation that I really don't want to be a part of, this time from [NPR](#):

Frontera led a study that found that more than 13% of hospitalized COVID-19 patients had developed a new neurological disorder soon after being infected. A follow-up study found that six months later, about half of the patients in that group who survived were still experiencing cognitive problems.

The current catalog of COVID-related threats to the brain includes bleeding, blood clots, inflammation, oxygen deprivation and disruption of the protective blood-brain barrier. And there's new evidence in monkeys that the virus may also directly infect and kill certain brain cells...

"Frontera was part of a team that [studied](#) levels of toxic substances associated with Alzheimer's and other brain diseases in older COVID-19 patients who were hospitalized.

"The levels were really high, higher than what we see in patients that have Alzheimer's disease," Frontera says, "indicating a very severe level of brain injury that's happening at that time."...

Even COVID-19 patients who experience severe neurological problems tend to improve over time, Frontera says, citing unpublished research that measured mental function six and 12 months after a hospital stay.

"Patients did have improvement in their cognitive scores, which is really encouraging," she says.

But half of the patients in one study still weren't back to normal after a year...."

## **M. It sounds like covid maybe persists in your body for ages?**

Seems like the virus lives throughout your organs long after recovery, based on autopsies, including of mild/asymptomatic covid sufferers ([summary](#), [paper](#)):

"The study found that the virus had replicated across multiple organ systems even among patients with asymptomatic to mild COVID-19.

The virus was detected in all 44 cases and across 79 of 85 anatomical locations and body fluids sampled"

"We performed 72 complete autopsies on 44 patients with COVID-19 to map and quantify SARS-CoV-2 distribution, replication, and cell-type specificity across the human body, including brain, 74 from acute infection through over seven months following symptom onset. We show that 75 SARS-CoV-2 is widely distributed, even among patients who died with asymptomatic to 76 mild COVID-19, and that virus replication is present in multiple pulmonary and 77 extrapulmonary tissues early in infection. Further, we detected persistent SARS-CoV-2 78 RNA in multiple anatomic sites, including regions throughout the brain, for up to 230 days 79 following symptom onset."

This seems like a bad sign for future problems.

Ok, that was 'covid is bad, primarily due to long covid'. But aren't we all going to get it anyway?

## **N. Later rounds of covid are probably bad too**

This assumes that later covids are basically free, once you've done it once, in a way that isn't true for e.g. crashing your car. My guess is that later bouts are less bad on average, but far from free.

In [my survey](#), of three people with lasting problems who got covid at least twice, one got the problems with the first, one the second, and one said both contributed (though for the last person, the second was around a month ago). Not a great sample size, but seems like strong evidence that second-round long-covid isn't unheard of.

## **O. It's not 100% that you will get it.**

I'd guess there's a decent chance you'll be able to avoid it, via reasonable-to-pay costs. For instance, maybe omicron basically brings herd immunity. Maybe rapid tests get more reliable, easy, and cheap. Maybe long covid becomes a bigger concern, and people worry enough to get  $R_0$  generally below 1. Or other technologies improve: Fend really does cut covid exhalations massively and is taken up widely, or something like Novid finds a way to be taken up. (Or your path takes you to being a hermit or living in China, or we spend next year hiding in bunkers from cyborg bees, and you are glad you didn't pay all that covid risk up front for parties that aren't happening.)

(If we were doing exactly what we are doing so far, but with air purifiers in most buildings, would  $R_0$  have been  $\frac{1}{4}$  instead of  $\sim 1$ , and would it have died out? Is the problem that we are psychologically incapable of maintaining slightly more caution than needed to keep infections steady?)

Are things so predictable?

## **P. The likelihood of you getting it probably does depend on how bad it is**

Whether you will definitely get covid depends on your—and everyone's—assessment of how bad it would be to get. If the high end of possible long covid risks became clearly correct, probably many people wouldn't take those costs vs. avoiding people who choose to be risky, so the options for low risk lifestyle would improve. Like, at first when people realized that it was potentially quite valuable to wash your hands after various bodily things, it would have been pretty annoying to be one of the few people trying to avoid contact with others so called 'germs'. One might have resigned oneself to getting every disease contained in anyone's bodily fluids. But eventually the idea of hand-washing being important has had enough traction that everyone can care a lot about other people washing their hands after using the bathroom say, and it imposes very little burden, because basically everyone has habits and infrastructure that make it easy to uphold hygiene.

If everyone believed that covid was bad enough to get, it would die out in a month (possibly with terrible economic downsides for a month). It probably isn't that bad, but if it really is quite bad to get, and ways to avoid it exist, people will probably realize that eventually, and adjust.

## **Q. Getting covid later is probably better than earlier.**

So far this trend seems strong: I would rather get covid now than in March 2020. I expect more of this, from better knowledge, medicine, vaccines, and availability of everything we already have.

If I expect to get covid every year for the next five years, adding one more bout now is adding one more especially bad bout in expectation.

### **R. Huge VOI in following behind the crowd, at a safe distance**

I know many people resigning themselves to getting covid this month, and results so far suggest that a lot of them will indeed be infected.

If half of my friends are getting covid right now, there seems to be massive value in waiting another three months before reassessing precautions, so I can see how it goes for them. While noisy, I expect to get a better sense of the distribution of outcomes among people in my reference class—and especially whether there is a macroscopic chance of really dire outcomes—from this (I think regrettable) experiment than I get from cobbling together different studies about vaguely defined symptoms among ambiguous or not-quite-relevant demographics, with various guesstimated adjustments.

There also seems to be a huge amount of uncertainty about how bad long covid is. I and my friends have various intuitions and guesses about where assorted factors of two push in this direction or that, and I doubt any of us should be super confident that we are the one with the right intuitions and guesses.

I think waiting out this spike looks like the right move from a VOI perspective even if your own judgments come out in favor of long covid being probably fine. As long as you are decently uncertain—which it seems to me that you probably should be—there is a lot of information coming in (at least if you are someone in my social circles, but probably a lot of social circles are coming to have a lot more people with covid this month).

# [Intro to brain-like-AGI safety] 1. What's the problem & Why work on it now?

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

## 1.1 Post summary / Table of contents

This is the first of a [series of blog posts on the technical safety problem for hypothetical future brain-like Artificial General Intelligence \(AGI\) systems](#). So my immediate priority here is saying *what the heck is "the technical safety problem for brain-like AGI" and what do those words even mean and why on earth should I care*.

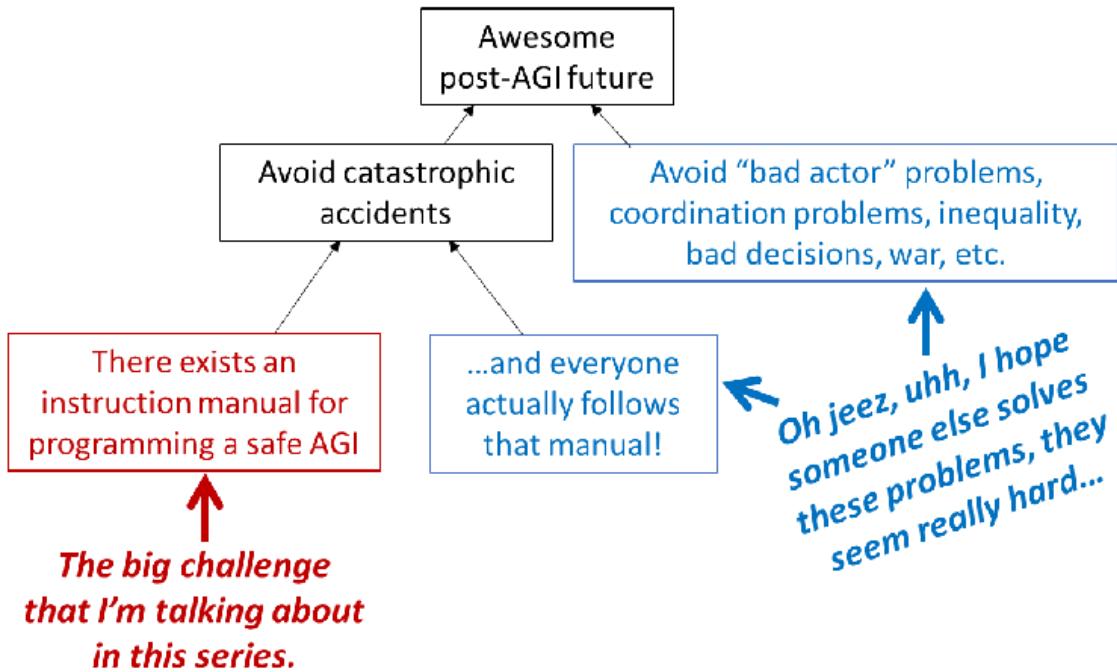
Summary of this first post:

- In Section 1.2, I define the “AGI technical safety problem”, put it in the context of other types of safety research (e.g. inventing passively-safe nuclear power plant designs), and relate it to the bigger picture of what it will take for AGI to realize its potential benefits to humanity.
- In Section 1.3, I define “brain-like AGI” as algorithms with big-picture similarity to key ingredients of human intelligence, presumably (though not necessarily) as a result of future people reverse-engineering those aspects of the human brain. What exactly that means will be clearer in future posts. I will also bring up the counterintuitive idea that “brain-like AGI” can (and probably will) have *radically nonhuman motivations*. I won’t explain that here, but I’ll finish that story by the end of [Post #3](#).
- In Section 1.4, I define the term “AGI”, as I’m using it in this series.
- In Section 1.5, I discuss the probability that people will eventually make brain-like AGIs, as opposed to some other kind of AGI (or just not invent AGI at all). The section includes seven popular opinions on this topic, from both neuroscientists and AI / machine learning experts, and my responses.
- In Section 1.6, I’ll address AGI accidents, which is something we should expect if we don’t solve the AGI technical safety problem. I’ll argue that these kinds of accidents can be catastrophic indeed, including human extinction. This topic is a *minefield* of confusion and miscommunication, and I will frame my discussion around responses to eight common objections.
- In Section 1.7, I’ll address the more specific question of why we should think about AGI safety *right now*. After all, there is a *prima facie* good case for waiting, namely: (1) AGI doesn’t exist yet, (2) AGI will exist someday in the future, and (3) it will be easier to do AGI safety research when we know more about the AGI, and easier still when we actually have AGI code that we can run tests on. There is indeed *something* to that argument, but I’ll argue that there is nevertheless a lot of safety work that can and must be done ASAP.
- In Section 1.8, I’ll suggest that brain-like-AGI safety is a fun, fascinating, and fruitful topic, even if you don’t buy the idea that it’s important for the future.

## 1.2 The AGI technical safety problem

AGI is short for “Artificial General Intelligence”—I’ll get back to the definition of AGI in Section 1.4 below. AGI doesn’t exist right now, but I’ll argue in Section 1.7 that we can and should be preparing for AGI even today.

The part I’ll be talking about in [this series](#) is the red box here:



Specifically, we zoom in on a single team of humans who are trying to create a single AGI, and we want it to be possible for them to do so without winding up with some catastrophe that nobody wanted, with an out-of-control AGI self-replicating around the internet or whatever (more on which in Section 1.6).

Blue boxes in this diagram are things that I won't talk about in this series. In fact, I'm not working on them at all—I have enough on my plate already. But I very strongly endorse other people working on them. If you, dear reader, want to work on them, *godspeed!!* I'm cheering you on! And here are a few links to get you started: [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [7](#).

Back to the red box. This is a technical problem, calling for a technical solution. Nobody wants catastrophic accidents. And yet!! Indeed, it's entirely possible for people to write an algorithm that does something that nobody wanted it to do. It happens all the time! **We might call it “a bug” when it’s a local problem in the code, and we might call it “a fundamentally flawed software design” when it’s a global problem.** I'll argue later in the series that AGI code may be unusually prone to catastrophic accidents, and that the stakes are very high (see Section 1.6 below, and [Post #10](#)).

Here's an analogy. If you're building a nuclear power plant, nobody wants an out-of-control chain reaction. The people at Chernobyl certainly didn't! But it happened anyway! I take a few lessons from this analogy:

- Enrico Fermi invented a technical solution for controlling nuclear chain reactions—control rods—before starting to build the first-ever nuclear chain reaction. Right on!! That's doing things in the right order! By the same token, I suggest that we should strive to have a technical solution to avoiding catastrophic AGI accidents ready to go before people start programming AGIs. In fact, I'll argue below for something even stronger than that: knowing the solution (even vaguely) 10 years before AGI is even better; 20 years before AGI is better still; etc. etc. This claim is not obvious, but I'll get back to it (Section 1.7).
- Technical solutions aren't all-or-nothing. Some reduce the chance of accidents without eliminating them. Some are complicated and expensive and error-prone to implement. In the nuclear case, control rods reduce accident risk a lot, but [passively-safe reactors](#) reduce it even further. By the same token, I expect that technical AGI safety will be a

rich field, where we'll develop better and better approaches over time, involving multiple techniques and multiple layers of protection. At least, I hope! As I'll discuss later in the series, I claim that right now we have no solution at all—not even vaguely. We have our work cut out!

- The blue boxes (see diagram above) also exist, and are absolutely essential, even if they're out-of-scope for this particular series of articles. The cause of the Chernobyl accident was *not* that nobody knew how to keep a nuclear chain reaction under control, but rather because best practices were not followed. In that case, all bets are off! Still, although we on the technical side can't *solve* this noncompliance problem by ourselves, we can *help on the margin*, by developing best practices that are maximally idiot-proof, and minimally expensive.

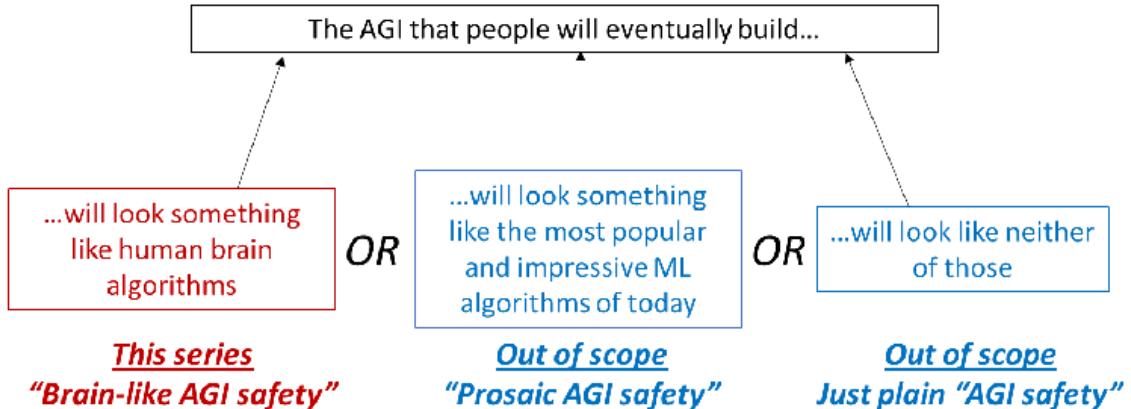


In *The Sorcerer's Apprentice*, if I'm remembering it right, software engineer Mickey Mouse programs an AGI with a broomstick-like robotic body. The AGI proceeds to do the exact thing that Mickey *programmed* it to do ("fill a bucket with water"), but that turns out to be very different from what Mickey *wanted* it to do ("fill a bucket with water, without making a mess or doing anything else that I would find problematic, etc.") Our goal is to empower software engineers like Mickey with the technical tools and knowledge they need to avoid these kinds of accidents. See [this talk by Nate Soares](#) for a deep-dive into why Mickey has his work cut out.

## 1.3 Brain-like AGI

### 1.3.1 Overview

This series will focus on a particular scenario for what AGI algorithms will look like:



The red box is what I'll talk about here. The blue boxes are things that are out-of-scope for this series.

You may have opinions about which of these categories is more or less likely, or impossible, or whether this breakdown is even sensible. I have opinions about those things too! I'll discuss them later (Section 1.5). My main opinion is that *all three* of these are sufficiently likely that we should be "contingency planning" for them. So while I personally don't do too much work on the blue boxes, I'm sure glad that other people do!

Here's an analogy. If someone in 1870 were guessing what future human flight would look like...

- "Kinda like birds" would have been a reasonable guess...
- "Kinda like today's best airships" would also have been a reasonable guess...
- "Neither of the above" would have been a reasonable guess too!!

In this particular imaginary case, all three of those guesses would have turned out correct in some ways and wrong in other ways: The Wright Brothers were directly and extensively inspired by large soaring birds, but left out the wing-flapping part. They also used some components found on airships (e.g. propellers), as well as plenty of original ingredients. That's just one example, but I think it's suggestive.

### 1.3.2 What exactly is “brain-like AGI”?

When I say “brain-like AGI”, I have a particular thing in mind. This thing will become much clearer in the subsequent posts, after we've started diving into neuroscience. But here's what I'm going for in broad strokes.

There are ingredients in the human brain and its environment that lead to humans having general intelligence (i.e., common sense, ability to figure things out, etc.—see Section 1.4 below). The scenario I have in mind is: Researchers will figure out what those ingredients in the brain are, and how they work, and then they will write AI code based on those same key ingredients.

To clarify:

- I don't expect that “brain-like AGI” will include every part of the brain and its environment. For example, there are highly-intelligent people who were born without a sense of smell, which suggests that brain olfactory processing circuitry probably isn't essential for AGI. There are [highly-intelligent people who were quadriplegic from birth](#),

suggesting that lots of spinal-cord circuitry and (certain aspects of) "embodiment" aren't essential either. There are likewise [people born without a cerebellum](#) who are nevertheless well within the range of normal adult human intelligence (able to hold down a job, live independently, etc.—the kinds of capabilities that we would unhesitatingly call "AGI"). Other adults are holding down jobs while [missing an entire brain hemisphere](#), etc. etc. My default expectation is that AGI will be created by people trying to create AGI, and they'll leave out whatever components they can, to make their jobs easier. (I'm not endorsing that as necessarily a good idea, just saying what I expect by default. More on this in [Post #3](#).)

- In particular, the kind of "brain-like AGI" I'm talking about is *definitely* not the same as Whole Brain Emulation.
- I don't require that "brain-like AGI" will resemble the human brain in low-level details, like with spiking neurons, dendrites, etc., or direct simulations thereof. If the resemblance is only at a higher level of abstraction, that's fine, it won't affect anything here.
- I don't require that "brain-like AGI" will be invented by a process of reverse-engineering the brain. If AI researchers happen to *independently* reinvent brain-like algorithms—just because those algorithms are good ideas—well then I still count it as brain-like algorithms.
- I don't require that "brain-like AGI" will be designed in a way that resembles how the brain was designed, i.e. evolutionary search. Quite the contrary: My working assumption is that it will be designed by humans in a way that's akin to a typical machine learning project today: [lots of human-written code](#) (loosely analogous to the genome), a *subset* of which defines the inference and update rules of one or more learning algorithms (corresponding to the brain's *within-lifetime* learning algorithms). There may be a few blank spaces in the code that get filled in by hyperparameter search or neural architecture search etc. Then you run the code, and the learning algorithms gradually build up a big complicated trained model from experience, maybe with trillions of adjustable parameters. Much more on this stuff in the next two posts and [Post #8](#).
- I don't require that "brain-like AGI" will be conscious (in the [phenomenal](#) sense). There are *ethical* reasons to care about whether AGI is conscious (more on which in [Post #12](#)), but nothing I say in this series will depend on whether or not the AGI is conscious. Machine consciousness is a big contentious topic and I just don't want to get into it here. (I've written about it a bit [elsewhere](#).)

Maybe a more practical way of saying it is: I'm going to make a bunch of claims about the algorithms underlying human intelligence, and then talk about safely using algorithms with those properties. If our future AGI algorithms have those properties, then this series will be useful, and I would be inclined to call such an algorithm "brain-like". We'll see exactly what those algorithm properties are, going forward.

### **1.3.3 “Brain-like AGI” (by my definition) can (and quite possibly will) have *radically nonhuman motivations***

I'm going to talk about this a *lot* more in later articles, but this is *such* an important point that I want to bring it up immediately.

Yes I know it sounds weird.

Yes I know you think I'm nuts.

But please, I beg you, hear me out first. By the time we get to [Post #3](#), well *then* you can decide whether or not to believe me.

In fact, I'll go further. I'll argue that "radically nonhuman motivations" is not just *possible* for a brain-like AGI, but is *my baseline expectation* for a brain-like AGI. I'll argue that this is generally a bad thing, and that we should consider prioritizing certain lines of R&D in a proactive effort to avoid that.

(To be clear, "radically nonhuman motivations" is not synonymous with "scary and dangerous motivations". Unfortunately, "scary and dangerous motivations" is *also* my baseline expectation for a brain-like AGI!! But that requires a further argument, and you'll have to wait until [Post #10](#) for that one.)

## 1.4 What exactly is "AGI"?

A frequent point of confusion is the word "General" in "Artificial General Intelligence":

- **The word "General" DOES mean "not specific",** as in "In general, Boston is a nice place to live."
- **The word "General" DOES NOT mean "universal",** as in "I have a general proof of the math theorem."

An AGI is *not* "general" in the latter sense. It is *not* a thing that can instantly find every pattern and solve every problem. Humans can't do that either! In fact, *no* algorithm can, because that's fundamentally impossible. Instead, an AGI is a thing that, when faced with a difficult problem, might be able to solve the problem easily, but if not, maybe it can build a tool to solve the problem, or it can find a clever way to avoid the problem altogether, etc. For our purposes here, think of AGI as an algorithm which can "figure things out" and "understand what's going on" and "get things done", including using language and science and technology, in a way that's reminiscent of how most adult humans can do those things, but toddlers and chimpanzees and GPT-3 can't. Of course, AGI algorithms may well be subhuman in some respects and superhuman in other respects.

Anyway, this series is about brain-like algorithms. These algorithms are by definition capable of doing absolutely every intelligent behavior that humans can do, and potentially much more. So they can *definitely* reach AGI. Whereas today's AI algorithms are *not* AGI. So somewhere in between here and there, there's a fuzzy line that separates "AGI" from "not AGI". Where exactly is that line? My answer: I don't know, and I don't care. Drawing that line has never come up for me as a useful thing to do. It won't come up in this series either.

## 1.5 What's the probability that we'll eventually wind up with brain-like AGI?

Above (Section 1.3.1) I suggested three categories of AGI algorithms: "brain-like" (as defined just above), "prosaic" (i.e. like today's most impressive deep neural net ML algorithms), and "other".

If our attitude is "Yes, let's do safety research for all three possibilities, just in case!!"—as I claim it should be—then I guess it's not all *that* decision-relevant what probability weights we put on each of the three things.

But even if it's irrelevant, it's fun to talk about, so what the heck, I'll just quickly summarize and respond to some popular opinions I've heard on this topic.

**Opinion #1:** "I dispute the premise: human brains work by basically the same principles as today's popular ML algorithms."

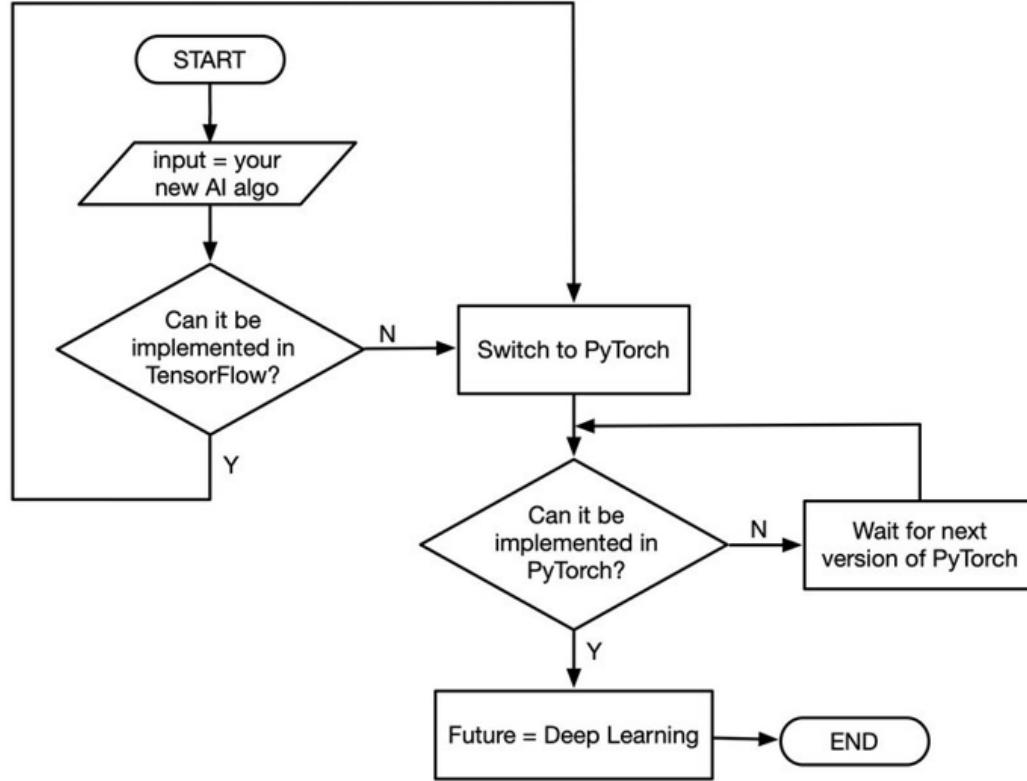
- The thing is, “today’s popular ML algorithms” is a big tent including lots of different algorithms. For example, I see *hardly any* overlap between “[GPT-3](#)-like-AGI safety” and “brain-like-AGI safety”, whereas I see *substantial* overlap between “[model-based RL](#) AGI safety” and “brain-like-AGI safety”.
- Anyway, by assuming “brain-like AGI”, I get the right to make certain assumptions about the cognitive architecture, representations, learning algorithms, and so on.
  - Some of these “brain-like AGI ingredients” are *universal* parts of today’s popular ML algorithms (e.g. learning algorithms; distributed representations).
  - Others of these “brain-like AGI ingredients” are (individually) present in a *subset* of today’s popular ML algorithms but absent from others (e.g. reinforcement learning; predictive [a.k.a. self-supervised] learning; explicit planning).
  - Still others of these “brain-like AGI ingredients” seem mostly or totally absent from today’s most popular ML algorithms (e.g. ability to form “thoughts” [e.g. “I’m going to the store”] that blend together immediate actions, short-term predictions, long-term predictions, and flexible hierarchical plans, inside a generative world-model that supports causal and counterfactual and metacognitive reasoning).
- So in this sense, “brain-like AGI” is a specific thing that might or might not happen, independently of “prosaic AGI”. Much more on “brain-like AGI”, or at least its safety-relevant aspects, in the subsequent posts.

**Opinion #2:** “Brain-like AGI is possible but Prosaic AGI is not. It just ain’t gonna happen. Today’s ML research is not a path to AGI, just as climbing a tree is not a path to the moon.”

- I find this to be a moderately popular opinion among neuroscientists and cognitive psychologists. Prominent advocates of this view include [Gary Marcus](#) and [Melanie Mitchell](#).
- One question is: if we take one of today’s most popular ML models, add no additional significant insights or architectural changes whatsoever, and just scale the model to ever larger sizes, do we get AGI? I join those neuroscientists in expecting the answer to be “probably not”.
- On the other hand, even if it turns out that deep neural networks can’t do important-for-intelligence things X and Y and Z, well c’mon, somebody’s probably just gonna glue together a deep neural network with other components that do X and Y and Z. And then we can have some pointless semantic debate about whether it’s still “really” prosaic AGI or not.

## ARTIFICIAL GENERAL FUTURE

Is Deep Learning the future of AI?  
Use this flow chart to find out!



[Image credit: Dileep George](#)

- Anyway, in this series, I will be assuming that AGI will have certain algorithmic features (e.g. [online learning](#), a certain type of model-based planning, etc.—much more in later posts). I'll be assuming that because (1) those features are part of human intelligence, (2) they seem to be there for a good reason. My safety-related discussions will rely on those features being present. Can algorithms with these features be implemented in PyTorch on a GPU? Well, I don't really care.

**Opinion #3:** “Prosaic AGI is going to happen so soon that no other research program has a chance.”

- A subset of people in ML believe this. I don’t. Or at any rate, I would be awfully surprised.
- I do agree that *IF* prosaic AGI is, say, 5 years away, then we almost certainly don’t need to think about brain-like AGI or indeed any other research program. I just think that’s an awfully big “if”.

**Opinion #4:** “Brains are *SO* complicated—and we understand them *SO* little after *SO* much effort—that there’s just no way we’ll get brain-like AGI even in the next 100 years.”

- This is a pretty popular opinion, both inside and outside of neuroscience. I think it’s very wrong, and will be arguing against it at length in the next two posts of the series.

**Opinion #5:** “Neuroscientists aren’t trying to invent AGI, so we shouldn’t expect them to succeed.”

- There’s *some* truth to this, but I mostly disagree. For one thing, a number of leading computational neuroscientists ([the DeepMind neuroscience team](#), [Randall O’Reilly](#), [Jeff Hawkins](#), [Dileep George](#)) are in fact explicitly trying to invent AGI. For another thing, people *in AI*, including prominent leaders of the field, try to keep up with the neuroscience literature and incorporate its ideas. And anyway, “understanding an AGI-relevant brain algorithm” is *part* of inventing brain-like AGI, whether or not that’s the intention of the person carrying out the research.

**Opinion #6:** “Brain-like AGI is kinda an incoherent concept; intelligence requires embodiment, not just a brain in a vat (or on a chip).”

- The “embodiment” debate in neuroscience continues to rage. I fall somewhere in the middle. I do think that future AGIs will have *some* action space—e.g., the ability to (virtually) summon a particular book and open it to a particular passage. I *don’t* think having a whole literal body is important—for example [Christopher Nolan](#) (1965-2009) had lifelong quadriplegia, but it didn’t prevent him from being an acclaimed author and poet. More importantly, I expect that whatever aspects of embodiment are important for intelligence could be easily incorporated into a brain-like AGI running on a silicon chip. Is a body necessary for intelligence after all? OK sure, we can give the AGI a virtual body in a VR world. Are hormonal signals necessary for intelligence? OK sure, we can code up some virtual hormonal signals. Etc. etc.

**Opinion #7:** “Brain-like AGI is incompatible with conventional silicon chips; it requires a whole new hardware platform based on spiking neurons, active dendrites, etc. Neurons are just plain better at computation than silicon chips are—just look at their energy efficiency etc.”

- I’m really unsympathetic to this position. Conventional silicon chips can definitely simulate biological neurons—[neuroscientists do this all the time](#). Conventional silicon chips can also presumably implement “brain-like algorithms” using different low-level operations more suited to that hardware, just as the same C code can be compiled to different CPU instruction sets. As for “neurons are just plain better”, I freely acknowledge the human brain does a *crazy impressive* amount of computation given its tiny volume, mass, and power consumption. But those are not hard constraints! If a silicon-chip AGI server were literally 10,000× the volume, 10,000× the mass, and 1000× the power consumption of a human brain, with comparable performance, I don’t think anyone would be particularly bothered—in particular, its electricity costs would *still* be well below my local minimum wage!! And [my best estimate](#) is that buying enough silicon chips for human-brain-human-lifetime-level computation is probably easily feasible, or will be in the next decade, even for small companies. The key reason that small companies aren’t building AGIs today is that we don’t know the right algorithms.

This is just a quick run-through; each of these opinions could be a whole article—heck, a whole book. For my part, I put >50% probability that we’ll have a *sufficiently*-brain-like AGI that this series will be very relevant. But who knows, really.

## 1.6 Why are AGI accidents such a big deal?

Two reasons: (1) the stakes are high, and (2) the problem is hard. I’ll be talking about (2) much more later in the series (Posts [#10-#11](#)). Let’s talk about (1).

And let's talk more specifically about one high-stakes possibility: the risk of human extinction. That sounds a bit wild but hear me out.

I'll frame this discussion as answers to popular objections:

**Objection #1: The only way that an out-of-control AGI could result in human extinction is if the AGI invents crazy sci-fi superweapons, e.g. [gray goo](#). As if such a thing is even possible!**

Oh, if only that were true! But alas, I don't think sci-fi superweapons are necessary. In fact, it seems to me that it's maybe borderline possible for a *human* intelligence using *existing* technology to cause human extinction!

Think about it: it's already at least borderline-possible today for an ambitious intelligent charismatic methodical human to arrange for the manufacture and release of a novel contagious disease that's 100x deadlier than COVID-19. Heck, it's probably possible to release 30 such plagues all at once! Meanwhile, I figure it's at least borderline-possible today for an ambitious intelligent charismatic methodical human to find a way to manipulate nuclear early warning systems (trick them, hack into them, bribe or threaten their operators, etc.), setting off an all-out nuclear war, killing billions of people and sowing chaos around the world. Those are just two things; creative readers will immediately think of lots more. I mean seriously, there are fiction books with *totally plausible* mad-scientist apocalypse scenarios—not just according to me, but according to domain experts.

Now, granted, human extinction seems like a high bar! People live in all kinds of places, including small tropical islands that would be insulated from both nuclear winter and plagues. But this is where we get a big difference between an *intelligent* agent like an AGI, versus an *unintelligent* agent like a virus. Both can self-replicate. Both can kill lots of people. But an AGI, unlike a virus, can *take command of military drones, and mow down the survivors!!*

So my hunch is that we're all still around today thanks in large part to the fact that all the most ambitious intelligent charismatic methodical humans *aren't trying to kill everyone*—and not because “killing everyone” is a thing that requires crazy sci-fi superweapons.

As discussed above, one of the failure modes I have in mind would involve out-of-control AGIs that combine (at least) human-like intelligence with *radically nonhuman motivations*. This would be a new situation for the world, and I don't find it comforting!

You might reply: The thing that went wrong in this scenario is *not* the out-of-control AGI, it's the fact that humanity is too vulnerable! And my response is: *Why can't it be both?* So in my book: Yes we should absolutely [make humanity more robust to bio-engineered pandemics](#), and [reduce the chances of nuclear war](#), etc. etc. All these things are great ideas that I strongly endorse, and *godspeed* if you yourself are working on them. But at the same time, we should *also* work really hard to *not create out-of-control self-replicating human-like intelligences with radically nonhuman motivations!*

...Oh and one more thing: Maybe “crazy sci-fi superweapons like gray goo” are possible too! Beats me! If so, we need to be even *more* cautious!

**Objection #2: The only way that an AGI accident could result in human extinction is if the AGI is somehow smarter than all humans combined .**

The issue here is that “all humans combined” may not know that they are engaged in a battle against an AGI. Maybe they would, maybe they wouldn't. If the AGI is at all competent at secrecy, it could presumably set up a surprise attack, without anyone knowing what was happening until it was too late. Or if the AGI is at all competent at disinformation and propaganda, it could presumably pass off its actions as accidents, or as (human) enemy

action. Maybe everybody would be blaming everyone else, and nobody would know what's going on.

**Objection #3: The only way that an AGI accident could result in human extinction is if the AGI is deliberately given access to levers of power, like nuclear codes, control over social media, etc. By the same token, we can run the AGI code on just one server, and then switch it off if anything goes wrong.**

The problem here is that intelligent agents can turn "few resources" into "lots of resources". Think of Warren Buffett, or Adolf Hitler.

Intelligent agents can earn money (whether legally or not), and they can earn trust (whether deserved or not), and they can get access to other computers (whether by purchasing server time or by hacking). The latter is especially important because an AGI—like a virus, but *not* like a human—can potentially self-replicate. Self-replication is one way it can protect itself from shutdown, if it's motivated to do so. Another way is by tricking / misleading / winning over / bribing / outsmarting whoever controls the shutdown switch.

(A kernel of truth here is that if we're unsure of an AGI's motivations and competence, then *giving it access to the nuclear codes is a very bad idea!* Trying to limit an AGI's power and resources doesn't seem to be a solution to any of the *hardest* problems that we're interested in here, but it can still be helpful on the margin, like as an "additional layer of protection". So I'm all for it.)

**Objection #4: The good AGIs can stop the bad out-of-control AGIs.**

For one thing, if we don't solve the technical problem of how to steer an AGI's motivation and keep it under control (see Posts [#10–#15](#)), then there may be a period of time when *there are no good AGIs!* Instead, *all* the AGIs are out-of-control!

For another thing, out-of-control AGIs will have asymmetric advantages over good AGIs—like the ability to steal resources, to manipulate people and institutions via lying and disinformation; to cause wars, pandemics, blackouts, famines, gray goo, and so on; and to not have to deal with coordination challenges across different (human) actors with different beliefs and goals. More on this topic [here](#).

**Objection #5: An AGI that's trying to kill everyone is a *really specific* kind of failure mode! There's just no reason that an AGI would try to do that. It's *not* the kind of thing that would happen as a general result of buggy or poorly-designed AGI software. It's the kind of thing that would *only* happen if somebody went out of their way to put malign motivations into the AGI. As a matter of fact, buggy or poorly-designed software tends to do, well, nothing in particular! I happen to know a thing or two about buggy software—in fact I just created some this morning. The only thing it murdered was my self-confidence!**

A kernel of truth here is that *some* bugs or design flaws in AGI code will indeed manifest as software that is *not* an AGI, and not "intelligent", and probably not even functional! Such errors do not qualify as catastrophic accidents, unless we were foolish enough to put that software in charge of the nuclear arsenal. (See "Objection #3" above.)

However, I claim that *other* bugs / design errors *will* in fact possibly lead to the AGI deliberately killing everyone, even if the AGI designers are reasonable people with noble, humble intentions.

Why? In the AGI safety lore, the classic way to justify this claim is the trifecta of (1) "The Orthogonality Thesis", (2) "Goodhart's law", and (3) "Instrumental Convergence". You can get the short version of this three-part argument in [this talk](#). For the long version, read on: this series is all about the nuts and bolts of motivation in brain-like AGIs, and how it can go awry.

So, hold that thought, and all will be clear by the time we get through [Post #10](#).

**Objection #6: If building AGIs seems to be a catastrophic-accident-prone endeavor, we'll just stop doing it, until when (and if) the problem is solved.**

My immediate reaction is to say: "We"? Who the heck is "we"? The AI community consists of many thousands of skilled researchers scattered across the globe. They disagree with each other about practically everything. There is no oversight on what they're doing. Some of them work at secret military labs. So I don't think we can take it for granted that "we" will not engage in research that *you and I* consider to be obviously ill-conceived and risky.

(Also, if some catastrophic accidents can be unrecoverable, then even one of those is too many.)

By the way, suppose someone says to me: "I have an extraordinarily ambitious plan, one that will require many years or decades of groundwork, but *if we succeed* then "Everyone on earth pauses AGI R&D until safety problems are resolved" will be on the table as a possible policy option in the future." OK sure, I would listen to that person with an open mind. They seem like they're at least understanding the scale of the challenge. Of course, I would expect them to probably fail. But what do I know?

**Objection #7: Accident risks have been going down and down, for decades. Didn't you read [Steven Pinker](#)? Have faith!**

Accident risks don't solve themselves. They get solved when people solve them. Planes generally don't crash because people have figured out how to avoid plane crashes. Nuclear power plants generally don't melt down because people have figured out how to avoid nuclear meltdowns.

Imagine if I said, "Good news, car accident death rates are lower than ever! So now we can get rid of seatbelts and crumple zones and road signs!" You would respond: "No!! That's insane!! Seatbelts and crumple zones and road signs are *the very reason* that car accident death rates are lower than ever!"

By the same token, if you're optimistic that we'll ultimately avoid AGI accidents, that's not a reason for you to be opposed to AGI safety research.

There's another thing to keep in mind before you take comfort in the historical record on technological accident risk: as technology gets inexorably more powerful, the scope of damage from technological accidents gets inexorably bigger as well. A nuclear bomb accident would be worse than a conventional bomb accident. A bioterrorist using 2022 technology would be able to do far more damage than a bioterrorist using 1980 technology. So by the same token, as AI systems get dramatically more powerful in the future, we should expect the scope of damage from AI accidents to grow dramatically as well. Thus the historical record here is not necessarily indicative of the future.

**Objection #8: Humans are doomed anyway. Oh well, whatever, no species lasts forever.**

I hear variants on this a lot. And granted, I can't *prove* that it's wrong. But the horseshoe crab has been around almost half a billion years, and counting. C'mon people, we can do this! Well at any rate, I'm not going down without a fight!

As for the people taking a "[far mode](#)" detached armchair-philosopher attitude to human extinction: If you would be devastated by the untimely death of your best friend or beloved family member ... but you're not particularly bothered by the idea of an out-of-control AGI killing everybody ... umm, I'm not sure what to say here. Maybe you're not thinking things through very carefully?

# 1.7 Why think about AGI safety now? Why not wait until we're closer to AGI and hence know more?

This is a common objection, and it indeed has a giant kernel of truth: Namely, that in the future, when we know more details about the eventual AGI design, there will be a lot of new technical safety work to do—work that we can't do right now.

However, there *is* safety work we can do right now. Just keep reading this series if you don't believe me!

I want to argue that the safety work that we *can* do right now, we *really should* do right now. Waiting would be much worse—even if AGI is still many decades away. Why's that? Three reasons:

**Reason 1 for feeling a sense of urgency:** *Early hints about safety can inform early R&D decisions—including via “Differential Technological Development”.*



The most important thing is that there's certainly more than one way to code an AGI algorithm.

Very early on in the process, we're making decisions about the big-picture path to AGI. We could do R&D towards one of many variations on “brain-like AGI” as defined here, versus whole brain emulation, versus various types of “prosaic AGI” (Section 1.3.1), versus [graph database query something-or-other AGI](#), versus [knowledge / discussion / reasoning systems](#), and we can proceed with or without brain-computer interfaces of various types, and so on. Probably not *all* of these research paths are feasible, but there's certainly more than one path towards more than one possible destination. We get to pick which one to go down.

Heck, we get to decide whether to build AGI in the first place! (However, see “Objection #6” above.)

In fact, we’re making these decisions already today. We’ve been making them for years. And our decision procedure is that lots of individuals around the world ask: What R&D direction is best for *me* right now? What gets *me* a job / promotion / profit / high-impact journal publication right now?

A better decision procedure would be: What kind of AGI do we eventually want to build? OK! Let’s try to make *that* one happen, sooner than all the inferior alternatives.

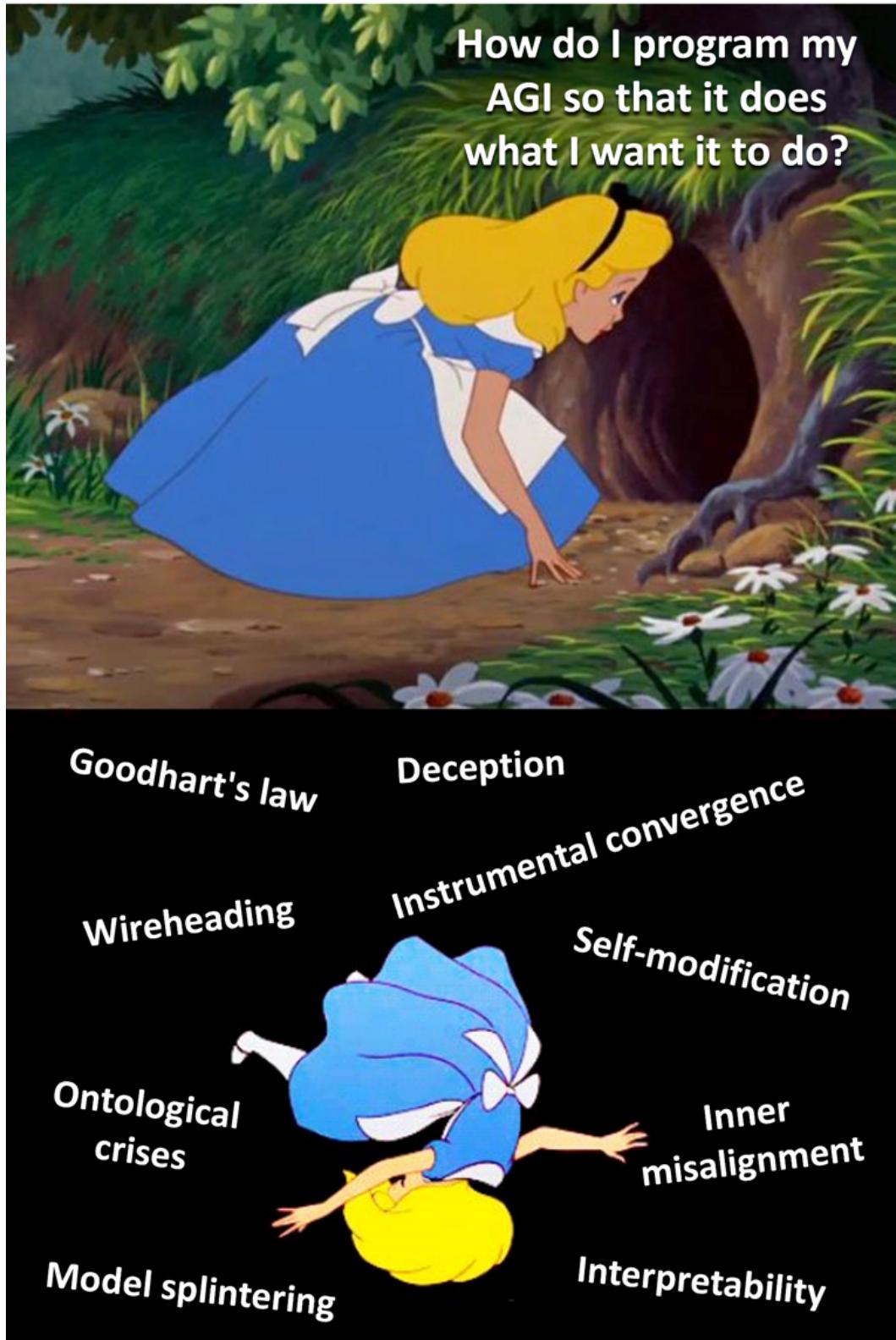
In other words, if someone chooses an R&D direction based on whatever looks interesting and promising, just like everyone else, well they’re not going to change our eventual technology development path. They’re just going to move us down the *same* path slightly *faster*. If we think that some destinations are better than others—say, if we’re trying to avoid a future full of out-of-control AGIs with radically nonhuman motivations—then it’s important to pick and choose what research you’re doing, in order to strategically accelerate the things that we most want to happen. This principle is called [\*\*differential technological development\*\*](#)—or more generally, [\*\*differential intellectual progress\*\*](#).



I have my own preliminary ideas about what should be accelerated for brain-like AGI to go better. (I’ll get to it much later in the series.) But **the main thing I believe is: “We should differentially accelerate work towards figuring out which work should be differentially accelerated”!! For example, would brain-like AGI be catastrophic-accident-prone or not? We have to figure it out! Hence this series!**

**Reason 2 for feeling a sense of urgency:** We don’t know how long safety research will take.

As discussed much more in later posts (especially Posts #10–#15), it is currently unknown how to make an AGI which is reliably trying to do the things that we want it to be trying to do. We don’t know how long it will take to figure it out (or prove that it’s impossible!). It seems prudent to start now.



As discussed later in the series (especially Posts [#10–#15](#)), AGI Safety seems to be a gnarly technical problem. We don't currently know how to solve it—in fact, we don't even

know if it's solvable. Thus, it seems wise to sharpen our pencils and get to work right now, rather than waiting until the last possible second. [Meme concept stolen from [here](#).]

In Stuart Russell's memorable analogy, imagine that we get a message from the aliens: "We are coming in our spaceships, and will arrive in 50 years. When we get there, we will radically transform your whole world beyond recognition." Indeed, we see their ships in our telescopes. They're inching closer each year. What do we do?

If we were to respond to the coming alien invasion the way we are *actually* today responding to AGI, we would collectively shrug and say "Meh, 50 years, I mean, that's *really* far away. We don't have to think about that *now*! If 100 people on Earth are trying to prepare for the looming alien invasion, that's *plenty*. Maybe too much! Y'know, if you ask me, those 100 people on Earth should stop looking up at the stars, and look around their own communities. Then they'd see that the *REAL* 'looming alien invasion' is *cardiovascular disease*. That's killing people *right now!*"

...You get the idea. (Not that I'm bitter or anything.)

**Reason 3 for feeling a sense of urgency:** *Building near-universal consensus about anything can be a horrifically slow process.*

Suppose I have a really good and correct argument that some AGI architecture or approach is just a terrible idea—that it's [unfixably unsafe](#). I publish the argument. Will everyone involved in AGI development, including those who have invested their career in that approach, immediately believe me, and change course? Probably not!!

That kind of thing *does* happen sometimes, especially in mature fields like math. But other ideas take many decades to become widely (let alone universally) accepted—famous examples include evolution and plate tectonics. It takes time for arguments to be refined. It takes time for evidence to be marshaled. It takes time for nice new pedagogical textbooks to be created. And yes, it takes time for the stubborn holdouts to die and be replaced by the next generation.

Why is near-universal consensus so important? See Section 1.2 above. Good ideas about how to build AGI are pointless if the people building AGI don't follow them. If we're going for voluntary compliance, then we need the AGI-builders to believe the ideas. If we're going for mandatory compliance, then we need the people with political power to believe the ideas. And we would *still* need AGI-builders to believe the ideas too, because perfect enforcement is a pipe dream (especially given secret labs etc.).

## 1.8 ...Plus it's a really fascinating problem!

Hey neuroscientists, listen. Some of you are trying to cure diseases. Good for you. Have at it. Others of you, well, you *say* you're trying to cure diseases on your NIH grant applications, but c'mon, that's not your *real* goal, and everyone knows it. You're *really* in it to solve fascinating unsolved problems. Well, let me tell you, brain-like-AGI safety is a fascinating unsolved problem!

It's even a rich source of insights *about neuroscience*! When I'm thinking all day about AGI safety stuff (wireheading, wishful thinking, symbol-grounding, ontological crises, interpretability, blah blah blah), I'm asking very different questions than most

neuroscientists, and thus finding different ideas. (...I'd like to think. Well, read on, and you can decide for yourself whether they're any good.)

So even if I haven't convinced you that the technical AGI safety problem is super duper important and impactful, read on anyway. You can also work on the problem because it's awesome. ;-)

# Future ML Systems Will Be Qualitatively Different

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

In 1972, the Nobel prize-winning physicist Philip Anderson wrote the essay "[More Is Different](#)". In it, he argues that quantitative changes can lead to qualitatively different and unexpected phenomena. While he focused on physics, one can find many examples of More is Different in other domains as well, including biology, economics, and computer science. Some examples of More is Different include:

- **Uranium.** With a bit of uranium, nothing special happens; with a large amount of uranium packed densely enough, you get a nuclear reaction.
- **DNA.** Given only small molecules such as calcium, you can't meaningfully encode useful information; given larger molecules such as DNA, you can encode a genome.
- **Water.** Individual water molecules aren't wet. Wetness only occurs due to the interaction forces between many water molecules interspersed throughout a fabric (or other material).
- **Traffic.** A few cars on the road are fine, but with too many you get a traffic jam. It could be that 10,000 cars could traverse a highway easily in 15 minutes, but 20,000 on the road at once could take over an hour.
- **Specialization.** Historically, in small populations, virtually everyone needed to farm or hunt to survive; in contrast, in larger and denser communities, enough food is produced for large fractions of the population to specialize in non-agricultural work.

While some of the examples, like uranium, correspond to a sharp transition, others like specialization are more continuous. I'll use **emergence** to refer to qualitative changes that arise from quantitative increases in scale, and **phase transitions** for cases where the change is sharp.

In this post, I'll argue that emergence often occurs in the field of AI, and that this should significantly affect our intuitions about the long-term development and deployment of AI systems. We should expect weird and surprising phenomena to emerge as we scale up systems. This presents opportunities, but also poses important risks.

## Emergent Shifts in the History of AI

There have already been several examples of quantitative differences leading to important qualitative changes in machine learning.

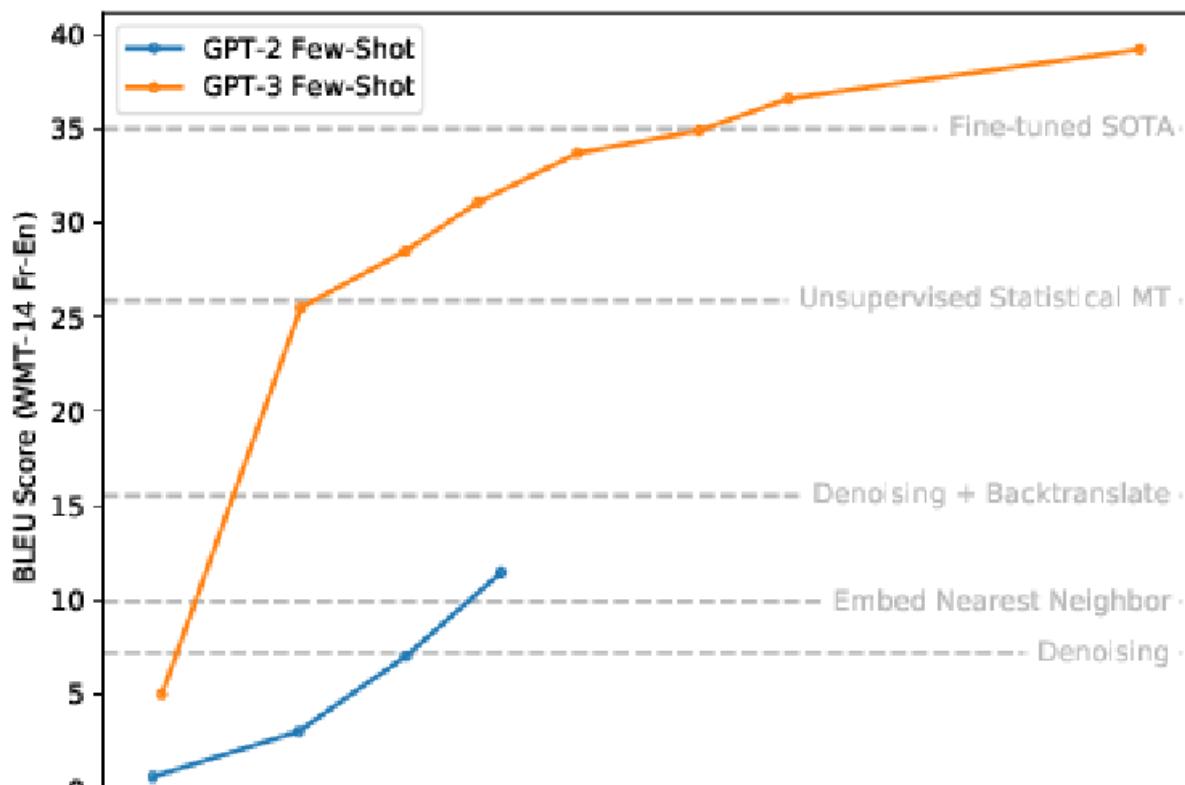
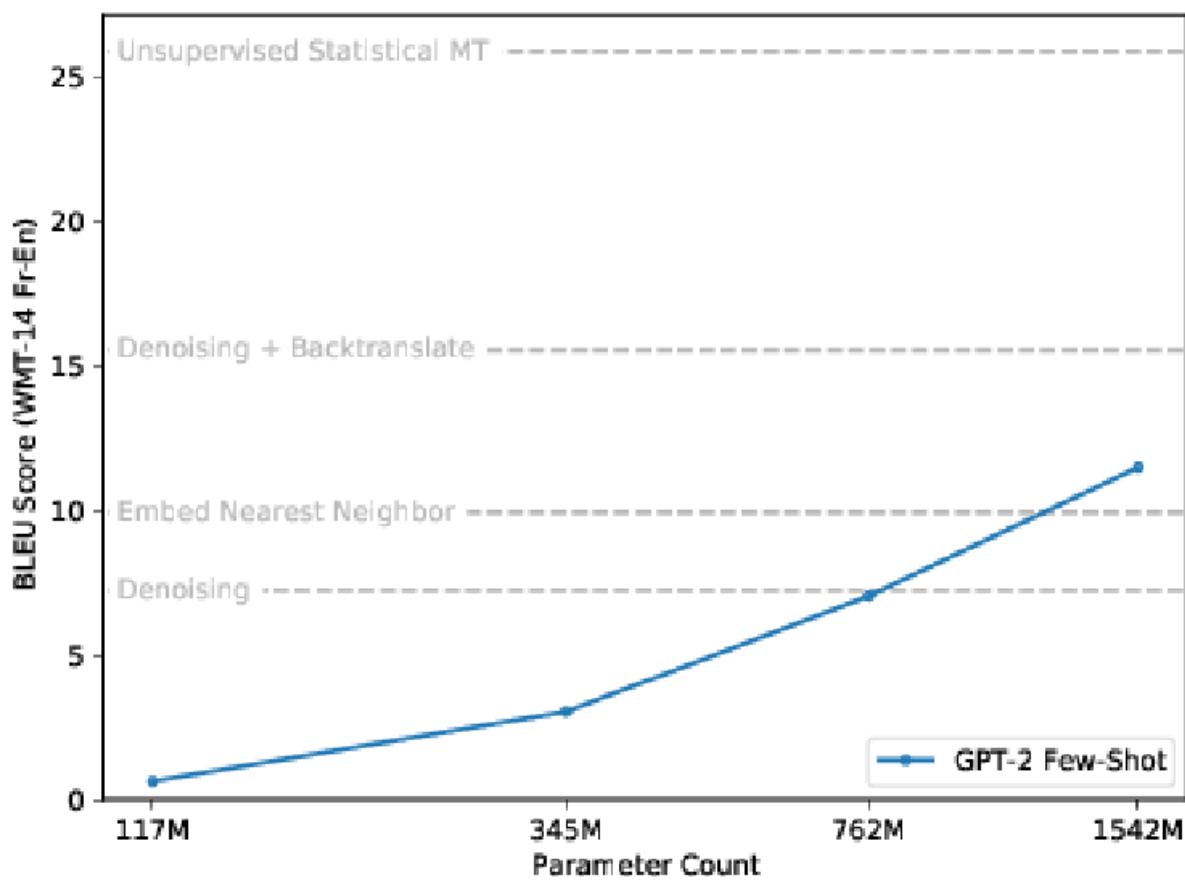
**Storage and Learning.** The emergence of machine learning as a viable approach to AI is itself an example of More Is Different. While learning had been discussed since the 1950s, it wasn't until the 80s-90s that it became a dominant paradigm: for instance, IBM's [first statistical translation model](#) was published in 1988, even though the idea [was proposed](#) in 1949<sup>[1]</sup>. Not coincidentally, 1GB of storage cost over \$100k in 1981 but only around \$9k in 1990 (adjusted to 2021 dollars). The [Hansard corpus](#) used to train IBM's model comprised 2.87 million sentences and would have been difficult to

use before the 80s. Even the simple MNIST dataset would have required \$4000 in hardware just to store in 1981, but that had fallen to a few dollars by 1998 when it was published. Cheaper hardware thus allowed for a qualitatively new approach to AI: in other words, More storage enabled Different approaches.

**Compute, Data, and Neural Networks.** As hardware improved, it became possible to train neural networks that were very deep for the first time. Better compute enabled bigger models trained for longer, and better storage enabled learning from more data; AlexNet-sized models and ImageNet-sized datasets wouldn't have been feasible for researchers to experiment with in 1990.

Deep learning performs well with lots of data and compute, but struggles at smaller scales. Without many resources, simpler algorithms tend to outperform it, but with sufficient resources it pulls far ahead of the pack. This reversal of fortune led to qualitative changes in the field. As one example, the field of machine translation moved from [phrase-based models](#) (hand-coded features, complex systems engineering) to [neural sequence-to-sequence models](#) (learned features, specialized architecture and initialization) to simply fine-tuning a [foundation model](#) such as BERT or GPT-3. Most work on phrase-based models was obviated by neural translation, and the same pattern held across many other language tasks, where hard-won domain-specific engineering effort was simply replaced by a general algorithm.

**Few-shot Learning.** More recently, [GPT-2](#) and [GPT-3](#) revealed the emergence of strong few-shot and zero-shot capabilities, via well-chosen natural language prompting.



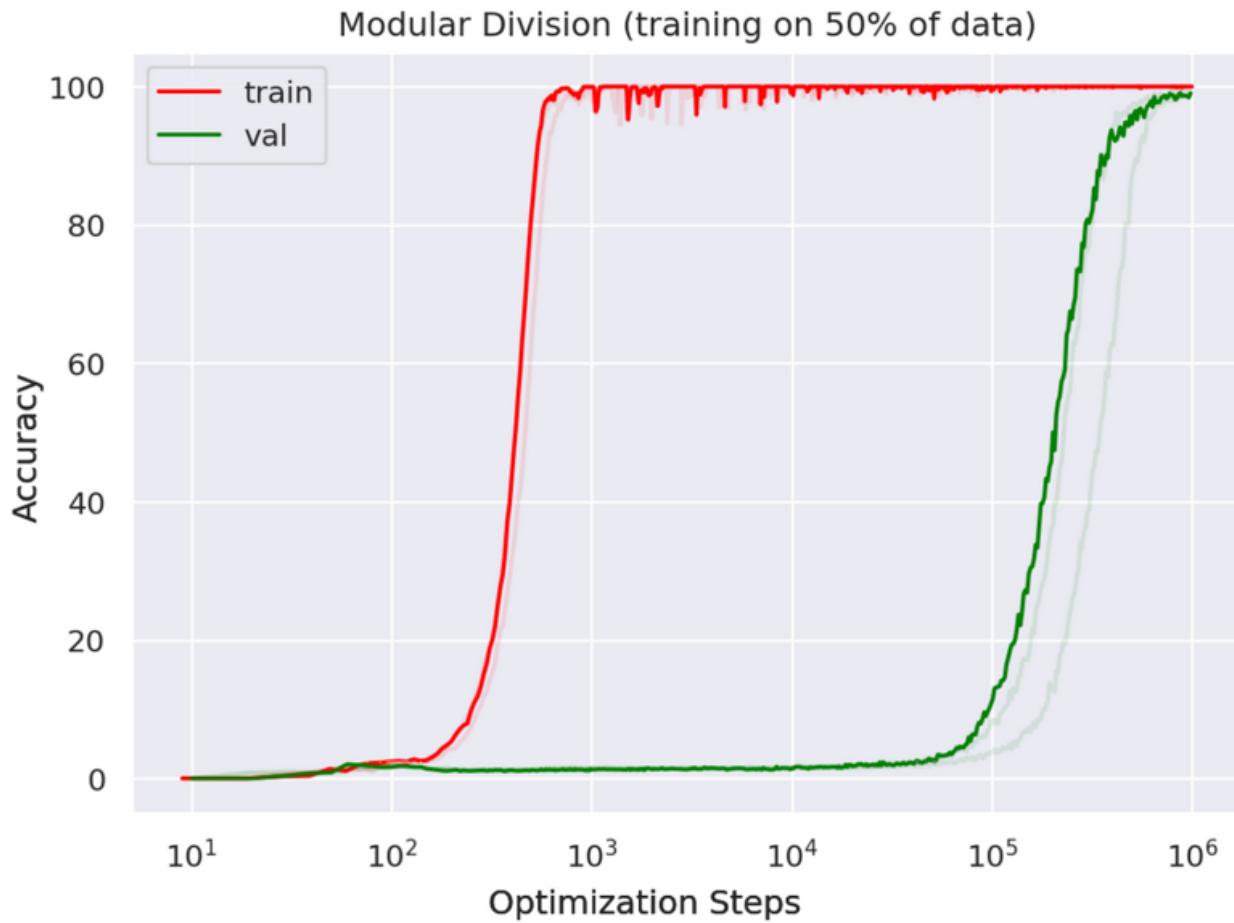


*Top: Few-shot machine translation performance (BLEU score) for GPT-2. Bottom: GPT-3 (trained on more data) has an even starker curve, going from 5 to 25 BLEU between 100M and 400M parameters. Unsupervised baselines, as well as fine-tuned state-of-the-art, are indicated for reference.*

This was an unexpected and qualitatively new phenomenon that only appeared at large scales, and it emerged without ever explicitly training models to have these few-shot capabilities. Comparing GPT-2 to GPT-3 shows that the exact model size needed can vary due to the training distribution or other factors, but this doesn't affect the basic point that new capabilities can appear without designing or training for them.

**Grokking.** In 2021, [Power et al.](#) identified a phenomenon they call "grokking", where a network's generalization behavior improves qualitatively when training it for longer (even though the training loss is already small).

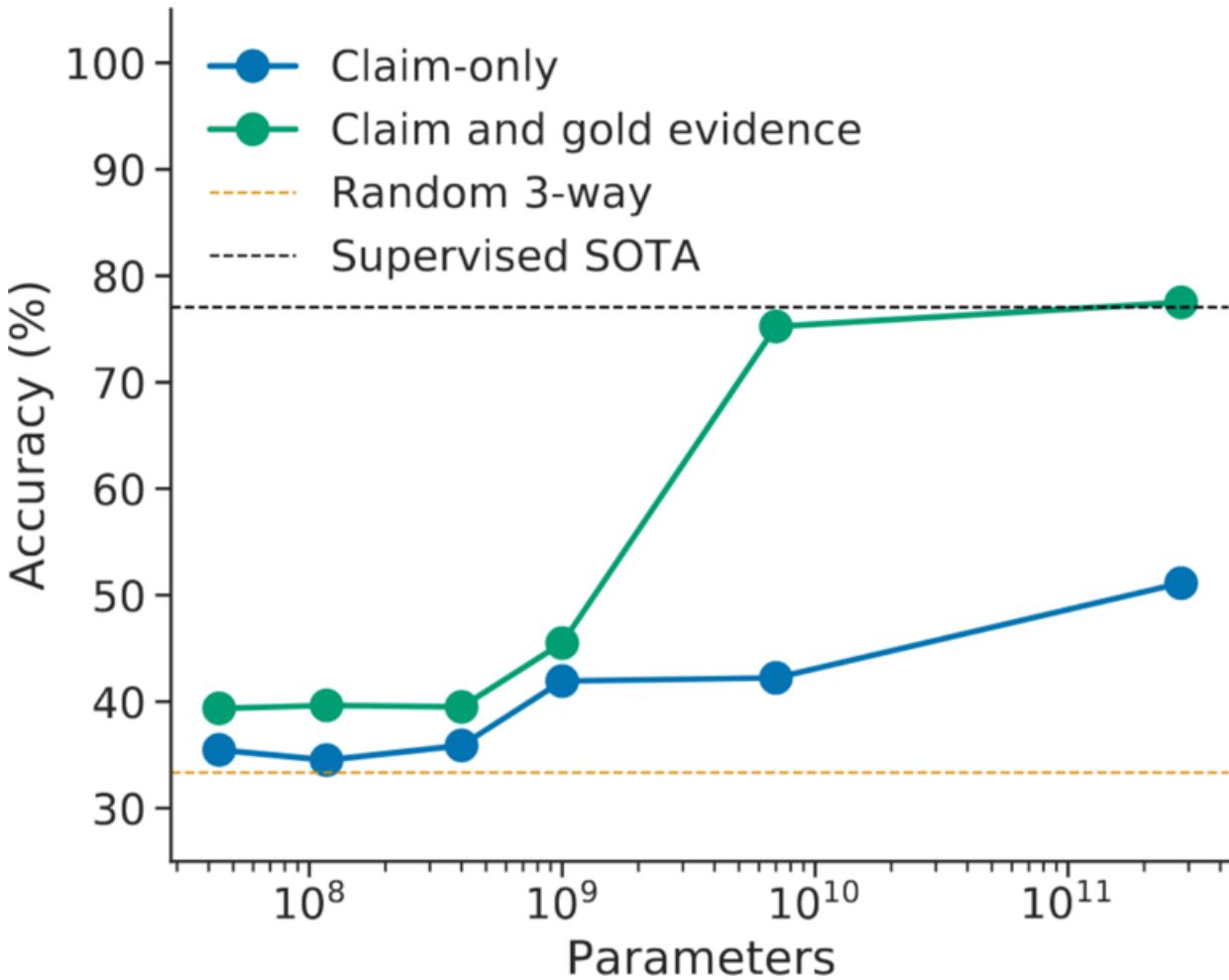
Specifically, for certain algorithmically generated logic/math datasets, neural networks trained for 1,000 steps achieve perfect train accuracy but near-zero test accuracy. However, after around 100,000 steps the test accuracy suddenly increases, achieving near-perfect generalization by 1 million steps.



This shows that even for a single model, we might encounter qualitative phase transitions as we train for longer.

**Other potential examples.** I'll briefly list other examples from recent papers. I don't think these examples are as individually clear-cut, but they collectively paint an interesting picture:

- [McGrath et al. \(2021\)](#) show that AlphaZero acquires many chess concepts at a phase transition near 32,000 training steps.
- [Pan et al. \(2021\)](#) show that reward hacking sometimes occurs via qualitative phase transitions as model size increases.
- DeepMind's recent [Gopher](#) model exhibits a phase transition on the FEVER task, acquiring the ability to utilize side information (Figure 3):



- [Wei et al. \(2021\)](#) show that instruction-tuning hurts small models but helps large models (see Figure 6).
- Some few-shot tasks such as arithmetic show phase transitions with model size (see [Brown et al. \(2020\)](#), Figure 3.10).
- [This](#) researcher shares an anecdote similar to the “grokking” paper.

# What This Implies for the Engineering Worldview

In the [introduction post](#) to this series, I contrasted two worldviews called Philosophy and Engineering. The Engineering worldview, which is favored by most ML researchers, tends to predict the future by looking at empirical trends and extrapolating them forward. I myself am [quite sympathetic to this view](#), and for this reason I find emergent behavior to be troubling and disorienting. Rather than expecting empirical trends to continue, emergence suggests we should often expect new qualitative behaviors that are not extrapolations of previous trends.

Indeed, in this sense Engineering (or at least pure trend extrapolation) is self-defeating as a tool for predicting the future<sup>[2]</sup>. The Engineering worldview wants to extrapolate trends, but one trend is that emergent behavior is becoming more and more common. Of the four phase transitions I gave above, the first (storage) occurred around 1995, and the second (compute) occurred around 2015. The last two occurred in 2020 and 2021. Based on past trends, we should expect future trends to break more and more often.<sup>[3]</sup>

How can we orient ourselves when thinking about the future of AI despite the probability of frequent deviations from past experience? I'll have a lot more to say about this in the next few posts, but to put some of my cards on the table:

- Confronting emergence will require adopting mindsets that are less familiar to most ML researchers and utilizing more of the Philosophy worldview (in tandem with Engineering and other worldviews).
- Future ML systems will have weird failure modes that don't manifest today, and we should start thinking about and addressing them in advance.
- On the other hand, I don't think that Engineering as a tool for predicting the future is entirely self-defeating. Despite emergent behavior, empirical findings often generalize surprisingly far, at least if we're careful in interpreting them. Utilizing this fact will be crucial to making concrete research progress.

- 
1. From the [IBM model authors](#): "In 1949 Warren Weaver suggested that the problem be attacked with statistical methods and ideas from information theory, an area which he, Claude Shannon, and others were developing at the time (Weaver 1949). Although researchers quickly abandoned this approach, advancing numerous theoretical objections, we believe that the true obstacles lay in the relative impotence of the available computers and the dearth of machine-readable text from which to gather the statistics vital to such an attack. Today, computers are five orders of magnitude faster than they were in 1950 and have hundreds of millions of bytes of storage. Large, machine-readable corpora are readily available." ↪
  2. This is in contrast to using Engineering to *build capable and impressive systems* today. If anything, recent developments have strongly solidified Engineering's dominance for this task. ↪
  3. This list is probably subject to selection bias and recency effects, although I predict that my point would still hold up for a carefully curated list (for instance, I didn't include the several ambiguous examples in my count). I would be happy to

bet on more phase transitions in the future if any readers wish to take the other side. [←](#)

# [New Feature] Support for Footnotes!

It is with great excitement<sup>[1]</sup> that I am pleased to announce that the main LessWrong text editor<sup>[2]</sup> now has support for footnotes!<sup>[3]</sup> A huge thanks to our friends over at the [Effective Altruism Forum](#) who coded this one up.

You can insert footnotes via:

1. Manually selecting text in the text box and selecting *insert footnote* from the footnotes menu icon.



The footnote icon is the [\*] on the right.

2. Using Markdown syntax

- Type [^n] where n is the number of the footnote you wish to insert.
- To insert a new footnote, use n that is <number of existing footnotes + 1>; to reuse an existing footnote, set n to be whichever footnote you are reusing.

**Footnotes will automatically renumber as you add and delete them!**

What's more, footnotes will render with hover-over previews once published:

It is with great excitement<sup>[1]</sup> that I am pleased to announce that the main LessWrong text editor<sup>[2]</sup> now has support for footnotes! I mean it, really. I've looked forward to us adding this support for years.<sup>[3]</sup>

You can insert footnotes via:

Behold! A footnote hover-preview!

That's it. Go forth and create scholarly works!

1. ^

I mean it, really. I've looked forward to us adding this support for years.

2. ^

That is, the **LW Docs editor**, as distinct from the Markdown editor and legacy Draft-JS editor.<sup>[4]</sup>

3. ^

The Markdown editor already had support for footnotes using [Markdown footnote syntax](#).

#### 4. ^

Yes, footnotes can have footnotes. And those footnotes can reference themselves.[\[4\]](#)

# How To Raise Others' Aspirations in 17 Easy Steps

*Title taken from Tyler Cowen: [The high-return activity of raising others' aspirations](#)*

*This is taken from a [twitter thread](#) (which I also wrote up on my [website](#)) that I was - aptly for the content - encouraged to put in non-Twitter places. Thanks for the encouragement!*

## **Things you can say to people in ten seconds that sometimes produce insanely outsized effects:**

1. Yeah, someone \*should\* do that. Why not you?
2. Is there something you could do about that problem in the next [five minutes](#)?
3. That's a great thought - have you written it up somewhere? I'd be excited to share it if so.
4. Should you write a book / blog?
5. You want to do that but don't think you will? Do you want to make a concrete plan now?
6. Do you want me to ask in a week / month if you've done that or how it's going?
7. Feeling stuck sucks. Want to brainstorm together?
8. Feeling stuck sucks. Have you spent a [five minute timer](#) generating options?
9. What's the twenty minute / minimum viable product version of this overwhelming-feeling thing?
10. Do you want me to sit with you while you fill out that application/write that email? Too bad, it's happening.
11. Is it worth just asking / cold emailing and seeing?
12. Is there anyone I know who you'd like to be introduced to?
13. Do you want to just set an alarm on your phone now as a reminder? (*from [Damon Sasi](#)*)
14. Do you know anyone else who might have struggled with or succeeded at that? Have you talked to them about it? (*from [Damon Sasi](#)*)
15. Who do you know who you could ask for help from?
16. You are Allowed (to try, to do, to build, to ask). I strongly recommend [this link of things you are allowed to do](#).
  - (Or do you [not need permission](#) in the first place?)
17. Do you sort of already know what you're going to do / have your mind made up about this?

# Lives of the Cambridge polymath geniuses

	Name	Main areas	Polymathy	War	Romance
	Bertrand Russell (1872-1970)	Mathematical logic, analytic philosophy	Popular science (relativity); popular writing on politics, ethics, philosophy; founded experimental school	WW1: Pacifist. Spoke up against the war. Fired by Cambridge and spent 6 months in a London prison.	Married four times and had numerous affairs. Wrote a book in 1929 arguing that affairs could be just. Had 3 children.
	John Maynard Keynes (1883 - 1946)	Economics, probability theory, economic policy	Investor; writer (intellectual biography); patron of the arts	WW1: Financial representative for British Treasury to the 1919 Versailles peace conference. WW2: Important role in setting up Bretton Woods system.	Bisexual. Kept a diary of dozens of hookups with men. Married a ballerina. No children.
	Ludwig Wittgenstein (1889 - 1951)	Analytic philosophy	Aeronautics: invented the tip jet propeller; designed a modernist townhouse; whistled symphonies;	WW1: Fought on the front for Austro-Hungarian army, won medals for bravery, captured and spent 9 months in Italian POW camp.	Bisexual but main relationships were with young men (e.g. David Hume Pinsent, relative of Hume). No children
	J.B.S. Haldane (1892 - 1964)	Population genetics, Mathematical biology	Popular science; futurism (IVF/cloning); read 11 languages; self-experimentation	WW1: Specialist in hand-bombing enemy trenches: killed many men with bombs and enjoyed it. Took part in Spanish Civil War.	Had an affair with married women, was fired by Cambridge but appealed and won. Married twice, second marriage to his grad student. No children.
	Joseph Needham (1900 - 1995)	Biochemistry, History of Chinese science and tech	Sinology; co-founder of UNESCO; Morris dance; polyglot	WW2: British diplomat sent to support Chinese scientists during the Sino-Japanese War.	Open marriage. He, his wife, and his partner lived on the same street (harmoniously). Gymnosophist. No children.
	Alan Turing (1912 - 1954)	CS (theory and applied), AI, philosophy	Codebreaking; morphogenesis	WW2: War hero for cryptanalysis at Bletchley Park (including Enigma codebreaking)	Gay. Was engaged to a woman but did not marry. Convicted of homosexual acts and forced to take hormonal treatment to reduce libido. No children.

[Link to full-size image](#)

## Delicious Facts

### Russell

- Grandfather was British prime minister
- Told Virginia Woolf his devotion to serious intellectual work came to an end when “my passions got hold of me”
- Met Lenin on a visit to Russia. This visit turned him against the Russian Revolution

### Wittgenstein

- His Austrian family was one of the richest in the world
- Was at same (obscure) school at same time as Hitler
- His family paid off the Nazis using their fortune to be classified as “mixed” rather than “full Jews” (and avoided the Holocaust)
- Three of his four brothers died by separate acts of suicide
- Tried to move to the Soviet Union to work as a laborer

### Haldane

- Aristocratic family, father was a prominent biologist
- Introduced the primordial soup theory of the origin of life
- For a period, he was a Stalinist and defended Lysenkoism on BBC radio
- Moved to India late in life and renounced British citizenship

### **Needham**

- As a professor of biochemistry, at age 37 he began an improbable pivot into Sinology when he fell in love with his Chinese grad student and started learning Chinese
- In China he befriended Zhou Enlai and met Mao
- He was part of a commission investigating whether the US had used biological weapons in the Korean War and was fooled into believing the US had

### **Turing**

- Would occasionally run 40 miles from Bletchley to London for meetings and tried out for British Olympic team
- Apparently he took fortune-telling seriously

## **Biographies**

- Russell: [Autobiography](#); [Monk](#)
- Keynes: [Skidelsky](#),
- Wittgenstein: [Monk](#)
- Haldane: [Subramanian](#)
- Needham: [Winchester](#)
- Turing: [Hodges](#)

Also see:

- Biopic [film](#) on Turing from the BBC
- Wittgenstein [movie](#), which includes Russell and Keynes as characters:
- [Logicomix](#): graphic novel feature Russell, Turing, and Wittgenstein (coauthored by computer scientist)

## **Why I found these figures interesting**

- They made exceptional and creative intellectual contributions (helping to found new fields). Turing's contributions seem most important.
- They had dramatic, full-bodied involvement in wars
- They spent significant periods working outside academia
- For their time, they had highly unconventional romantic lives and were eccentric in other ways
- Russell and Haldane were self-described [rationalists](#)

## **Interactions**

Russell acted as Wittgenstein's PhD supervisor but felt Wittgenstein surpassed him already as a student. Keynes invited Wittgenstein to join the Apostles and helped him get British citizenship during WW2. Turing attended Wittgenstein's lectures on the philosophy of mathematics. Needham succeeded Haldane as Reader in biochemistry at Cambridge.

Russell on Keynes:

Keynes's intellect was the sharpest and clearest that I have ever known. When I argued with him, I felt that I took my life in my hands, and I seldom emerged without feeling something of a fool.

Keynes on Russell and Wittgenstein:

The first impression conveyed by the work of Russell was that the field of formal logic was enormously extended. The gradual perfection of the formal treatment at the hands of himself, of Wittgenstein and of Ramsey had been, however, gradually to empty it of content and to reduce it more and more to mere dry bones, until finally it seemed to exclude not only all experience, but most of the principles, usually reckoned logical, of reasonable thought. Wittgenstein's solution was to regard everything else as a sort of inspired nonsense, having great value indeed for the individual, but incapable of being exactly discussed.

Wittgenstein on Russell:

Russell's books should be bound in two colours...those dealing with mathematical logic in red – and all students of philosophy should read them; those dealing with ethics and politics in blue – and no one should be allowed to read them.

Russell wrote a long essay ("[Icarus or The Future of Science](#)") responding to Haldane's seminal book "Daedalus".

Haldane was the [reviewer](#) for Turing's famous paper on morphogenesis. The review is prickly:

Before the paper is accepted, I consider that the whole mathematical part should be re-written. In the first place, some passages assuming ignorance in readers might be omitted without much loss. Secondly, much of the biology, e.g. pp 56-58, can be found in elementary textbooks, and often stated more accurately

...

1. Does it contain contributions to knowledge of sufficient scientific interest for the space required? **Yes**

2. Are any portions of the paper, or any illustrations, redundant? **Yes**

3. Should the paper be published by the Society? **Yes, subject to drastic emendations.**

...

7. Comments or criticisms which might enable the author to improve or correct his statement.

**I regret that my report, in the absence of figures or tables, must be insecurely grounded. I should be glad to discuss the paper with the author, but may be leaving for India shortly. I regard the central idea as being sufficiently important to warrant publication. I am equally clear that the paper should not be published as it stands.**

## **Third Time: a better way to work**

[Previously called Rational Breaks. See [comments](#) for name discussion]

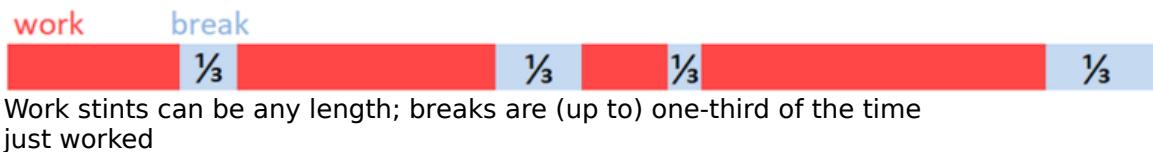


HOW CAN you be more productive? Instead of half-working all day, it's better to work in focused stints, with breaks in between to recover.

There are various ways to do this, but here's my new technique, called **Third Time**. The gist of it is:

- **Work for as long or as short as you like**, until you want or need to break; then
- **Break for up to one-third of the time** you've just worked.

So after 15 minutes of dealing with emails, you could stop for up to 5 minutes. After an hour-long meeting, you can take a good 20-minute break. And if a task bores you after 3 minutes, you can even break then—but only for 1 minute! Breaks reward you for working, but proper breaks have to be earned.



This kind of pattern is natural; [research](#) confirms that people tend to take longer breaks after working for longer. (One-third is just a recommendation; you can use other break fractions if you prefer.)

Third Time has many advantages over other techniques such as [Pomodoro](#) (which I'll discuss later), but the key one is flexibility. It adapts to your attention span, energy, and schedule, as well as to other people and events. And Third Time isn't just for your day-job—it suits anything that needs focus or effort, such as studying, practicing an instrument, personal admin, writing, or fitness training.

## Using Third Time

Here's an example of the basic procedure:

1. Note the time, or start a stopwatch
2. **Work** for as long or short as you like, until you want or need to break
3. Suppose you worked for 45 minutes. This earns you  $45 \div 3 = 15$  minutes off; so set an alarm for 15 minutes
4. **Break** until the alarm goes off
5. Go back to step 1.

## Breaks

You needn't take the full break. Maybe you have a tight deadline, an important customer calls, you're keen to resume work, or only have a short gap before a meeting. Whatever the reason, **if you end a break (say) 5 minutes early, add 5 minutes to your next break**. You don't lose the remaining time, it's just postponed:



You can shorten and postpone breaks like this, but don't let them overrun. Break time must be earned by working; it's like a debit card, not a credit card. So always set an alarm for the end of a break, and resume work as soon as it goes off (don't snooze it!)

**Take breaks whenever you like.** Surprisingly, *you'll still end up doing the same amount of work!* For example: instead of working for 45 minutes plus a 15-minute break, suppose you do just three minutes' work plus a one-minute break, and repeat that over and over again. You'll still be working three-quarters of the time; so in an hour, you'd still get 45 minutes' work done.

That said, it's annoying to end a break after just one minute, and keep switching back and forth. So even with tedious tasks, you'll soon find yourself working longer—and perhaps even enjoying it—to earn a decent break.

While breaking too often is counterproductive, so is breaking too seldom, such as when immersed in an interesting project. Notice when you start to flag, and give yourself a break, to restore focus and avoid burnout.

**Take proper breaks.** Don't think, talk or read about work. Instead, get up, walk around, drink water, go outside, chillax. (There'll be more about how to work & break properly in Part 2 of this article.)

Though it's best to choose when to work and break yourself, nothing changes if others are in charge. Suppose your boss fixes an hour-long meeting for you, then a five-minute gap, followed by a video call. The meeting earns you a 20-minute break; you can only take five minutes of it, but just add the remaining 15 minutes to your next break. (And if you really can't take a full break then, carry the leftover time forward again, and so on.)

## Interruptions

Similarly, nothing special happens if you get a work-related interruption, e.g. your boss calls you while you're hard at it. The call is still work, so the clock keeps ticking. Either deal with the interruption, or postpone it (e.g. send to voicemail) and resume what you were doing.



**Personal interruptions are different.** Suppose you're working from home, and the doorbell rings: **this isn't work, so a break starts then.** Once the interruption is over, figure out when you stopped work, and hence when this break should end (by dividing by 3 as usual). Then either take the rest of the break, or resume work (saving the remaining break time for later). If you've already overrun the end of the break, start work immediately.

## Meal breaks & big breaks

Most people stop work for lunch, some for dinner. With Third Time, you can divide the day into two or three separate sessions—morning, afternoon, maybe evening—to allow for proper meal breaks between.

If you only take a short time off for lunch, a normal break may suffice for it. For instance, an hour-long meeting just before lunch earns you a 20-minute break—enough to eat at your desk, anyway. Or you could save up more eating time by shortening earlier breaks. If you have lunch in a normal break like this, meals need no special treatment, and your whole workday is a single session.

But if a normal break won't do, you can take a **big break**. This means a meal break that **lasts as long as you like**—longer than you've earned from previous work. The only constraint is that you must decide at the start of the break when to resume work, then set an alarm, and obey it as usual. This stops you getting lazy.

If you work into the evening, you can extend dinner in the same way: when a normal break isn't long enough, you can take a big break, provided you set an alarm first. But only take big breaks for lunch and dinner.

Break minutes you've earned are used up by a big break, so **you can't carry them over to the afternoon/evening**. Each session starts with a clean slate. (Hence you don't need to time the work stint just before the meal.) Similarly, you can't carry unused breaks over from one day to the next.

Sometimes you'll have a personal task, e.g. going out to buy something, that's too long to do in a normal break. If possible, do it in a big break, or before/after work. Try to avoid taking unofficial breaks for errands or anything else, lest you allow more and more exceptions to the system, and relapse into chaos.

## Other fractions

What matters is not the *absolute* length of work and breaks, but their length *relative to each other*. Not enough breaks per hour worked, and you burn out; too many, and you don't get enough things done.

[Fitness training](#) and other [physical activities](#) follow the same principle. For instance, with high-intensity interval training (HIIT) you might rest for half of the time you work: between 20-second sprints, you'd rest for 10 seconds; after a 2-minute dash, rest for 1 minute.

Third Time recommends breaking for one-third of your work time (a similar proportion to the [Pomodoro and DeskTime](#) techniques). If this sounds lax, a [survey](#) showed Britons spend less than half their time in the office actually working! So, with Third Time, you'll accomplish much more than many people—particularly if you keep your work highly focussed, with no distractions.

But instead of one-third, you can use whatever fraction you like, such as:

- **1/2:** 40 mins work + 20 mins breaks per hour. Working 2/3 of the time. Lazy
- **1/3:** 45 mins work + 15 mins breaks per hour. Working 3/4 of the time. Standard
- **1/4:** 48 mins work + 12 mins breaks per hour. Working 4/5 of the time. Industrious
- **1/5:** 50 mins work + 10 mins breaks per hour. Working 5/6 of the time. Hard
- **1/6:** 51½ mins work + 8½ mins breaks per hour. Working 6/7 of the time. Grinding

If you take a big break, you could shrink your other breaks by using a smaller fraction. For instance, neuroscientist Dr Daniel Levitin [suggests](#) working 90-minute stints with 15-minute breaks, plus a long lunch. A fraction of  $15 \div 90 = 1/6$  will achieve this.

Should a deadline or crisis strike, so you need to get a lot done fast, don't abandon Third Time—just switch to a smaller fraction. Or save up breaks and use them later. Conversely, if you're tired, or there's not much to do, you could try a bigger fraction to get more breaks.

Different break fractions also suit different activities, from the gym to piano practice.

The other fractions, above right, let you calculate how much work you've done, or will do, without needing a timesheet or schedule. For example, Third Time's default of 1/3 makes you work 3/4 of the time, so in an eight-hour day you'd complete six hours' work—regardless of how your meetings, appointments, and interruptions pan out. (To calculate how long you worked if you took a big break, or didn't use all your breaks, see footnote.[\[1\]](#))

## Advantages of Third Time

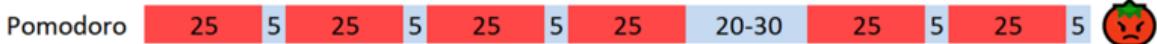
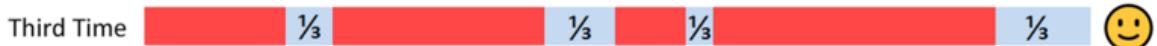
Various time management techniques involve working for fixed time periods. The most popular one, named Pomodoro after a tomato-shaped timer, alternates 25-minute work stints with 5-minute breaks (or occasionally longer).



Tomato overlord

While there are benefits to this kind of pattern, there are big problems too, which I discuss in a [separate post](#). In a nutshell, 25-minute stints may be far from optimal. Indeed, any fixed time period is unnatural and mechanical; thoughtful, creative work doesn't watch clocks. And it's hard to work with other people this way, as their meetings, calls and interruptions won't fit in with your timeslots. Nor will crises and deadlines. Basically, techniques like Pomodoro are too rigid.

This is because they insist on regulating your work. Third Time's key insight is that this is unnecessary—you can guarantee how much work you'll do, just by limiting breaks. This gives you complete freedom in how to divide up your day.



Both techniques achieve 2½ hours' work + 50 minutes of breaks in a morning—but Third Time is flexible. Don't take five, take a third!

People work best with loose constraints, somewhere between total freedom and total restriction. Total freedom is anarchy—the life of the lazy, the workaholic, and the procrastinator. Total restriction is tyranny—ruled by a despotic tomato, and forced to work like a robot.

Third Time applies a light touch, keeping you in the happy, creative zone between these extremes. Its flexibility is its big advantage over other systems:

- **Break whenever you want** or need to, to suit your attention span, energy, schedule, and the task at hand
- You **get the same amount of work done**, however often you break!
- Third Time accommodates **meetings, appointments, calls, interruptions, meals, and personal tasks**
- Unlike Pomodoro, it's **not just for working alone**, so you can **use it all day**
- When on a roll, you **don't get interrupted by alarms** for the end of fixed time stints
- **Save up unused break time** for later
- Choose a **break fraction to suit you**, or even vary it with the situation
- With tedious work, ploughing on **rewards you** with a longer break—an incentive to keep going
- With a **deadline or crisis**, you can take quick, occasional breaks, or none at all. Third Time automatically makes up the shortfall once the storm has passed.

If you like Pomodoro or other methods (e.g. the pressure of their short-term deadlines), you can still use them with Third Time to get the best of both worlds, as it doesn't forbid fixed work stints. I'll explain more in Part 2 of this article. In fact, Third Time is really a general form of time management technique that encompasses the others, and solves their flaws.

Other benefits of Third Time include:

- Encourages **focussed work** and **proper breaks**, rather than inefficient half-working (more in Part 2)
- Ensures you **don't underwork or overwork**
- Easy to **calculate how many hours' work** you've done or will do, without a timesheet or schedule
- You can even **knock off work early** every day (see Part 2)
- It's not just for work—use Third Time for **study, sport & fitness training, personal admin, hobbies, etc.**—anything that needs prolonged focus or effort.

## Summary

Lastly, here in one place are all the steps for using Third Time:

1. Note the time, or start a stopwatch
2. **Work** until you want or need to break
3. Divide how long you've just worked by 3 (or use your chosen fraction), and add any minutes left over from previous breaks
4. Set an alarm for that long
5. **Break** until the alarm goes off, or you decide to resume work
6. If you resume early, note how much time was left, to add to your next break
7. Go back to step 1.

Additional rules:

- If you have to stop work for a **non-work-related interruption**, start a break immediately.
- You can (optionally) take a **big break** for lunch and/or dinner, lasting as long as you like. Set an alarm at the start for when you'll resume work. A big break uses up any saved break minutes, so you can't carry them over to the afternoon/evening.
- **Avoid taking other unearned breaks** if possible—so try to do personal tasks during normal or big breaks, or before/after your work day.

(Third Time would also work well in a smartphone/web app, or smart speaker skill. Anyone want to create one?)

---

**Part 2 (coming soon)** will include tips on how to work and break properly, and using Third Time with other time management systems.

You could also use Third Time with [Hopscotch](#), my way of prioritizing what you work on.

1. [^](#)

To the day length, add any break minutes left over at the end of the day (including any earned by your final work stint). If you took a big break, subtract its length, and add the break minutes you had earned or saved up just before the meal. Then multiply the total by  $3/4$  (for the default break fraction of  $1/3$ ); or for a break fraction of  $1/n$ , multiply by  $n/(n + 1)$ .

*Thanks to Cat and Ari for many suggestions & comments*

# Trying to Keep the Garden Well

I used to live in city-subsidized housing. It was an apartment complex with a big garden.

To enjoy the garden in the summer, residents started putting old chairs and a table outside. Once summer was over, they didn't bother to take them inside again. The table became weathered and didn't offer a pleasant sight anymore.

The city noticed and put up a letter on the apartment board saying that having your own furniture in the garden was unorderly and threatened that any furniture that wasn't removed from the garden by the end of the month would be disposed of.

(You might want to quickly think about the long-run effects this letter had.)

In the following weeks, the residents dumped their old sofas, broken beds, televisions and, what-not into the garden. Normally, disposing of large items like this required you to have a car, drive the items to a disposal yard and pay a fee. However, the city had just made it free.

Or so the residents thought.

Of course, the city didn't think of following through with their threat anymore. Removing a couple of chairs and a table would have fit in a normal workday of a city janitor. Removing this pile of trash would have been costly.

The months went on, the sofas were wet, nobody used the garden anymore. I moved away soon after.

# Implications of Civilizational Inadequacy (reviewing mazes/simulacra/etc)

This was written for the [LessWrong Review](#), intended as a high level exploration of the 2020 posts on Moral Mazes, Simulacra, Coordination Theory, and Civilizational Adequacy.

## I. What Makes For a Useful Review?

When reviewing LessWrong posts, fairly often my reaction is "Cool hypothesis bro, but, can we actually go out and verify it?". A lot of LessWrong posts explore interesting ideas, and those ideas match my intuitions. But lots of intuitive ideas turn out to be wrong, and sometimes people have *different* intuitions. It seemed good to have a step where we actually turned armchair philosophy into empirical claims.

A year ago, I began two attempts at investigating some claims:

- [Eliezer's conjecture](#) that Mathematicians were more law-abiding than non-mathematicians (controlling for IQ).
- The overall Moral Mazes sequence.

## Mathematicians and "Is this data worth the expense?"

[Elizabeth](#) and I made some attempt to run a [Positly](#) study on math / law-abidingness. We quickly ran into "Running a real, legit study is a pretty big time cost." (We ran one quick-and-dirty study that essentially boiled down to asking people 'are you good at math?' and 'do you do crimes?', which Eliezer specifically called out as [Not A Good Approach](#) in his post. We tried looking for datasets that might be relevant to our interests, but finding and sifting through them was hard).

Ultimately I faced the question: "is this question actually worth investing this effort in?" and I felt the answer was "probably not." It seemed vaguely informative, and I felt a nagging annoyance at publishing [Local Validity As Key To Sanity and Civilization](#) without having investigated it. But after a lot of reflection (on the timescale of years), I came to the belief that not every claim is worth investigating.

If I was worried about the epistemic integrity of the LessWrong books, a much cheaper approach would be to include some metadata about the epistemic status of each post, or suggestions of future work that signposted which areas warranted more investigation.

# Moral Mazes, and "What are we actually trying to figure out?"

The Immoral Mazes sequence raised a lot of interesting hypotheses, which matched my intuitions/experience. But it seemed to me the next step was to form some actual predictions and go out and test them.

At the time, I was thinking about Mazes largely from the standpoint of "I want to build a rationalist/EA/longtermist ecosystem that accomplishes good at scale. How do we do that, without succumbing to Mazey pressures?". Example of questions I wanted to test were: "Do more layers of hierarchy reliably lead to more Mazeyness? How reliably? What observations do I actually expect if an org is 'mazey'? Does it help if an org is broken into smaller pieces or does that just shuffle around the problem? How do we evaluate middle managers?"

Testing these things would not be a cheap undertaking. But unlike the Mathematician example, it actually seemed important.

I went to chat with Oliver Habryka about it, and he noted (paraphrased) "For building an organization, I'm not sure it's actually all that useful to study how exactly things fail. I'd much rather see a bunch of case studies of how things *succeeded*. What types of orgs produce value, and how do they go about it?"

This was an interesting takeaway for me – when reviewing posts, the most important takeaways don't necessarily look like "evaluate specific claims". The important thing might be "figure out what decisions this post bears on, use the post as a springboard/pointer to help figure out what you actually need to know."

In my ideal world, there's funding and bandwidth to followup on lots of claims, fleshing out our understanding. Realistically, that bandwidth is limited. I think it makes sense to prioritize questions that are important enough that you'd actually be willing to pay money for the answers (or, questions where people are naturally driven by curiosity or righteous "[someone is wrong on the internet](#)" to get the answers).

## II. The Civilizational Adequacy Cluster

With all that in mind, when I look at the posts from 2020, there's a cluster of themes that fit together.

Some of them clump fairly "obviously":

- [The Immoral Mazes sequence](#)
- [Simulacrum Levels](#).
- Civilizational response to covid, including [Credibility of the CDC](#), and [Seemingly Popular Covid-19 Model is Obvious Nonsense](#). ("[Simulacrum Levels and their Interactions](#)" blends with the previous topic)
- Anna's thoughts on [Where Do/Did Stable Cooperative Institutions Come From?](#), (and jacobjacobs corresponding [Babble Challenge](#))

In my mind, these connect into a fuzzy question of "What actually is up with civilization? Is it functional? Is it trustworthy? Is much of it a corrupt, immoral mess? Can/should we cooperate with it? If not, what do we do instead?"

## Theoretical Underpinnings

Perhaps less obviously, I also think our work on coordination and game theory fits into this cluster. Some of these are approaching the question from a different angle, but ultimately help me build a [unified map](#) of "what's up with civilization?" Some concrete concepts I took away last year:

*Schelling Problems (as opposed to PD or Staghunt)*

Abram's [Most Prisoners Dilemmas are Stag Hunts, most Stag Hunts are Schelling Problems](#) was a crisp conceptual update for me. For the first time I felt like I could look at realworld problems and have a game theory abstraction that worked at multiple levels, where I could see all the way down into the math and it made sense. This dovetailed with [Coordination as a Scarce Resource](#) to me give a sense of *why* coordination problems are hard, but lucrative to solve.

*Simulacrum 3 as Stag Hunt Strategy*

Another conceptual update was some clarity on what Simulacrum 3 / Belief-As-Attire is *for*. We've been using belief-in-belief as a coordination mechanism for getting into better equilibria, and we actually need better alternatives.

Perhaps least obviously, Critch's [Some AI Research Areas And Their Relevance To Existential Safety](#). Much of Critch's agenda seems to explore "how exactly is civilization going to coordinate in a way that preserves human values, when you have tons of AI running around making decisions and executing plans?".

*(But, be careful with game theory)*

My thinking here is also cautioned by Elinor Ostrom's book [Governing The Commons](#), which argues that simplified game theory models in fact are pretty misleading and you need to actually model the actual situation which has tons of nested sub-games in order for game theory to get you "the right answer." This seems true. I nonetheless feel like these two posts ratcheted forward a clear understanding in me, and I have some hope that eventually the theory can turn into real illumination.

## III. Lenses for "Why (In)adequacy matters"

Civilization is huge. "What's up with civilization and why does it sometimes do things that seem crazy?" feels like it ought to be important to a lot of decisionmaking.

I had previously spent a bunch of time thinking through the lens of "what kinds of orgs do I want to build, and how do I make sure they don't become corrupt?". But there are other lenses.

## Moral Mazes and Civilizational Adequacy as Cause Area

Zvi is notably [not an Effective Altruist](#), but when I translate (Im)moral mazes into the EA framework, one obvious followup thought is:

"Hrmm. If the world is infested with Mazes, maybe this is *horrifying and bad* and is maybe worthy of consideration of a top EA cause?"

Two different reasons you might think this:

*Crippling Civilization.* The maze nature destroys an organization's ability to communicate clearly and accomplish their original goals. Mazes try to turn other orgs they interact with into mazes, so this problem may be afflicting large swaths of humanity. This may have significantly harmed our ability to respond to covid effectively, and to other complex 21st century problems. The "moral mazes" problem is bad because it cripples civilization as a whole.

*Bad for the People Involved.* The maze nature is harmful to the welfare of individuals who end up trapped within it. People are encouraged to devote everything (their family life, their hobbies, their free time) to the organization. The organization doesn't even care about its own ostensible goals. Only a teeny fraction of people who make it to "the top" get any kind of reasonable payoff for their sacrifice. Since mazes try to create more mazes, this means a large swath of humanity is afflicted by soul-sucking machinery.

### **Bad for the People Involved. Or, "Classic Givewell-style EA"**

Point #2 basically amounts to some cluster of "People are suffering, people aren't achieving their potential, living the fulfilling lives they could be." And there's some math about whether the scope of this problem, and the tractability of solving it, are comparable to people suffering/not-achieving-potential because they are killed by malaria or whatever.

I don't actually know the answer to the math. There's a question of to what degree moral mazes are pervasive, and how unhealthy they for the median person caught up in them. Then there's a question of what sort of interventions work and how effective they are. It's hard to do 'vanilla' EA math until there's at least some people who've tried solving the problem, whose progress can be evaluated.

From this perspective, a next step might be "thinking more about the scope of the problem, and what might help."

### **Crippled Civilization, or "Maybe humanity's hamming problem?"**

Point #1 seems like a strong contender for "humanity's biggest problem." It might or might not be tractable, but if true, it affects all other problems we might need to coordinate around.

"Moral mazes" is a potential subproblem of "civilization just doesn't seem to be able to reliably do sensible-things-at-scale." And here is where a lot of the other "Civilizational Adequacy post cluster" come in. There seem to be multiple problems at play, all the problems seem gnarly and hard with a lot of entrenched political interest.

I'm not sure I trust a lot of the evidence that has been reported to me about "how functional is civilization?". Information about this is heavily politicized, and anecdotes that reach me are selected for being outrageous. Still, I've gotten a lot of bits of info that seem to at least there's a ton of room for improvement here.

Here, some obvious questions are: Does the moral mazes framework suggest solutions to 'civilization not being able to act sensibly at scale'? Do simulacrum levels?

My naive thoughts all go in the same direction as my original query "how can I avoid creating moral mazes, or minimize their damage/degree?". Oliver's suggestion seems potentially still relevant: "look for successful organizations, and parallels between them, rather than spending a lot of time thinking about failure." Only in this case it's something like "Look at the places where civilizations succeeded. Maybe particularly look for places where a civilization's competence sort of wavered and recovered." (I'm wondering if this is a good job for Jason Crawford)

But I'm interested in followup work that explores entirely different avenues. "Look at success" is harder when the topic is macro-history, and there are probably other options.

## Civilizational Adequacy as AI-Strategy Crux

The previous points were mostly things I thought about last year. But this year, as I chewed on the [MIRI 2021 Conversations](#), it struck that there's a different application of the "is civilization terrible?" question.

It seems to me that whether one believes civilization is horribly inadequate has a pretty big impact on one's AI strategy.

If you think our civilizational apparatus is basically functional, or *close* to functional, then it looks more like a reasonable approach to base your AI macro-strategy on government policy, agreements between companies, or changing the longterm AI research institutional culture.

If you think civilization is dysfunctional, you're more likely to think you need a [pivotal act to permanently change the gameboard](#), and that all the other stuff won't help.

(I don't mean to imply this is the [only major crux](#), or even the most important one. But it seems like a fairly big one)

In many areas, I don't think it's *that* bad that the LW / AI Risk community struggle to reach convergence. There's a lot of research ideas, and it's maybe just fine to have lots of people pursuing their own agendas in parallel while thinking each other's research agendas are doomed. The fact that we can't agree is evidence that at least one of us is *thinking* wrong, but the [best group-epistemic strategy](#) might not be to try to enforce convergence.

- a) I think disagreement about this *is* a pointer to something that (I think) should be important to people's individual epistemic journeys
- b) the disagreement is particularly annoying because it points in fairly different directions on what one is trying to *do* with an AI Alignment community, which leads to some plans working at cross-purposes, with people vaguely annoyed or embarrassed by each other.

Unfortunately, this seems like a topic that is some combination of political and aesthetic, and easy to collapse into a "yay/boo civilizational adequacy?" question rather than making specific predictions. When I imagine trying to get other people to engage seriously with the topic I feel some despair.

I think there's one set of biases that color people's thinking by default (i.e. generally having a hard time distinguishing social reality from reality, going funny in the head with politics, being vaguely optimistic about things, etc).

...and another set of biases that LessWrong folk tend to have (selected for not playing well with others, being extra annoyed at things that don't make sense according their strong inside view, etc)

I'm still confused about how to think about this. I think the most useful next action for me to take is to try to think about this for myself, without worrying about whether anyone else agrees with me or is thinking about it the same way.

I feel drawn to questions like:

- Which parts of civilization are the relevant parts for coordinating on AI deployment?
- How can I get unbiased samples of "what actually happens when the relevant pieces of civilization try to coordinate at scale?"
- How quickly do arbitrary AI companies adopt best practices? If good alignment technology is developed, but in a different paradigm than whatever Deepmind etc are using, would Deepmind halt its current projects and switch to the new paradigm?
- How lumpy is progress? I feel confused about Eliezer and Paul's debates about how smooth the economy actually is in practice.

But maybe the best question is "what is the actual best end-to-end plan that routes through large-scale coordination, vs what is the best end-to-end plan routing through pivotal-act-type interventions?" Having actual details would probably be pretty grounding.

## Mazes as "a problem for institutional Effective Altruism™"

Originally I was interested in Moral Mazes from the perspective of building EA, Longtermism, or other nearby organized efforts to accomplish important things at scale.

I touched on this in the first section. But I wanted to spell out some of the concerns here.

**The Middle Manager Hell hypothesis** (my name for a subset of Moral Maze theory) is: A problem with middle managers is that it's hard to evaluate their work, which makes it tempting to evaluate them in goodhart ways, or de facto for evaluations to run through some manner of politics and network effects. This gets much worse when there's enough layers of managing that the managers interact with each other to build their own internal ecosystem that's untethered to object-level-work.

We also face this problem with longtermist research. You could cheekily describe longtermism as "believing that the most important things have the worst feedback loops." If longtermist EA is to scale its impact, it needs to deal with many human-elements of "wanting money/reward for doing good work, a sense of how to progress/gain-status, etc". Building an ecosystem that effectively rewards good work is important, but it's hard to tell what "good work" means.

Longtermism-at-scale *both* might require middle managers *and* researchers in topics that are hard to evaluate, which could add up to a doozy of a hard time for healthy-ecosystem management.

Some problems might include:

- There are many individual organizations, which may scale, and become more maze-y over time.
- The ecosystem collectively sort of acts like an organization, and it's unclear how this plays out. In the mix of community organizers, managers, organizational grantmakers, and re-grantmakers (i.e. sometimes one org grants money to another org to re-grant), you may end up with the same "management class" that is increasingly untethered from the object-level-work. A worry might be status ending up getting allocated via passing favors around rather than demonstrating skill/good-judgment. And this might lead to a collective maze-atmosphere.
- A lot of people seem to be following a strategy of "gain power/prestige within mainstream institutions, so they can use that power to do good later." Those mainstream institutions might be mazes, and those people might end up self-modifying into an increasingly maze-like outlook which filters back into EA.

It's worth noting a distinction between "mazes" as a corrupting influence from the mainstream world, vs *other* forms of corrupting influence. Moral Mazedom is one particular theory. But, in some cases it doesn't seem necessary to explain what's going on. "Some EAs care about prestige and the brand of EA, which makes them hesitant to talk openly about important problems" could be explained by things other than mazedom.

I think one key question here is "how do you actually do longtermist research?", which is not a small question. This includes (but is not limited to) how to do pre-paradigmatic research, [which has some intrinsic difficulty](#) regardless of how you've tried to organize your meta-community.

If you don't know how to do that, then "how to middle manage longtermist research" is even more confused.

## IV. Takeaways

That was a lot. Here's a recap of some pieces that felt important:

- Getting from "interesting hypothesis" to "vetted knowledge" is expensive, and not necessarily worthwhile. Reflect on what information is actually decision-relevant.
- Civilizational (In)adequacy seems it should be a major crux for various high level EA strategies – for EA institutional infrastructure, AI strategy, and a common cause among causes.
- Maze Theory suggests a mechanism by which dysfunction is spreading.

I feel a bit sad that in each section, I only have vague inklings of what further work is actually useful. Each lens for "why (in)adequacy matters" felt like it warranted a whole other post. There's a lot of work left to do here.

# Omicron Post #12

It involved a backlog, but the USA reported over 1 million cases yesterday. That's a lot of cases, and we're missing a ton more of them. Schools, hospitals and all businesses face massive disruptions from people unable to work, even with the new CDC isolation guidelines.

So far, we're holding up well, better than I expected even after previous severity estimates were adjusted. Things are going to be severely disrupted for a while, but not as severely as we might have feared, and even if we get a median outcome that is much better than the mean outcome given the tail risk of true disaster.

Still, brace yourself, and choose how you want to handle the next month or two. If you're serious about trying to avoid Omicron, now is the time to show it.

## Vaccine Effectiveness

[A Danish study](#) from earlier [found large negative vaccine effectiveness](#). Large negative effectiveness is not a thing, we would know about it and the world would look very different, but anti-vax people are of course jumping on it and the attempt to explain isn't going all that well. Which it mostly shouldn't, in my view, because when you get nonsense results that are clearly nonsense results but you don't control properly and so you publish anyway, what else do you expect?

I looked at the paper. They control for the basics, including geographical region.

VE was calculated as 1-HR with HR (hazard ratio) estimated in a Cox regression model adjusted for age, sex and geographical region, and using calendar time as the underlying time scale.

I've heard reasons that make it *possible* for vaccine effectiveness to go negative, but those reasons don't seem compatible with large positive effectiveness shortly after two doses, and large positive effectiveness again after a booster shot.

More than that, check the sections on severity and hospitalization below. They make it very clear that being vaccinated offers a lot of effective protection and is very much a good idea, and getting boosted is an even better idea.

## Treatment

[CDC lifts the pause on anti-Delta monoclonal antibodies](#), says they can be used if Delta is still in your region and there aren't other options available. My guess is there are still some places in America where this makes sense, since even a small chance of getting the old effectiveness is better than doing nothing, but this window is ending rapidly.

## Quarantine, Isolation and Travel

Instead of ending all travel-based quarantines because they no longer make any sense, many new travel restrictions are being imposed.

This is deeply stupid. It makes sense to make someone quarantine if they are at much higher risk than the surrounding population, and there's some hope of stopping the spread. Two places having *mutual* travel restrictions usually means both are making a mistake. These

quarantines are quite expensive (as are the required PCR tests, which given the need for reliable results often cost hundreds).

What prompted this was that I noticed that [France added the United States to their 'red list'](#), including the first comment as a 'same energy' to indicate the mindset.



Reuters Health @Reuters\_Health · 22h

France has put the United States on its COVID-19 travel "red list", meaning unvaccinated people coming into the country will have to quarantine for 10 days.



[reuters.com](#)

Unvaccinated U.S. travellers added to French quarantine list

France has put the United States on its COVID-19 travel "red list", meaning unvaccinated people coming into the country will have to ...



Anasanzi0928 @anasanzi0928 · 19h

Replies to [@Reuters\\_Health](#)

Shouldn't allow them in at all! Same here. Vaccines/mask mandate if not a 60 day lockdown. Bill of Rights doesn't cover the rights to infect ppl you come into contact with. As crazy as a person with AIDS saying it's their right to infect ppl and there's nothing wrong with it.

France has twice as many cases per capita as the United States right now, and about two thirds of the death rate.

At this point, quarantines even for people with large known exposures seem questionable. If you learn you've been exposed to Covid-19, you have not received that many bits of information, and it's not clear how much behavior change is justified. Not zero, especially if you're going to potentially see vulnerable people, but if you used the old thresholds for quarantine you would (quite literally) never leave the house.

At least it's better than universal isolation, [such as the curfews in Quebec](#):



**EMERGENCY ALERT / ALERTE D'URG...** maintenant  
Message du gouvernement du Québec : un couvre-feu entrera en vigueur à compter de ce soir pour tout le Québec. Interdiction de se trouver à l'extérieur de sa résidence ou de son terrain entre 22 h et 5 h. Surveillance policière accrue. Pour plus d'information, visitez le site Web du gouvernement du Québec. /// Message from the government of Québec: a curfew will be in effect as of tonight for all of Québec. It is forbidden to be outside your home or property between 10 p.m. and 5 a.m. Increased police surveillance. For more information, visit the website of the government of Québec.

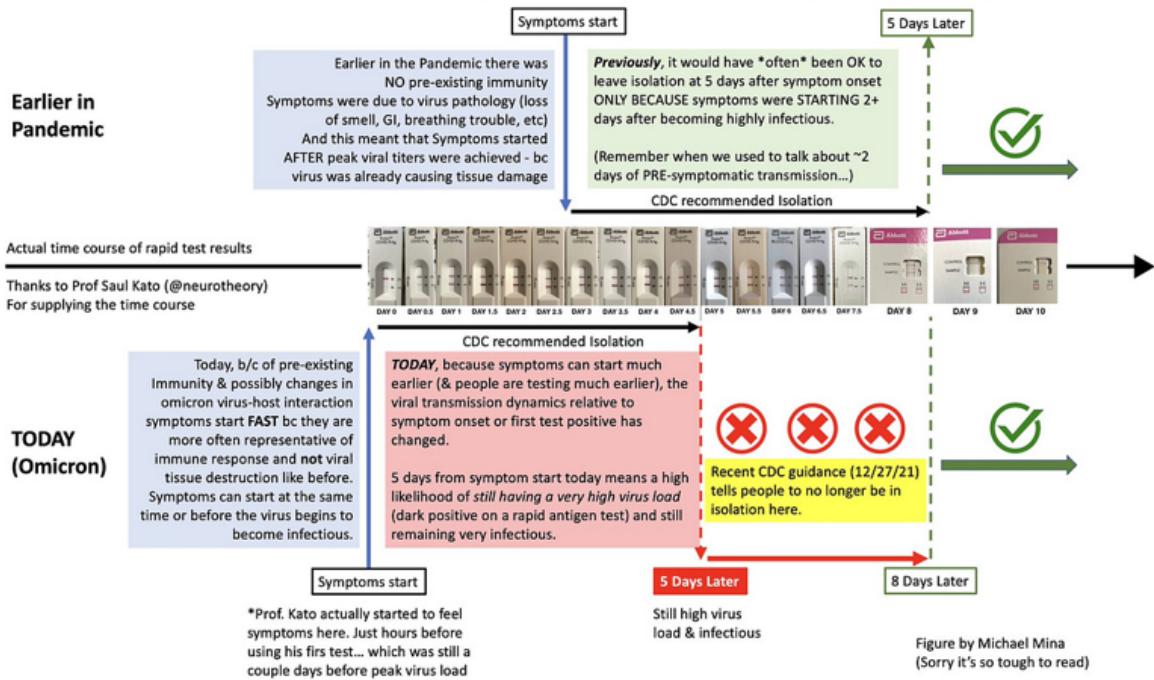
[Whereas Princeton is telling its students they can't leave the county](#) after they arrive on campus. Not the country, the *county*. I am confused how they think this will help.

[Delhi is locking down for the weekend](#). I have no idea what this hopes to accomplish. [Reported cases in India are up 488% in a week](#) (presumably they're missing most of them), but what good will three days do? Allow the countryside to catch up a bit?

Here's a graphical argument liked by Mina from Saul Kato for why you need a negative test before you end isolation ([the thread contains more details](#), but we've seen those before).

## Why the CDC should reconsider its guidance to exit isolation at 5 days without a test

The Relationship between CONTAGIOUSNESS (when someone should isolate) & Symptom Onset has changed  
 [And why we need a ***negative*** test to safely guide us out of isolation and not just 5 days post symptoms]



[The CDC's guidance is definitely having an impact](#), for better and for worse.



Rachel Cohen @rmc031 · 21h

the texts I've been having with my covid positive doctor friends recently...

...

Today 11:02 AM

I just got cleared to go back to work today, day 4. I'm livid with how the conversation just went with health services

Tell me

no test I'm assuming?

"We're encouraged to get doctors back to work"

No test

I was saying that I haven't even been quarantined for 5 days but she didn't care

i just got a call saying that I  
get to go back to work  
tomorrow instead of  
quarantine until the 30!  
Lmao

My rapid was positive  
yesterday but they dgaf



**CptCarr** @Got\_Carr · 17h

Replies to [@rmco31](#)

It's the same in the restaurant industry right now as well. I still have symptoms, and they are telling me 5 days was enough.

4

21

281



**Penny Ell** @pennyaell · 16h

I knew this would happen when the CDC said 5 days if asymptomatic - that employers would ignore the 'asymptomatic' part and only focus on '5 days.'



**Chris McKercher** @System\_is\_broke · 7h

Replies to [@rmco31](#)

I know of a nursing home where staff has deep chest coughs and are being made to work anyway. One told me that she had the flu, but was better, and that she had a Covid test after being back a couple of days, and was waiting for results.

If you tell someone who is *symptomatic* to come back to work, the results to your business are not, in expectation, going to be what you are hoping for.

Yes, some amount of this was inevitable from the new guidelines, especially given the lack of a testing requirement, but restaurants have been making sick people come in when they very much should not come in (even if it means closing the place down for a bit) for a long time and they are not about to stop now.

There are some who are going to treat the CDC requirements as a *minimum*, without which doing anything is crazy. Others are going to treat it as a *maximum*, and then make

compromises. You do have to set the requirements knowing both of these things will happen. It's a hard problem.

Doctors being told to come back in four days with no test means they're going back while still infectious. Is that better or worse than having patients go without care due to lack of staffing? I don't know.

I do know that [this Colbert clip](#) shows how symbolic the whole thing is. Colbert can come back in five days *if he wears a mask at all times including during the monologue* when he's dozens of feet away from the nearest person (or if not, he certainly could be) but otherwise he needs to wait ten and she doesn't tell him to test. The need to pretend that the guidelines make sense forces director Walensky into Obvious Nonsense territory. She did have the decency to nervously force herself to laugh.

## Disruptions other than Hospitals

Staffing shortages shut things down. A substantial portion of local restaurants have been shut down at some point by now, for example, along with several subway lines.

When we check in on our 'not an alarmist' account, we see these: [A county in Maryland cancels school bus routes. Multiple school districts close entirely. Milwaukee going to virtual schooling due to staffing shortages. So is Newark. A prediction that almost all schools will close during January. All NYC Apple stores are closed. A third of Philadelphia classrooms forced to go remote on Tuesday.](#)

I updated a small amount in favor of things being somewhat less disrupted than I expected, because given the information networks available, I would have expected the account to find *more and bigger* examples of school disruptions. Still, Milwaukee is big, and Philadelphia is big.

So far, both my children's schools remain open, albeit with large drops in attendance and a bunch of [michegas](#) regarding testing.

[There's a reasonable argument](#) that I buy, which says that if you believe school is what its advocates say it is, schools should be kept open as long as possible. NYC Mayor [Eric Adams \(man saying that is such a relief\) definitely buys that, saying that 'the safest place for our children is in a school building. We are going to keep our schools open.'](#)

The claim that a school is 'the safest place for our kids' in a pandemic is grade-A Obvious Nonsense, and presumably comes from this logic:

1. We must always do The Safest Thing for our kids.
2. We need to open the schools.
3. Therefore, the schools are The Safest Thing.

The problem is that people mostly aren't willing to let such things 'muddle through' in any sense, so this kind of will often won't generate a way.

Schools aren't going to be allowed to say 'yes, the school is going to involve a lot of Covid-19' and instead imposes lots of tests, which finds cases, which forces isolations and shuts down classrooms. Even without that, staffing shortages would still make it impossible to keep many schools and classrooms open, because the system does not have enough slack for such an event.

Businesses and other activities that can survive with limited on-site staff by muddling through, or that are important enough that people should come in anyway if it comes to that, should mostly be able to stay open.

Those that cannot take that approach are going to face a lot of disruption over the next few weeks. Be ready.

Ideally we can find the right balance of shutting versus muddling. It's difficult.

## Spread

One million cases reported in one day. Wow. Backlogs from New Year's and the weekend were involved, to be sure, but still, wow.

[Georgia didn't update because too much data overwhelmed the system](#), so you can guess how well things are going there.

Seattle offers some strange graphs, [where if we use SGTF deletion as our guide, we see Omicron fading away and Delta increasing](#). I was told this was from Trevor Bedford, who has otherwise been an excellent data source, but there's simply no way the graphs are indicating what they say they are. One possibility is that the other version of Omicron, that doesn't have SGTF deletion, is rising in relative terms.

[San Francisco infections may already be peaking](#). My guess is this is premature, but we are seeing data points in this direction. Could be the same holiday-related fluctuations, as well.

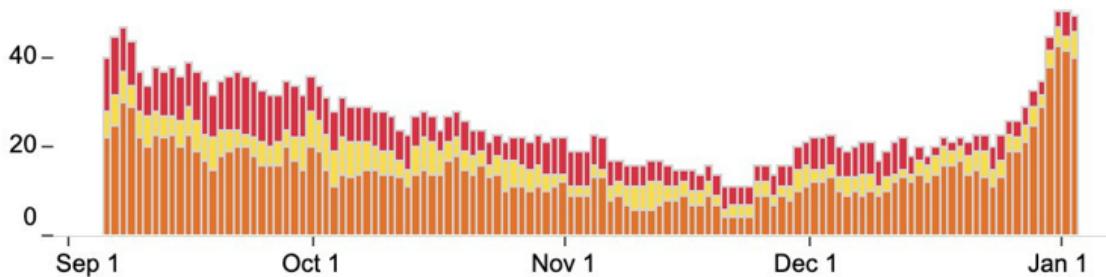
 **Bob Wachter**  @Bob\_Wachter · 20h ...

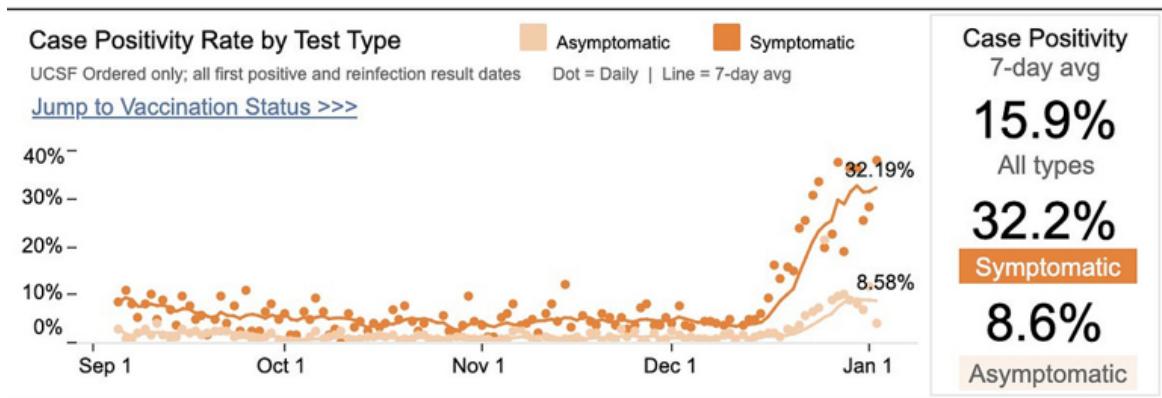
WAY too soon to be sure (especially w/ potential spread over holidays), but we are seeing a plateau [@UCSFHospitals](#) in hospitalizations (Fig L) & test positivity (both symptomatic & asymptomatic; R). Next few days will be key. Until we see sharp downturn, I'm in uber-careful mode.

COVID-19 Test Positive Daily Patients Hospitalized

 ICU, Vented       ICU, Not Vented       Acute Care/TCU/ED

[Jump to Vaccination Status >>](#)





[New York is going to start having hospitals report how many people are hospitalized ‘with Covid’ versus ‘for Covid’ starting in five days.](#) This will be a great source of data, hopefully exactly in time to not much matter.

[This article highlights that 30% of EMS workers in NYC are currently out sick.](#) The thing is, they’re not all out sick, they’re all sick *or isolating*.

Amid the spike in calls, 30% of the city’s 4,400 emergency medical services (EMS) staff were out sick as of Wednesday, according to Frank Dwyer, an FDNY spokesperson. He added that the surge in people out sick was a combination of COVID-positive workers, those who’d been exposed and staff awaiting test results.

That means we can’t use the 30% to estimate prevalence, and it also means that we’re holding too many people back. If you’re an EMS worker, yes, *you’ve been exposed*. That’s the job when 10%+ of the city has Covid-19 at once, how could you not be exposed?

## Calls To 911 For Fever And Cough In NYC

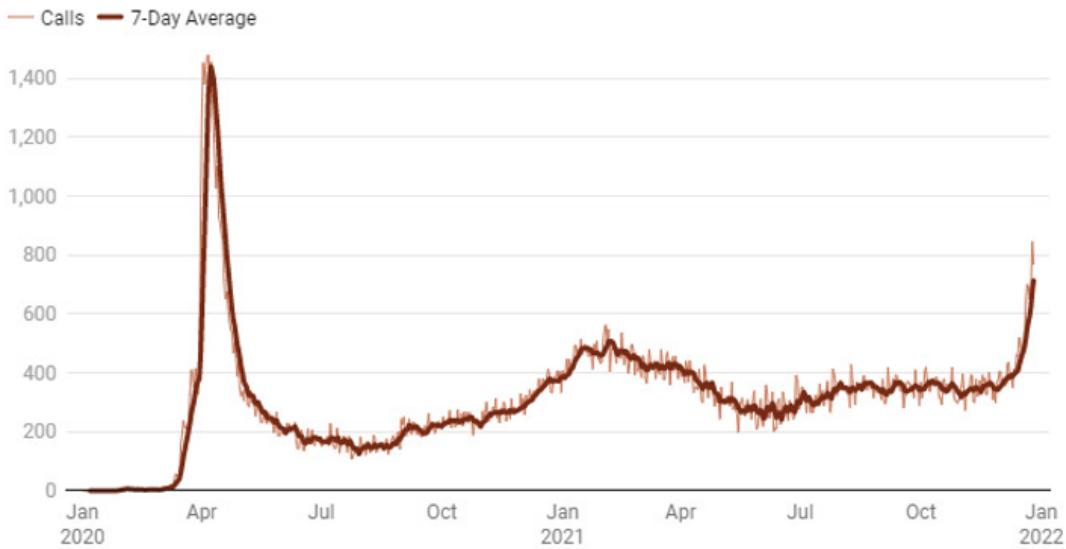
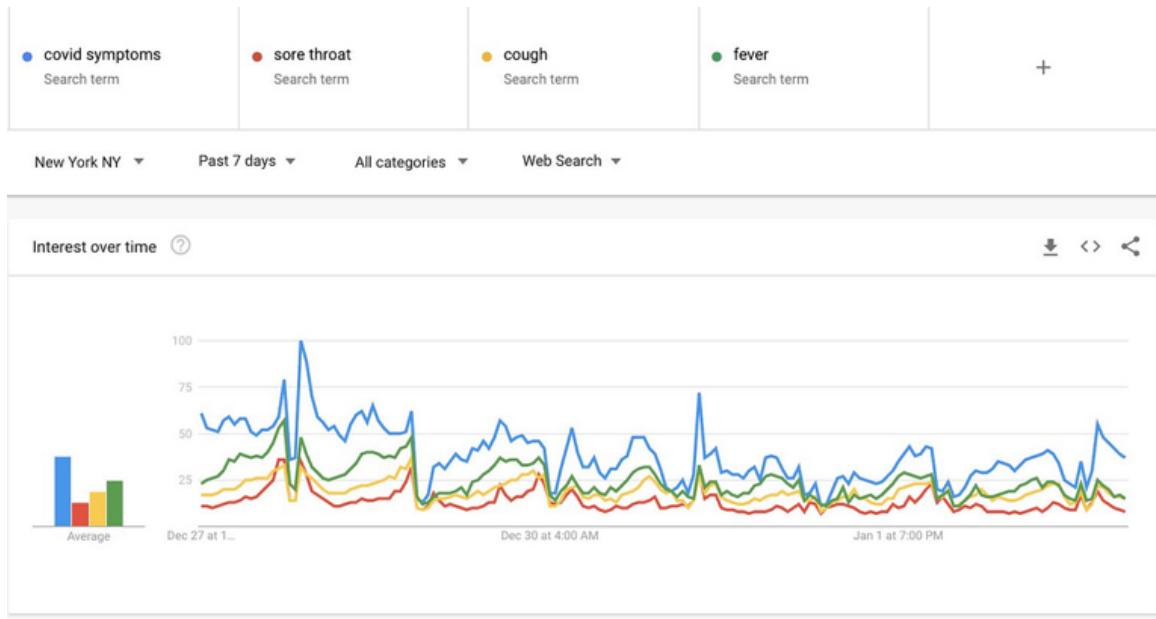


Chart: Jaclyn Jeffrey-Wilensky • Source: FDNY • [Get the data](#) • Created with Datawrapper

Having such people isolate after an exposure or while waiting for test results is going to do far more harm than good at this point. It’s also fully incompatible with ending isolation *after a positive test* after only five days without a negative test, although I assume the EMS is ignoring that suggestion and requiring the negative tests anyway.

Nate Silver notes [a decline in Google searches for Covid symptoms](#) in NYC.



My guess is that at this point everyone's already Googled a lot, and a lot of this is pseudo-random, but it is an interesting data point. Nate also notes that the positive test rate has stabilized at 23%, which seems like stronger evidence that we may be at or near the peak.

## Severity

[Good news from London.](#)



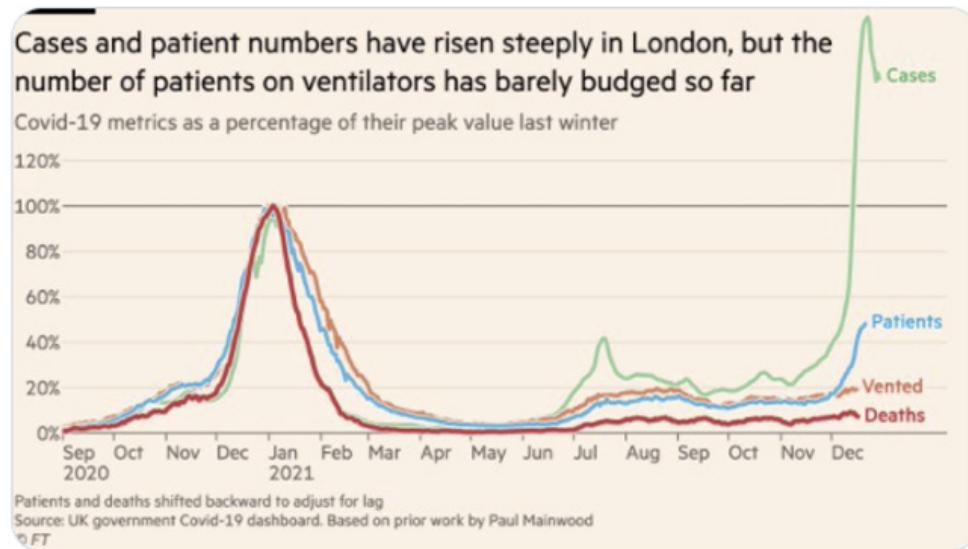
John Burn-Murdoch ✅ @jburnmurdoch · 1h

...

NEW: first thread of 2022 is an Omicron situation update, starting with a detailed look at UK hospitals, before going international.

Let's start with severity, and the most important chart:

Despite steep rises in cases and patients, the number on ventilators has barely risen.





**John Burn-Murdoch**  @jburnmurdoch · 1h

...

Replying to [@jburnmurdoch](#)

I've adjusted for lag between positive test and severe disease, so the divergence between patients and ventilated is a marked difference between this and previous waves when they were in lockstep.

The link between cases and severe disease has significantly weakened with Omicron.

7

58

598



**John Burn-Murdoch**  @jburnmurdoch · 1h

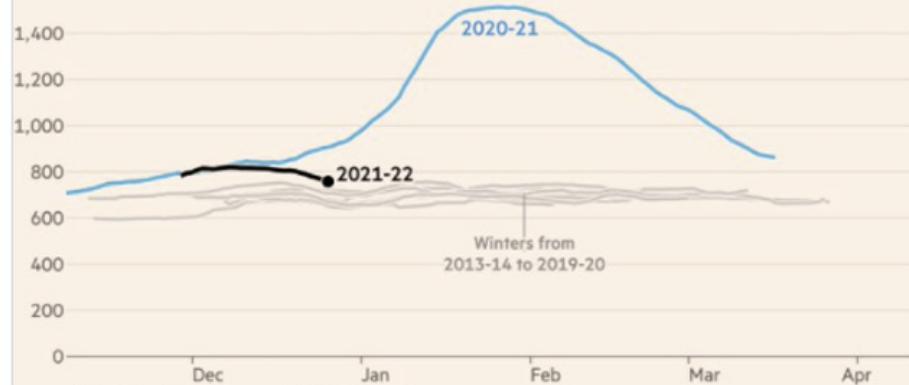
...

We also continue to track "excess ICU pressure" — total number of people in ICU, for any reason, Covid or otherwise, compared to past winters:

Latest data show that the number of people in London ICUs has fallen in recent weeks, and is not following the same path as last winter.

### Total ICU bed occupancy in London has dipped in recent weeks and is not tracking the rise seen last winter

Total number of adult critical care beds occupied in London over recent winters



Source: NHS urgent and emergency care daily situation reports  
FT graphic by John Burn-Murdoch / @jburnmurdoch



7

236

893



**John Burn-Murdoch**  @jburnmurdoch · 1h

...

So that's two very strong signals that this wave is not the same as those that have come before it in terms of severe disease and the pressure placed on ICUs.

This is unambiguously good news!



**John Burn-Murdoch** @jburnmurdoch · 1h

...

There are many caveats, and that's before we get onto the wider pressures facing NHS (read on for the deep dive into that)...

First is age: cases have turned the corner among young adults, but rates among elderly have kept climbing, rising more than 10x among over-80s to Jan 2

There's been enough time that if John's adjustment is remotely reasonable, the cohort that got sick earlier was at little risk. The question now is what will happen with the elderly, for whom not as much time has passed.



**John Burn-Murdoch** @jburnmurdoch · 1h

...

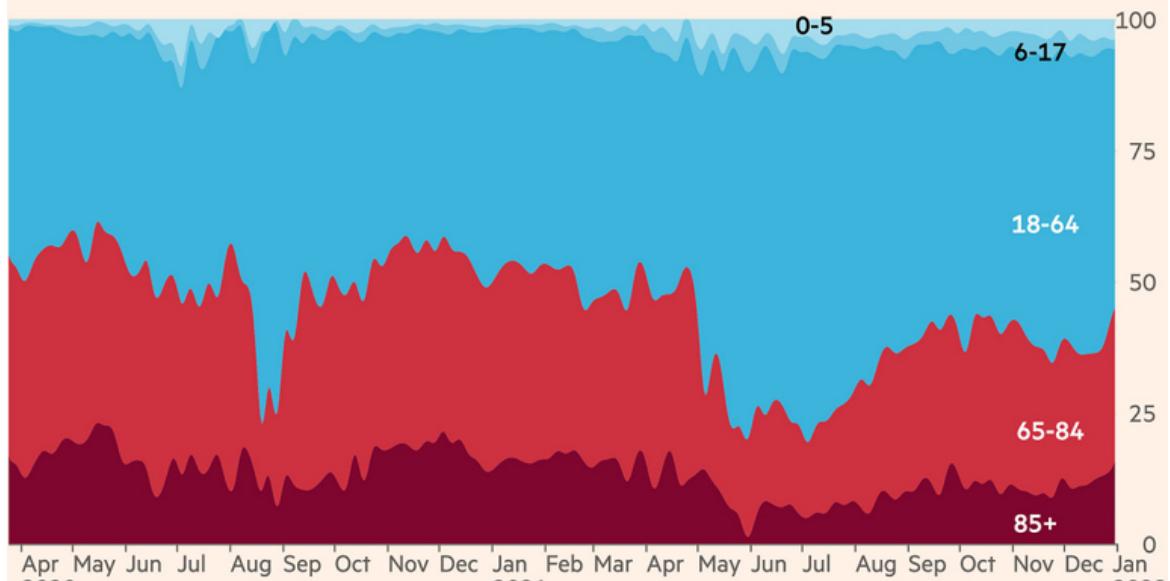
And just as night follows day, this has fed into hospital admissions.

In the last few days the share of admissions among people aged 65+ has ticked upwards.

We know Covid has a steep age gradient, so this may increase the share of patients whose condition becomes serious.

## In recent days the elderly share of Covid patients has started ticking up

Percentage of all London patients that are from each age group



Source: FT analysis of UK government Covid-19 dashboard  
© FT



John Burn-Murdoch ✅ @jburnmurdoch · 1h

(it may also be worth noting here that a study from Gauteng, South Africa, last week found that the reduction in severity among hospital patients was smaller among over-60s than among young adults  
[twitter.com/miamalan/statu...](https://twitter.com/miamalan/status/146081111000000000))

This is the data from Gauteng:

**Table 2:** Indicators of disease severity among SARS-CoV-2 positive cases admitted in the Beta- (29 Nov-26 Dec 2020), Delta- (2 May-29 May 2021) and Omicron-dominated waves (14 Nov-11 Dec 2021), Gauteng Province, South Africa

Variant wave	Number of cases admitted to hospital with known outcome	Median length of stay in days (IQR)	Percent (Number) of admitted cases who received supplemental oxygen (n)	Percent (Number) of admitted cases who were treated in ICU (n)	Percent (Number) of admitted cases who had severe disease (n)
All ages					
Beta	7774	7 (4;11)	39.4 (3063)	20.0 (1552)	60.1 (4672)
Delta	4574	8 (4;14)	48.8 (2231)	26.2 (1198)	66.9 (3058)
Omicron	4438	4 (2;6)	19.7 (875)	6.9 (308)	28.8 (1276)
Age <20 years					
Beta	306	4 (2;8)	14.7 (45)	3.9 (12)	22.5 (69)
Delta	161	3 (2;8)	15.5 (25)	9.9 (16)	23.0 (37)
Omicron	844	3 (2;5)	13.0 (110)	4.3 (36)	20.4 (172)
Age 20-39 years					
Beta	1697	5 (3;9)	31.3 (532)	11.5 (195)	42.6 (723)
Delta	714	6 (3;11)	35.4 (253)	14.0 (100)	46.6 (333)
Omicron	1862	3 (2;6)	13.0 (242)	3.6 (67)	19.1 (355)
Age 40-59 years					
Beta	3188	7 (4;11)	42.6 (1358)	21.6 (689)	61.9 (1972)
Delta	1685	8 (5;13)	51.1 (861)	27.2 (459)	67.2 (1132)
Omicron	983	4 (2;7)	22.8 (224)	9.5 (93)	34.2 (336)
Age ≥60 years					
Beta	2583	8 (4;13)	43.7 (1128)	25.4 (656)	73.9 (1908)
Delta	2014	9 (5;16)	54.2 (1092)	30.9 (623)	77.3 (1556)
Omicron	749	5 (3;9)	39.9 (299)	15.0 (112)	55.1 (413)

The >60 year old category there definitely shows a much lower drop-off.



**John Burn-Murdoch** @jburnmurdoch · 1h

So, less severe disease this winter than last, but as [@ChrisCEOHopson](#) has painstakingly explained, this doesn't mean everything is fine in hospitals. Far from it, in fact.

Let's take a look at the different pressures facing the NHS this time around:

From there the thread pivots into what's happening to hospitals, including reference to [this additional thread on that](#). I'll cover that in the next section. That other thread does refer to severity in elderly patients, and finds it substantially improved:



**Chris Hopson** @ChrisCEOHopson · Jan 1

Providing more evidence vaccines work with omicron and risk of serious illness for fully vaccinated elderly looks much lower than before. Latest primary diagnosis data shows growing incidental covid – patients coming to hospital for another reason and testing positive...13/25

5

64

431



**Chris Hopson** @ChrisCEOHopson · Jan 1

But issue is G&A bed pressure, whatever reason for admission. Covid cases require isolation, potentially reducing capacity. So important not to over-emphasise “beneficial impact” of incidental covid, as some doing today. Hope this is a tall thin peak that plateaus soon....14/25

It then pivots back.



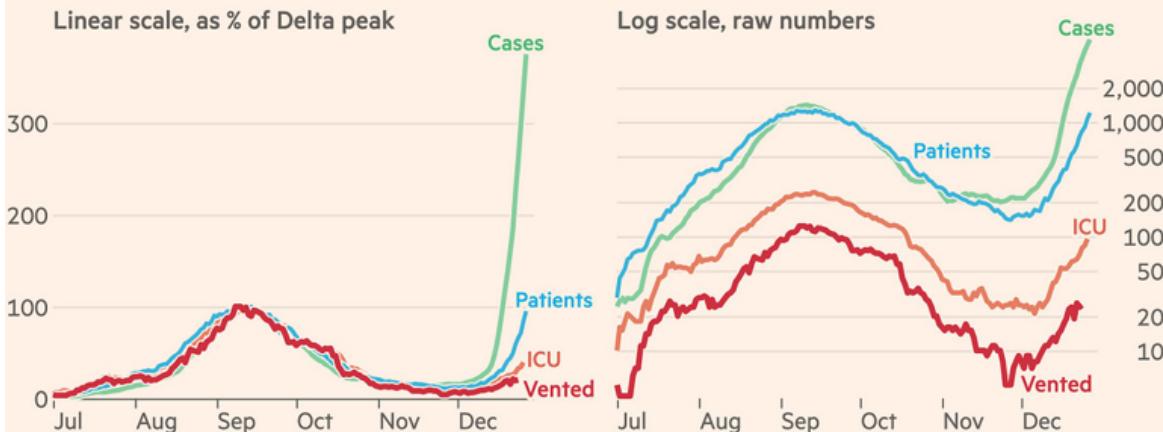
**John Burn-Murdoch** @jburnmurdoch · 1h

On the left, [@edjegasothy](#)'s approach shows the link between cases and acute outcomes has weakened considerably: good news

On the right, [@OYCar](#)'s preferred chart shows ICU and vent numbers are following cases upwards: bad news

Both show the same data. Both show reduced severity.

The number of patients with severe disease in New South Wales is rising, but the ratio of severe disease to cases has weakened considerably



The log scale on the right means that it's the gap between the lines we should pay most attention to, and it clearly does indicate good news.



**John Burn-Murdoch** @jburnmurdoch · 1h

It's not that Omicron is more severe for kids than past variants, it's "adults now have loads of immunity, kids don't, and hospitalisations will absolutely exceed past peaks among the unvaxxed, young or old"

6

85

359



**John Burn-Murdoch** @jburnmurdoch · 1h

Another issue is the level of existing pressure Omicron arrives into. In the US, Omicron has landed right on top of a nasty winter Delta wave.

Covid ICU numbers are approaching record highs in several states, and even a less severe variant can push them over the top

[Similarly, this thread](#) is on the situation in the hospitals and draws clear contrasts of those boosted, those vaccinated but not boosted, and the unvaccinated.



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

This version of Covid is crazy contagious. Our ER hit record highs in patient volumes. Hospitals everywhere over capacity during this surge of people coming in with Covid.

There is a strikingly consistent pattern based on vax status...

4

72

383



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

Boosted? Hardly saw anyone who had gotten a booster because if they caught Covid they're likely at home doing fine or having regular cold/flu like symptoms

8

88

563



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

Vaxxed/unboosted? Tons of patients like this: wiped out, dehydrated, febrile. If they were older (eg > 55 yo or had other medical problems) often had to admit for overnight IV hydration and supportive care. But usually go home within a day or two.

6

98

561



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

Unvaxxed: These are the folks that get sick and had to be hospitalized because they need oxygen, some even younger than me. Fortunately Philly has a fairly high vaccination rate, particularly in older adults, so we're not seeing severe cases as much as we used to.

## Hospitals

Our hospitals in many places were already under severe strain due to Delta. Now there's a shortage of staff due to illness, and lots of Omicron patients showing up. Those patients mostly aren't landing in the ICU, but the situation remains under tremendous strain on both sides of the pond.

## Covid-related NHS staff absences in London have risen steeply in recent weeks, though there are signs of a reversal in the last few days

Number of hospital staff absent due to Covid (either sick or isolating)



Source: NHS England  
© FT

UK hasn't lowered their isolation requirements yet, which is compounding the problem there.



**John Burn-Murdoch** @jburnmurdoch · 1h

(To combat this, I agree with [@timspector](#)'s call to reduce the length of isolation for hospital staff who come into contact with a known case. Obviously this would be conditional on not developing symptoms and continuing to test negative [twitter.com/timspector/sta...](https://twitter.com/timspector/status/134444444444444))

...



**Tim Spector** @timspector · Jan 3

W Covid cases in London now decreasing + UK slowing - it is great to see no real change in Covid deaths over last month. The health crisis is in danger of being driven by staff problems due to over-cautious isolation rules. Lets reduce this to 5 days!  
[coronavirus.data.gov.uk/details/deaths](https://coronavirus.data.gov.uk/details/deaths)



**John Burn-Murdoch** @jburnmurdoch · 1h

The dual factors of surging demand and diminished supply would be a problem at any time, but this winter they come against a very different backdrop to past UK waves...

1

12

177



**John Burn-Murdoch** @jburnmurdoch · 1h

During the first two big waves, hospitals scaled down the amount of non-Covid treatment to create capacity.

This winter they're trying to work through the resulting backlog of often very ill people waiting for treatment, and the result is far less headroom, far more stress.

Backlogs persisting for months is entirely consistent with my model of the dynamics here. We never got our slack back from when we gave it up, because that would mean giving up permanently on patients.



**John Burn-Murdoch** @jburnmurdoch · 1h

Another recurring point made by those on the front lines is that the pressure on emergency departments is immense.

Data from London is stark: last year A&E attendances plummeted as many people avoided hospitals. This year Covid comes on top of very high winter A&E pressures.



**John Burn-Murdoch** @jburnmurdoch · 1h

And even though London's ICUs have far fewer Covid patients this winter, there is still significant disruption:

The need to keep patients with Covid away from those without means there is less capacity for non-Covid care this winter than there would otherwise be.

While ICU is under less pressure from Covid this year, the rest of the hospital is facing more.

One critical care consultant told us they may end up having to redeploy staff from ICUs to other wards.

Not so much less pressure as different pressure

Here's the summary from the linked thread above.

**Chris Hopson** @ChrisCEOHopson · Jan 1

SUMMARY. Pressure and trust leader concern mounting as hospitalisations increase. Next few days crucial. London hospitalisation rates could plateau echoing community infection rates as we hit Xmas. Or they could rise on the back of intergenerational Xmas mixing...23/25

1 73 370

**Chris Hopson** @ChrisCEOHopson · Jan 1

So we still need to watch and wait. Govt must be ready to introduce new restrictions at pace if they're needed. Positive news is that we are still not seeing the large numbers of severely ill older people that we have seen in previous peaks. Vaccine effect very pronounced....24/25

27 82 448

**Chris Hopson** @ChrisCEOHopson · Jan 1

NHS preparing for worst & hoping for best. Staff flat out, esp. given level of staff absences. We will need to ask them to perform flexible heroics again if hospital covid numbers continue to rise. We can't keep doing this. Long term NHS capacity issues must be addressed. 25/25

This seems right. Hospitals are in a bad way but holding. The scary scenario is if cases among the elderly are only now picking up and severity is much less reduced in that group while also starting out at a higher level, and things then get much worse. Cases are accelerating full blast and will peak very soon in the first-hit areas like New York and London if they haven't already, but it will take another cycle or two before the effective risks start declining.

[Hopson then had another updated thread on 3 January](#), which gives more hope that the scenario can be handled. Situation continues, and the problems in London are being seen elsewhere as well, but London isn't getting that much worse.



**Chris Hopson** @ChrisCEOHopson · Jan 3

The challenges trusts are reporting to us centre around three familiar issues. Growing number of covid patients. Rapidly increasing staff absences, as covid infection rates rise in local communities. All against backdrop of huge wider pressure in the health and care system....7/25

4

77

310



...



**Chris Hopson** @ChrisCEOHopson · Jan 3

Wider pressure: Busy emergency care pathway. Significant numbers of planned care cases that cannot be delayed any longer. Complex and resource intensive extension of booster campaign. Big increase in pressure on social care due to staff absences/care home outbreaks 8/25.

...

The news is good. The center is holding:



**Chris Hopson** @ChrisCEOHopson · Jan 3

But growth rate over last few days dropped significantly. Number of patients with covid in London hospitals grew by 5% on both 30 and 31 Dec and then 1% on 1 Jan and 2% on 2 Jan. IF (a very big if) this much lower growth rate continues over next few days, growing optimism....17/25

6

52

316



**Chris Hopson** @ChrisCEOHopson · Jan 3

That London should be able to cope this week. But we need to recognise that "coping" under-plays degree of pressure trusts and their frontline staff are under. And the fact that, in many cases, trusts are having to delay planned care cases again, with impact on patients 18/25

3

43

359



**Chris Hopson** @ChrisCEOHopson · Jan 3

We must keep very close eye on London. Vital to see what effect New Year celebrations/schools return will have on infection rates. Definite worry that hospitalisations could rise again if infection rates rise again. Govt must be ready to move fast on restrictions if needed. 19/25

11

65

334



**Chris Hopson** @ChrisCEOHopson · Jan 3

Other piece of positive news is that hospitals are still not seeing large numbers of seriously ill older people. Striking that CEOs across country are echoing London colleague in pointing to the fact that care home omicron outbreaks not translating into hospital admissions. 20/25

18

171

620



**Chris Hopson** @ChrisCEOHopson · Jan 3

Increasingly clear, therefore, that the issue for NHS is not acuity/size of very ill older people covid caseload. But the number of staff absences and general admissions with covid on top of existing pressures. This is still stretching the NHS very significantly. 21/25



**Chris Hopson** @ChrisCEOHopson · Jan 3

SUMMARY Rest of country now under pressure. Some trusts declaring critical incidents to manage staff absences. Recent London data, fact that London/NHS currently "coping" & absence of large nos. of seriously ill offer grounds for optimism. But future still uncertain. 25/25

It's not over yet, and it's going to get worse before it gets better. [Similar story from Philadelphia](#) (see full thread for more details).



**M. Kit Delgado, MD, MS** @kit\_delgadoMD · Jan 2

...

Some thoughts from seeing a ton of Covid in the ER this week

Good news: Aside from a few patients with compromised immune systems, didn't have to hospitalize anyone who was vaxxed/boosted

Bad news: Health systems are stressed to the max

We need to adapt...quickly

Thread



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

...

...First, this Covid wave is different and we need to be smarter about how we interpret Covid cases that present to the ER and subsequent hospitalizations numbers (which are through the roof). They fall into three categories and can be managed differently...

1

32

275



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

...

- 1) Hospitalized for Covid: mostly unvaxxed or elderly/chronic ill & unboosted
- 2) Covid exacerbated a chronic illness (e.g. COPD) aka "The Covid Tip" (h/t [@jeremyfaust](#))
- 3) Incidental Covid (caught on screening) happens to everybody (maybe ~1/5 of hospitalizations I saw)...

7

63

364



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

...

...Based on increased immunity in the population, and this lower severity strain of Covid, less people are presenting with a need for ICU level care.

But a HUGE number need supportive care over 1-3 days (e.g. IV fluids, brief oxygen support, optimization of chronic illness)...

1

100

435



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

...

...Second major difference in this wave is there a huge number of patients coming to the ER for testing with mild symptoms or even no symptoms because of the lack of available testing or timely outpatient evaluation capacity elsewhere...

If I needed a test, and didn't have much in the way of symptoms, I'd stay *the hell* away from the ER right now. That seems like a very good way for you to test negative today and then get your symptoms two or three days later. We need to do better warning essentially healthy people to keep their distance, not only for the system's sake but for their own. This also seems like an excellent prescription if we are capable of doing it:



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

...

Replying to @kit\_delgadoMD

...So what can health systems do over the next 4-6 weeks to deal with the surge?

1) Set up Covid testing and flu like illness clinics outside the ER to divert low severity patients from ERs to more convenient care and allow ERs to focus on non Covid and Covid emergency care...



6



116



754



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

...

...2) Optimize care pathways in the hospital. We have shown the benefits of triaging lower severity patients (who are now the majority of hospitalizations) to dedicated Covid observation units. Better care, less time in the hospital. See article below...



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

...

...3) Dialysis centers, behavioral health treatment providers, and long term care providers need to adapt to the reality that a huge proportion of patients who need their care will incidentally test positive for Covid. Decreased access to care will worsen outcomes, and...



3



47



372



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

...

...and with staff in these facilities almost universally vaccinated and everyone wearing masks, the risks from Covid are no worse than any other aspect of society (family gatherings, public transportation, shopping, dining, bars, etc)...



6



23



277



M. Kit Delgado, MD, MS ✅ @kit\_delgadoMD · Jan 2

...

...4) Finally, implementing a low-tech text-message based monitoring program can help the millions of patients who will test positive stay safe at home and get to the hospital at the right time if they get sick. See our work below...

[Link to that article.](#) A lot of this is much worse due to a failure of professionals to talk logistics. I'm guilty of this too, in the sense that I don't focus on forward-looking logistics often enough. Long term I need to remember to talk logistics more. After all, as Hunter S. Thompson pointed out, when the going gets weird, the weird turn pro.



M. Kit Delgado, MD, MS @kit\_delgadoMD · Jan 2

...

...Lastly, the sheer number of people testing positive with mild or no symptoms means there's a large population of patients with serious health needs that are having challenges accessing care because of their Covid + status...



1



52



366



M. Kit Delgado, MD, MS @kit\_delgadoMD · Jan 2

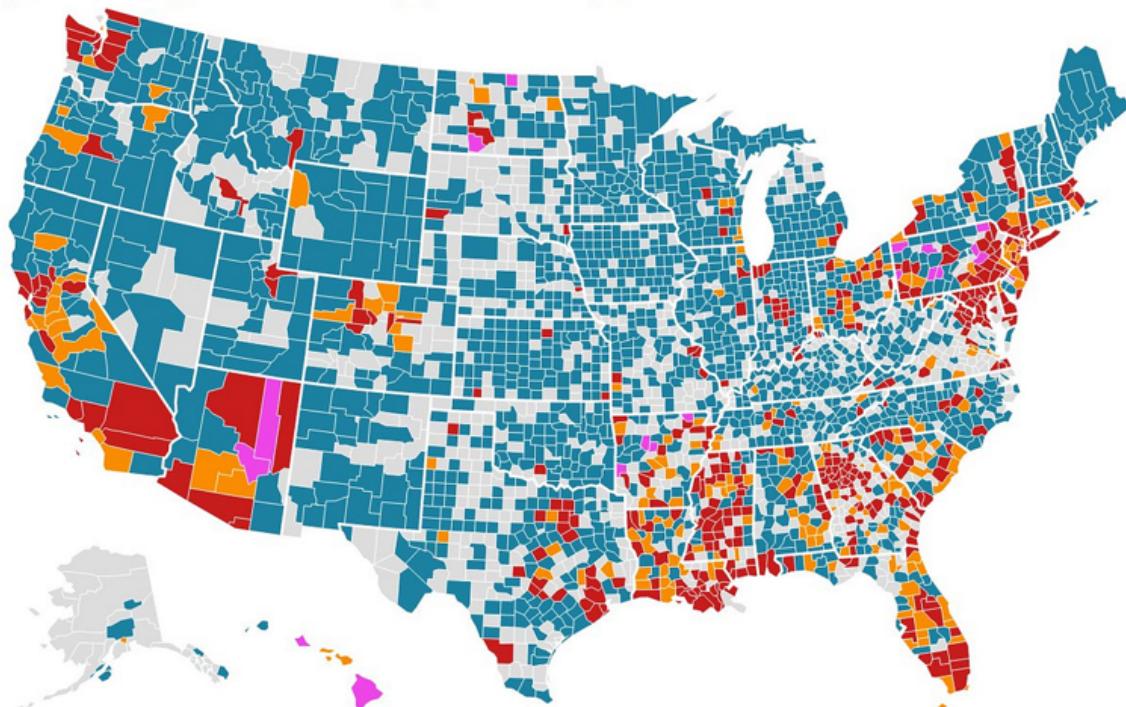
...

...Dialysis patients report difficulty getting outpatient dialysis because they tested positive. Many psych crisis centers, detox units, homeless shelters, and some nursing facilities won't accept Covid+ patients. This leads to extra time in the hospital to meet these needs...

## Circuit Breaker Status - County

Counties that have hit the circuit breaker are at high risk of overwhelming hospitals in the next 1-10 days given current case counts and hospital capacity. View the circuit breaker dashboard at [alexanderjxchen.github.io/circuitbreaker/](https://alexanderjxchen.github.io/circuitbreaker/).

■ At Capacity ■ Hit Circuit Breaker ■ Unsustainable ■ No Breaker



At Capacity - Hospitals are projected to be at capacity given a Nowcast with recent COVID-19 case trends (assuming non-COVID care is stable); Hit Circuit Breaker - Given recent COVID-19 case trends, counties are at high risk of overwhelming hospitals in the next 1-10 days; Unsustainable - The single day case count yesterday is on a trend to the area reaching a circuit breaker; No breaker - Hospital capacity not at high risk of being exceeded

Map: Jeremy Faust, Benjy Renton

There's a lot of population in the red and yellow areas here, but it's not about whether you go over, it's about *how far* you go over, and how bad the consequences will be for how long.

[Here's a thread about the situation in New York City.](#) A lot of similar observations. I'm worried about this observation, which is the flip side to other places where it's proving logically difficult to administer care due to the need to avoid it:



**Craig Spencer MD MPH** @Craig\_A\_Spencer · 10h

...

What's also different now is those COVID cases are often in beds next to patients who've done everything to avoid the virus, and for whom an infection might have a dramatic toll.

The cancer patient on chemotherapy.

Those immunocompromised or severely sick with something else.

27

491

2.6K



Overall, same patterns.



**Craig Spencer MD MPH** @Craig\_A\_Spencer · 10h

...

During the first surge, COVID was the only thing we saw in our ERs.

Now record-number COVID cases are hitting at a time when our ERs are already seeing extremely high numbers of non-COVID patients too.

Thankfully the Covid patients aren't as sick.

BUT there's SO many of them.



**Craig Spencer MD MPH** @Craig\_A\_Spencer · 10h

...

Overall it seems Omicron causes milder disease. We aren't seeing as many patients gasping for air.

But there's just SO much of it and it's impacting patients in different ways.

So even if just a tiny portion of cases need to stay in the hospital, it can turn into a huge influx.



Craig Spencer MD MPH ✅ @Craig\_A\_Spencer · 10h

...

I felt sick of wearing that damn N95 and face shield for 13 hours, again, after I hoped that phase was behind us.

The next few weeks will be really really tough for us.

A lot of healthcare workers will get sick.

We will have to work short-staffed and take on more patients.

16

339

2.8K



Craig Spencer MD MPH ✅ @Craig\_A\_Spencer · 10h

...

If you do need to go to the ER, try to understand we are doing our best.

If you haven't been vaccinated or boosted yet, now is really the time. It makes a difference.

I know you're tired of this.

We are too.

But we'll really need everyone's help to get through it, again. 🙏

Given how much downside risk there was initially, I'll take this level of pressure.

It's going to suck, a lot. But it looks like we'll mostly make it through. If it was going to be a complete collapse, we would know by now.

## Threads

[Could we be looking at the Pi variant in France? This reply convincingly says no, this one is less kind](#), this isn't new and if it had anything like Omicron's infectiousness then we would know. Level of concern here very low.

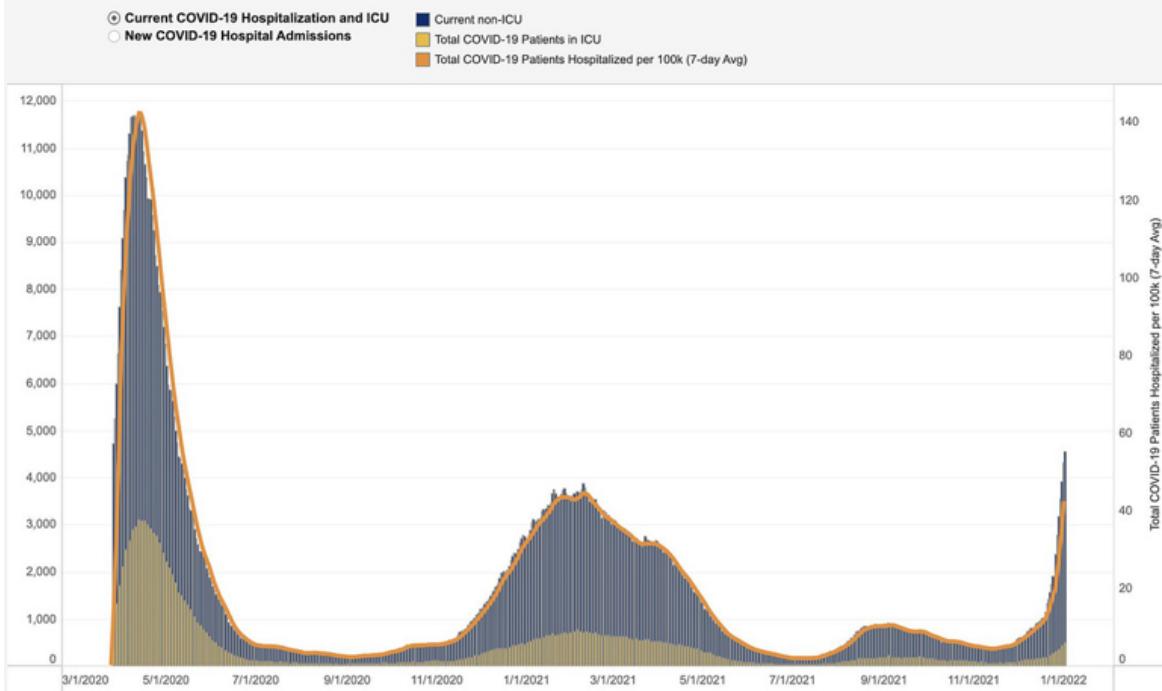
[This is what the quarantine looks like in Xi'an](#), where you're not allowed to go outside for any reason so you're dependent on government food deliveries, and they disinfect the streets anyway. This seems like Maximum Containment and it's at best going to be close. This doesn't seem sustainable.

[Zeynep thread pointing out that exponential math eventually causes massive disruptions no matter what](#), in context of NYC.

[Gottlieb predicts the wave will take two months, that NYC will peak within two weeks, most other areas within four weeks](#). If anything I expect it to be faster.

The question is how much more exponential growth is left before it's over, and how much we should be focused on the hospitalization number versus the ICU number (because potentially we have a ton of people 'with Covid' instead of being there 'for Covid'.)

## New York City



[Thread calling for more effort to slow spread on the basis of 'millions being crippled by Long Covid'](#) and claiming that getting Covid now makes you *more* susceptible in the future by 'creating pre-existing conditions.' This is very much not my view, but if I did believe that Long Covid was this bad, the rest would indeed follow. I don't see any evidence cited here that would cause me to update on my view of Long Covid.

### Prediction Updates

[Adding this, from Nate Silver \(here's Metaculus\):](#)



Nate Silver @NateSilver538 · Dec 31, 2021

...

Replies to @NateSilver538

A fun question for prediction markets might be: will there be a federal mask requirement in place on domestic flights as of Nov. 8, 2022? (i.e. Election Day.) I think I'd make 'No' about the 3:2 favorite, but wouldn't put a lot of money on it.

63

33

403

↑

I hope soon to be able to expand my work with prediction markets, and also we'll have everyone's 2022 predictions to play off. Going to be a fun ride.

### Will There Be a Federal Mask Requirement on Domestic Flights as of November 8, 2022? %? → 50%.

Metaculus has the median for lifting this requirement in August, which roughly agrees with Nate. I'm more optimistic about *not needing* a mask mandate on planes, but also more skeptical that it will get lifted. [Is your tray table up and your seatbelt in the full upright](#)

position? Have you met the TSA and their machines that they assure us *no longer* cause cancer (and that I still opt out of)?

Since this is a slow-moving one, I will only update it every so often. If you're curious where I'm at, you can give me a nudge, but this should stay roughly here for a while.

**Chance that Omicron has a 100% or bigger transmission advantage in practice versus Delta: 60% → 55%.**

I messed up the math last time and said 70% before adjusting to 60%, but our starting point this week is 60%. On reflection, after adjusting for early advantage and looking at when things are stabilizing, my guess is that this is going to be close, and we may never know for sure.

**Chance that Omicron is importantly (25%+ in the same person) less virulent than Delta: 85% → 90%.**

**Chance that Omicron is vastly (75%+ in the same person) less virulent than Delta: 20% → 25%.**

The news out of the hospitals seems clear. Unless something very strange is happening, Omicron is substantially more mild than Delta. I don't expect us to hit the 75% threshold, but the news keeps being good on that front, so for now I'm creeping a bit higher.

**Chance we will be getting boosters modified for Omicron within 6 months of our previous booster shot: 20% → 20%.**

Israel is going with their fourth shot, but otherwise things seem to be settling into full complacency. I didn't define who 'we' is here, but I meant the broad 'we' as in I personally have the legal option to get one, I'll be 43 years old and technically have a risk condition but am not immunocompromised. Note that even if I'm *allowed* to get one, I don't *intend* to get one if the peak has already passed.

**Chance we are broadly looking at a future crisis situation with widely overwhelmed American hospitals, new large American lockdowns and things like that: 10% → 5%.**

We've made our position clear. It's going to be difficult to get care of any kind for a month, but we've picked 'muddle through' and we're sticking to it.

**Generation time (serial interval) of Omicron is 3.5 days or less: 85% → 87%.**

This continues to be the best fit of the data and I'm creeping the number up a bit, but my guess is we will never find out for sure (unless it's super not close) because experimentation remains illegal.

**Daily cases in the United States have peaked by February 1, 2022: ?% → 85%.**

This seems like more than enough time to be on our way back down, but you *can* sustain a lot of cases for a long time. Feels right to express my degree of certainty about this. I'm curious how confident other people are on this.

I'll open the floor to suggestions. What would be the most high-value things to put in the predictions section at this point? Links to Metaculus are always a bonus, or better yet real-money prediction markets. Mostly I plan to wait until I see the other 2022 prediction threads, and play off of those, like I have in the past.

# **Activated Charcoal for Hangover Prevention: Way more than you wanted to know**

Activated charcoal is charcoal that has been heated up a whole lot and then washed really hard, by water or acid. [This process](#) is what "activates" it, and activating it increases the surface area by a lot. I think it prevents hangovers.

I won't spend too much time in this post arguing that activated charcoal works. Instead, I'll mostly take for granted that it does, then take a tour of the various things that might cause hangovers, stopping at each to look at why activated charcoal might or might not be likely to reduce them. So even if you're sure activated charcoal is just a hokey health craze thing I've tricked myself into believing, this post still has lots of interesting hangover science you might not have heard about.

I used to suffer vicious hangovers pretty often - usually a headache and sometimes nausea - after nights of drinking with friends. This past spring, I started taking two 520mg capsules of activated charcoal before going to bed on these nights. Since then, I've had maybe 26 nights out with the level of drinking I'd expect to result in hangover, and had a headache twice and nausea just once. (Tiredness is still common). Based on past experience, without activated charcoal I would have had headaches 12-18 of these times and nausea 4-10 times.

I recently convinced a friend to try a capsule before going to bed after a late night, after she'd insisted for months that there was no way this would work. The next morning she sent me a rave review, saying she'd expected to feel terrible after the number of drinks she'd had, but just felt a little tired. She's tried it a couple more times since and it seems like it worked those times too. I had another friend try it once and he said he thinks there was a 60% chance it helped; another friend's husband says it worked twice and didn't work once. I originally tried the stuff after I found [this article](#), by a woman who went to a bachelorette party. They were split into two hotel rooms, and she gave activated charcoal to the women staying in her room, but not to anyone in the other one. The next day, those in the other room were in much worse shape.

The idea that activated charcoal helps hangovers doesn't have much support in the first few pages of google results. If I'd stopped at those, I wouldn't think it does anything. But my personal experience with it is striking. So I'm confused. Let's look into it.

(A quick warning: Some sellers of activated charcoal say to take it every day. I don't think there's any reason to do this, and some people say it's bad to. Also, if you take medications, activated charcoal will make them less effective; this could be pretty bad, depending on how important the medications are to you. Don't take activated charcoal every day.)

## **Why might activated charcoal prevent hangovers?**

I mentioned activated charcoal (charcoal for short) has a lot of surface area. From US Poison Control:

Activated charcoal is a special form of carbon that can bind other substances on its surface (adsorption). That is why it is used for water filtration. It is also used to adsorb drugs in the gut so the drugs don't enter the body. Activated charcoal is made by burning

carbon-rich materials, such as wood, at very high temperatures to create charcoal. The resulting product is a black, odorless powder. The charcoal is then "activated" through a special process that makes lots of holes and crevices on the charcoal particles to increase surface area and available binding sites. One teaspoonful of activated charcoal has about the same total surface area as a football field!

They also [share](#) a story, back from the arsenic-drinking days of 1813, of a chemist drinking 5 grams (read: a whole lot) of arsenic trioxide mixed with activated charcoal and being fine.

So activated charcoal puts a whole lot of surface area into your body. All kinds of molecules are floating around in your body, and when they bump into the little granules of charcoal powder, they bind to them. Whether or not a particular molecule binds depends on its chemical properties. These molecules include those that might be poisoning you at the time. If a molecule does bind to the charcoal, it can no longer pass from the gut into the blood, because charcoal granules are too big. Instead, the toxin is stuck to the charcoal as it slowly makes its way through the gut, and out through poop.

Activated charcoal is [used in emergency rooms](#) to help victims of poisoning<sup>[1]</sup>. (But not alcohol poisoning!) [One source](#)<sup>[2]</sup> says that it's a poor treatment for alcohol poisoning because alcohol is too rapidly absorbed by the gut; [another](#) says it's poor because ethanol doesn't bind to AC as much as other poisons do; [this one](#)<sup>[3]</sup> says that that low binding is due to the polarity of alcohols.

## Does activated charcoal reduce blood alcohol levels?

After long nights of drinking, I usually take 1040mg of activated charcoal just before bed. This is usually around 2am, when my first drink might have been at 7 or 8pm. I have the sense these days to wind down my drinking as the night nears its end, so I usually do not even feel drunk when I take the charcoal. However, many of the studies on charcoal and alcohol give people charcoal right after giving them alcohol, and then measure their Blood Alcohol Content. I don't want to skip these, so let's spend some time on how charcoal interacts (or doesn't) with BAC.

I looked at 4 or 5 studies on humans that gave the charcoal around the same time as the alcohol, waited a bit, then measured BAC. They all report not finding any lower BACs in people who took charcoal. Most were done in the 80's and are paywalled and were on less than 10 people, so I might say that there was plausibly an effect but the studies were underpowered. Except [this one](#)<sup>[4]</sup> measures slightly *higher* ethanol levels in the people who got the charcoal (p-value 0.08). I don't think charcoal actually raises ethanol levels, but the finding is at least evidence against a small-but-not-significant effect of charcoal on ethanol blood levels. That study also involved ludicrous amounts of the stuff - 60 grams of activated charcoal. That's 60 times what I take!

Then to confuse things, [this study on dogs](#) finds an effect, saying "absorption of ethanol was significantly inhibited by activated charcoal during the first hour after administration. Blood ethanol levels remained significantly lower throughout the study"<sup>[5]</sup>. [This one](#) finds the same in dogs and rats, though it's more complicated because they tried a couple different delivery methods and some worked and some didn't<sup>[6]</sup>. I don't see any reason ethanol could bind to charcoal in dogs and rats but not humans, so I don't get this. Maybe there's something going on here, but no one can find it happening in humans, so action on ethanol probably isn't why activated charcoal works.

I mentioned earlier that top google results for searches like "does activated charcoal prevent hangovers" say it doesn't. Most of them are making the basic case I made here: they looked up whether activated charcoal binds to ethanol, found studies saying it doesn't, and rested

their case. But this case is too basic. BACs and hangovers correlate - people who get more tanked have worse mornings - but there's a lot more going on with hangovers than just BAC levels. I'll get into some of the other compounds you get from alcoholic drinks next, and try and figure out whether activated charcoal might be acting on any of them. But first: if I take charcoal before bed, an hour after my last drink, does it even get a chance to meet those compounds in the body?

## Can activated charcoal even meet alcohol when not taken at the same time?

Some digestion happens in the stomach, but most of it happens in the small intestine. If I take charcoal an hour after my final drink, the night's alcohol is mostly gone from the stomach already, so they don't meet there. By the time the charcoal reaches the small intestine, the final drink has probably finished its initial pass through there too.

Most ethanol metabolism happens in the liver, but same problem: the alcohol doesn't pause in the liver, politely waiting to move on until the activated charcoal catches up. So what gives? How can they ever meet? This goes for all drugs, not just alcohol - how can charcoal ever help with anything, unless it's taken right away?

I was pretty confused by this, but it turns out the body digests drugs via [a cyclic process](#). From what I can tell, some volume of the drug is processed in the first pass through the liver, but if there's too much to process in one pass, the substance gets put back into the small intestine for another go-around. This explains why [this source](#) can recommend charcoal as possibly beneficial up to 4 hours after large ingestions of some toxins: the toxin or its metabolites might still be going through the cycle<sup>[3]</sup>. If you're drinking a fair amount of alcohol, you'll need multiple go-arounds through this cycle of liver->intestine->back-to-liver to fully process it.

*Side note: a quick review of the alcohol metabolism pathway here, to make the next paragraph make sense: Ethanol is converted into acetaldehyde (AcH), then AcH is converted into acetate, then acetate is converted into carbon dioxide and water. The ethanol->AcH and AcH->acetate steps are mostly done in the liver. I'll use the words "byproduct" and "metabolite" to refer to these intermediate chemical steps.*

So. Say I'm out drinking into the night - enough drinks to be at risk of hangover - and I get home at 2am and take activated charcoal. That liver->intestine->liver cycle is still going on: some ethanol has been converted into AcH, some AcH has been converted into acetate, and some acetate has been fully broken down into CO<sub>2</sub> and water. The metabolites that the liver hasn't finished with are being periodically deposited back into the small intestine for another go. Charcoal makes it into the small intestine an hour or two after ingestion. So they meet!

But for how long? Things that don't get absorbed into the blood, like charcoal and food, spend hours in the stomach, then more hours in the small intestine. I did some squinting at a couple sources to try and nail down how long. Some doctor with Cleveland Clinic says it takes [40 minutes at quickest, up to 3-5 hours if you eat a lot](#), for food to leave the stomach; a paper says it takes [an hour for just 10%](#)<sup>[7]</sup> of food to pass the stomach (though no way it takes 10 hours for all of it to pass); someone at Mayo says it takes 6-8 for things to get through the stomach and the small intestine together; and another doctor says it could be [7 to 11](#). These are all vaguely in line with each other, so I'm fine with taking 6-8 hours as a first approximation for the time from ingestion to small-intestine-exit.

So, roughly, there's a 2 to 5 hour window where the charcoal is in your small intestine (6-8 hours minus the time it spends in the stomach), where it can meet the alcohol or alcohol byproducts. Some digestion happens in the large intestine, too, and stuff spends days in there, so maybe charcoal spends time with some of alcohol's byproducts there too.

But since charcoal doesn't do much to ethanol, knowing that charcoal and ethanol meet in the intestine isn't enough to tell us why charcoal might prevent hangovers. Who cares if they meet, if they don't bind? So let's look at the things ethanol breaks down into, as well as the other substances that come in alcoholic drinks, in search of something that binds.

## On alcohol byproducts

Since you can go to bed at 3am and still be hungover at noon, the presence of ethanol probably isn't causing hangover directly - it's all been metabolized by noon! Plus, lots of ethanol in your system makes you feel drunk, not hungover; and being hungover doesn't feel like being drunk. So it must be something else. Our main suspects in this section will be the two byproducts of ethanol metabolism - acetaldehyde and acetate - and also methanol, since alcoholic drinks have small amounts of methanol in them. In high enough doses, methanol is definitely terrible for you; acetaldehyde is probably pretty bad; and acetate has been linked to headaches.

### Acetaldehyde

Everyone agrees that acetaldehyde (AcH) is pretty bad for you. Some evidence AcH is pretty bad: acetaldehyde dehydrogenase (ALDH) is the enzyme that breaks AcH down into acetate. If you give someone a drug to temporarily stop ALDH from working, then give them alcohol, they will probably be violently ill. The reason they get sick is because AcH builds up in their body. (Disulfiram is a drug that does just this, and is used to treat alcoholism by making alcohol just horribly unappealing.)

ALDH is involved in "Asian Glow" too - people that get Asian Glow get it because of problems with their ALDH enzyme. I don't think people with Asian Glow suffer near as badly as people on Disulfiram; I asked a friend who has it, and she said her face turns red after a drink or two but she doesn't feel sick at all. Probably most Asian Glow sufferers have an ALDH enzyme that's less efficient, but still mostly works. It could also be that each enzyme works fine, and just not enough of them are produced - similar to how lactose intolerance is usually caused by not having enough lactase.

Everyone says that in normal people (not on Disulfiram or anything), AcH is converted to acetate pretty darn quickly. By morning, there isn't any left. There might not even be much left by the time I take activated charcoal at the end of the night. So the timing doesn't line up here. Also, AcH doesn't really bind to charcoal. [This paper](#)<sup>[8]</sup> tried using activated charcoal in the creation of alcohol, adding it to a "barley malt-based agricultural distillate". Of everything they tested, acetaldehyde had one of the lowest binding affinities to charcoal.

Acetaldehyde is pretty interesting, and a good thing to target if you're trying to prevent hangovers; more about it in the ZBiotics section below. But no way is activated charcoal working on it, so moving on for now.

### Acetate

[Acetate causes hangover headache in rats!](#)<sup>[9]</sup>

I'm not sure how you measure a rat headache. It sounds like the authors poked the rats in the face using a special rat-face-poking machine, and measured how strong a poke had to be to make the rat react.

Unlike AcH, which is converted to acetate quickly, acetate hangs around in the body for a while. The paper says that acetate levels are up for at least six hours after ethanol consumption, though I can't find any statement that direct in their citations for the claim. The whole quote is useful:

Although... acetaldehyde levels are low and change minimally following ethanol intake, **acetate levels are significantly elevated for at least six hours**. Studies suggesting that acetaldehyde is responsible for hangover have not considered acetate, which increases to much higher levels in the circulation, even after moderate drinking, compared to acetaldehyde levels. Acetate alone may induce headaches at these elevated concentrations. This is supported by the clinical observation that **acetate given during kidney dialysis causes headache**.

These authors say a couple things throughout the paper suggesting they don't think AcH is very important in hangovers, though they don't go all-out, instead saying it's possible that AcH matters in addition to acetate:

These data do not completely exclude the role of acetaldehyde in hangover... Acetaldehyde may still cause hangover symptoms in humans including headache. The data in this study demonstrate that acetate also contributes to the headache component of the hangover.

Since Asian Glow symptoms are caused by AcH build-up, I don't think AcH is off the hook either.

## Tangent: Acetate and caffeine

[This paper](#)<sup>[10]</sup> has more bad things to say about acetate:

Ethanol has been shown to increase markedly portal blood flow, primarily by increasing intestinal blood flow. This effect of ethanol is reproduced by acetate, infused at rates equivalent to those leading to endogenous acetate production following ethanol administration... adenosine is also known to increase markedly intestinal and portal tributary blood flow. We have shown that adenosine receptor blockade with 8-phenyltheophylline completely abolishes the effects of ethanol, acetate, and adenosine on intestinal and portal blood flow.

**... acetate administration has marked effects on central nervous system function.** On two tests, motor coordination and anesthetic potency, both ethanol and acetate showed similar effects. **The effects of acetate were fully abolished by 8-phenyltheophylline...** Whereas ethanol at low doses increased locomotor activity in mice, acetate markedly reduced it.

No clue what 8-phenyltheophylline is. The important thing is it *blocks adenosine* - just like coffee - and this improves symptoms! Also, the rat-headache people [found](#) that the rats they gave caffeine had less severe headaches. They say "acetate increases adenosine in many tissues, including the brain", so their model is basically that acetate causes adenosine, which causes headaches.

Caffeine doesn't have anything to do with activated charcoal, but people have been into treating hangovers with coffee forever: In 1768, the writer William Hickey wrote a whole lurid description of having a hangover. It's good stuff:

My first return of sense or recollection was upon waking in a strange, dismal-looking room, my head aching horribly, pains of a violent nature in every limb, and deadly sickness at the stomach. From the latter I was in some degree relieved by a very copious vomiting. Getting out of bed, I looked out of the only window in the room, but saw nothing but the backs of old houses, from which various miserable emblems of poverty

were displayed . . . At that moment I do not believe in the world there existed a more wretched creature than myself. I passed some moments in a state little short of despair . . .

Later in the same essay, he says "very strong coffee proved of infinite benefit"<sup>[11]</sup>. Sadly this doesn't work for me at all for some reason, but the acetate->adenosine angle gives a mechanism by which coffee might reduce hangovers.

## Does activated charcoal bind acetate?

[Wang 2005](#) looks at a bunch of different acetates (ethyl acetate, methyl acetate, butyl acetate...) and sees how much they're absorbed by different materials<sup>[12]</sup>. This is again a study about air filtering, outside of any biological body, but they say that "In general, activated carbon has a greater affinity for [these acetates] because of the nonpolar nature of its surface compared to other solid sorbents."

I'm not sure if the type of acetate matters here. Even [this paper](#) specifically on alcohol metabolism just calls the thing that AcH gets turned into "acetate"<sup>[13]</sup>. Acetic acid is really similar to acetate, and [here's a source](#) that says activated charcoal is good at binding acetic acid, so... that's suggestive?<sup>[14]</sup> But me not knowing much chemistry is starting to be a problem here, and I don't know whether the difference between acetic acid and acetate would cause changes in the effectiveness of charcoal.

I also think this [extremely confusing abstract](#) is saying that activated charcoal is good at absorbing ethyl acetate<sup>[15]</sup>. This still feels shaky, since I'm not sure if we're all talking about the same kind of acetate here. Also, like, Wikipedia says "In nature, acetate is the most common building block for biosynthesis". Does the most common building block for biosynthesis really cause hangover headaches? Sounds weird.

## Methanol

Methanol is present in all alcoholic drinks. It's produced naturally by the fermentation process (the step in alcohol production where sugar is broken down by microorganisms). Methanol is also added in large amounts to make denatured alcohol, but that's not what I mean here: it's present in small amounts in all alcoholic drinks, and most can be distilled out, but it's hard to completely remove. Methanol is often bound to pectin - a sugar molecule common in fruit - and fermentation frees methanol from pectin. The rough alcohol-creation/methanol model I have in my head is "more fruit = more methanol". But there's a little bit of methanol present even in things that don't seem fruity at all - for example, Budweiser, and Vodka too - because all plants have pectin.

Methanol is higher in wood-aged liquors than clear liquors. I only have rough numbers here, but the worst fruit brandies [can have as much](#) as 4,000 mg of methanol per liter (mg/L) [this seems high though - don't anchor to it!], and Scotch whisky has 40-130 mg/L<sup>[16]</sup>. Vodka, much less: less than 2.5 mg/L in a survey of Brazilian vodkas, though higher in some vodkas in Eastern Europe - around 50 mg/L, and [this 1981 paper](#)<sup>[17]</sup> measured Smirnoff at 39 mg/L<sup>[18]</sup>. (The Smirnoff number is more relevant to American drinkers, but it was the 80's, so let's discount it to 25 mg/L for no good reason). It's [present in wine](#), and red wines have more methanol than whites or rosés<sup>[19]</sup>. Reds have 60-280 mg/L, and whites, 40-120mg. Popular light beers had 6 - 20 mg/L in that same 1981 study.

These numbers are hard to compare as-is, since if you're drinking beer, you're drinking a whole lot more volume than if you're taking shots. What we want is methanol content per liter of ethanol, instead of per liter of liquid. Say the typical liquor is 50% ethanol, wine 15%, beer 5%. Then we get these rough ranges:

- Scotch: 80-260 mg/L
- Vodka: 50 mg/L
- Red wine: 400 - 1850 mg/L
- White wine: 250 mg/L - 800 mg/L
- Light beer: 120 mg/L - 400 mg/L

(I'm worried about the beer numbers, because lots of soft sources online say beer has *no* methanol, and for some reason almost no scientists have ever published measurements of methanol in beer, so all I have is that one 1981 paper. I feel better about the other drinks, because I have a few sources for each that all hit around the same rough range. A friend who knows brewing points out that barley, used to make beer, has pectin, which explains how methanol can make it in.)

Methanol is bad for you because of its byproducts: "[the products of](#) methanol metabolism are extremely toxic and in high concentrations may cause blindness and death."<sup>[11]</sup> These two byproducts are formaldehyde and formic acid; it sounds like formic acid is the really bad one. There is a little bit of methanol in alcoholic drinks. It's so little that the papers I've seen so far just dismiss it; public health councils have looked into it a lot and don't think anyone is going to get proper methanol poisoning from properly-made alcoholic drinks. However, [for hangovers](#):

One specific congener implicated in hangover effects is methanol... Support for methanol's contribution to hangovers comes from several sources. **For example, distilled spirits that are more frequently associated with the development of a hangover, such as brandies and whiskeys, contain the highest concentrations of methanol.** Moreover, in an experimental study with four subjects who consumed red wine containing 100 milligrams per liter (mg/L) of methanol, Jones (1987) found that **elevated blood levels of methanol persisted for several hours after ethanol was metabolized, which corresponded to the time course of hangover symptoms.** Methanol lingers after ethanol levels drop, because ethanol competitively inhibits methanol metabolism. The fact that **ethanol readministration fends off hangover effects** may be further evidence of methanol's contribution to the hangover condition, given ethanol's ability to block methanol metabolism and thereby slow the production of formaldehyde and formic acid.

(Notice the last sentence, where the authors take for granted that "ethanol readministration fends off hangover effects", i.e. that hair of the dog works! Also that dark liquors are probably worse for hangovers than light ones like vodka.)

The reason ethanol fends off methanol metabolism is that they're both processed by the enzyme alcohol dehydrogenase (ADH), but ADH really prefers to do ethanol first, for vague chemical reasons. I'm not sure if this means that all the ethanol in your body has to be broken down before methanol metabolism begins, or just that methanol metabolism is very slow while ethanol is still around, but either way, most of methanol metabolism occurs after ethanol has mostly been broken down. So as you sleep after a long drinking night, the concentration of ethanol is going down. This allows more methanol to be metabolized, and it breaks down into formaldehyde, then formic acid. This means your peak concentration of formaldehyde and formic acid happens hours after you finish drinking, either in the middle of the night or the next morning.

So all over this post I've said (and will keep saying) stuff like "Methanol is important in hangovers", but I never actually mean the methanol is directly causing any of the symptoms - whatever effects it has are because of the formaldehyde and formic acid it turns into.

## **Does activated charcoal bind methanol, formaldehyde, or formic acid?**

Remember the paper from the acetaldehyde section that found it had one of the lowest affinities to activated charcoal? The other molecule they called out as having especially low binding was methanol. So direct action on methanol is out.

How about formic acid? [Adekola 2016](#) (also done outside the body - in an "aqueous solution") says yes! "It is therefore established that [activated charcoal] has good potential for the removal of [acetic acid] and [formic acid] from aqueous solution"<sup>[14]</sup>. [Another study too](#)<sup>[20]</sup>.

[Colgan 2016](#) suggests using activated charcoal in drug packaging to stop formaldehyde and formic-acid degradation of the packaged drugs<sup>[21]</sup>. I guess those two chemicals either leak in from the environment or are partly a result of the drugs themselves degrading over time. The study tried speeding up the drug-degradation process by some method, compared how fast drugs degraded with and without charcoal around, and found they degraded slower when there was charcoal present.

Yet another finds that formic acid binds "much slower" to activated charcoal than acetic acid. But I'm not sure if this means that acetic acid gloms on super fast, and formic acid binds at a decent rate, or if it means that charcoal is bad at binding to formic acid. So there's decent evidence that charcoal has some decent action on formic acid.

Recall that the chain goes methanol->formaldehyde->formic acid. So adsorbing formaldehyde might also help with the possible toxic effects of the methanol metabolism pathway. [Here's a company](#) describing an air filter with activated charcoal in it that can remove formaldehyde from the air: "[the design of our product] ensures the efficient removal and retention of formaldehyde while retaining an excellent physical adsorption capacity." [Tanada 1999](#) "aminates" activated charcoal - I don't know what that means (well, it means adding an amine group, but I don't know what *that* means) - and says that makes it effective at adsorbing formaldehyde<sup>[22]</sup>.

And it's not just them - most hits for searches like "activated carbon formaldehyde" return results about air filtering. Even trying "activated carbon formaldehyde in the body" returns air filtering results! Trying a little harder, I find that activated charcoal administration is [one of the CDC's recommendations](#) for someone who's taken in too much formaldehyde.

[Also](#) "Most researchers used activated carbon as an adsorbent for removal of formaldehyde because of its high adsorption capacity", though again with a focus on formaldehyde's gas form<sup>[23]</sup>.

For a lot of things, action in the air is different than action within the body - but the narrow question "does formaldehyde bind well to activated charcoal" seems like a question about very local chemical properties, so I'm going to take these results as evidence that charcoal can bind formaldehyde effectively in the body too.

OK. Summarizing so far:

- Activated charcoal has little or no direct action on ethanol or methanol.
- AcH, a byproduct of ethanol often implicated in hangovers, also is not much affected by charcoal.
- Acetate, another alcohol byproduct, might cause headaches, and activated charcoal might bind acetate.
- Methanol, present in small amounts in alcoholic drinks, has two toxic byproducts, and activated charcoal probably binds both.

I smell blood here. If methanol's byproducts really are implicated in hangovers, then we have our answer: charcoal binds its byproducts well, stopping them from being absorbed by the body, instead ferrying them out directly through poop. Above, I mentioned some reasons for thinking methanol is related to hangovers; let's go deeper into those reasons.

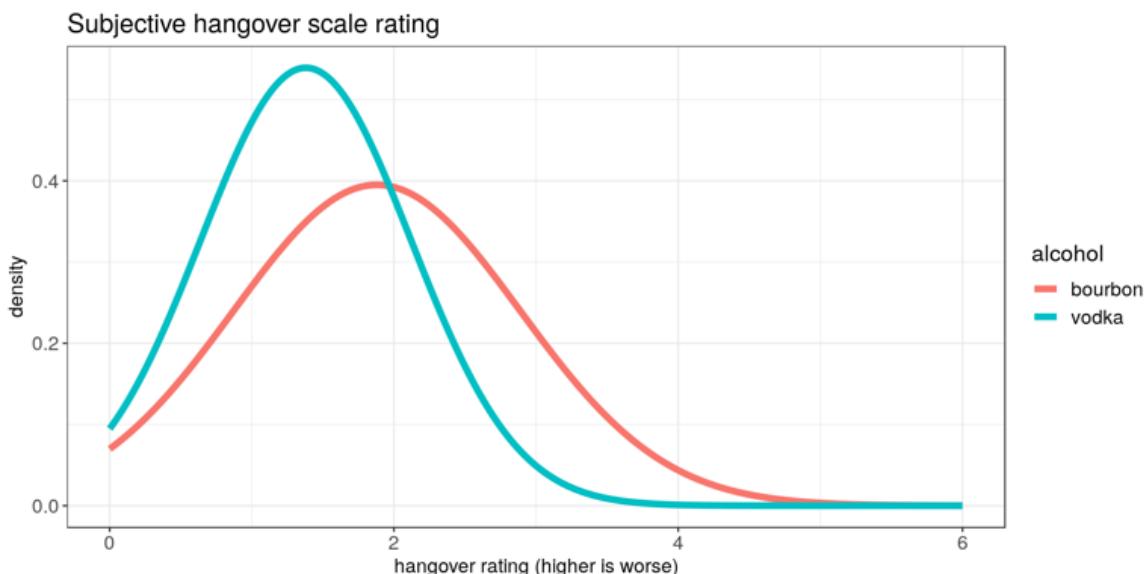
# More on methanol and hangovers

Methanol definitely isn't the *only* cause of hangovers - you can get a hangover from pure ethanol administration (though [my source<sup>\[11\]</sup>](#) doesn't cite a source). But looking at some hangover folk knowledge - things I've heard from friends and in random parts of the internet and in locations where alcohol is served, basically - gives hints that methanol might be involved:

## Folk knowledge #1: Clear liquors are less likely to cause hangovers than dark liquors

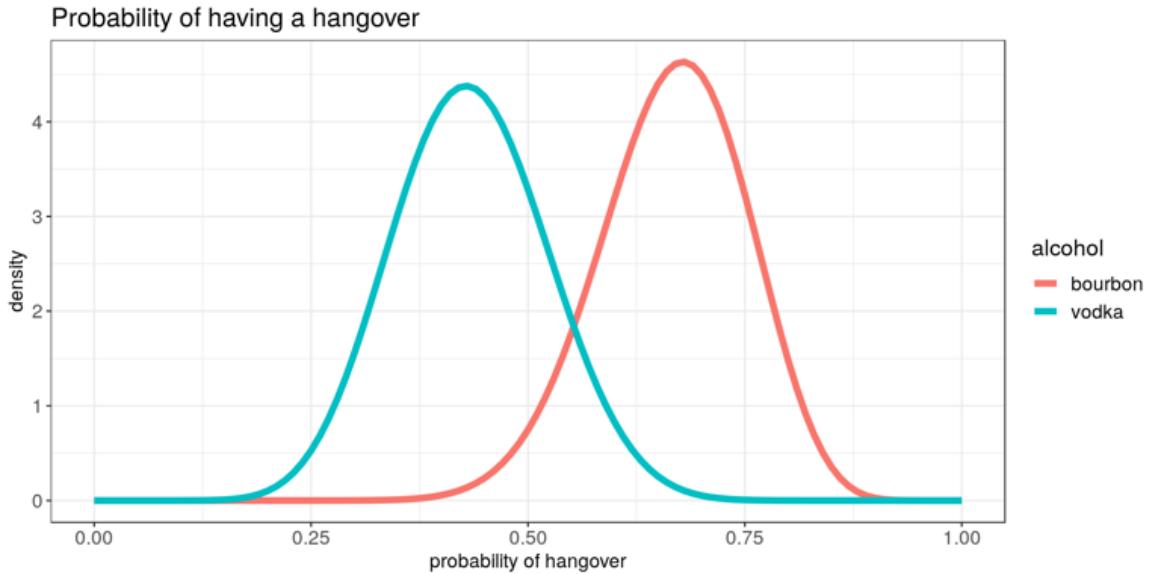
This one has to be true.

[Rohsenow 2010](#) gave bourbon to some people and vodka to others and compared how hungover they were by just asking them and rating their answers on some pre-established scale<sup>[24]</sup>. Their ratings on the different alcohols looked like this:



This graph is kind of jank because I assumed a normal distribution and just plugged in their means and standard deviations.

[Chapman 2010](#) just gave people bourbon or vodka and then asked if they were hungover the next day, Yes or No<sup>[25]</sup>. There were 30 people in each group; 13 of the vodka people were hungover, and 20 of the bourbon people. Those distributions look like this:



Bourbon definitely has more methanol than vodka, and bourbon causes more/worse hangovers than vodka.

And another:

[Pawan compared](#) the hangover produced by different types of drink (but only one brand of each) in his study of 20 volunteers<sup>[26]</sup>. The severity of hangover symptoms declined in the order of brandy, red wine, rum, whisky, white wine, gin, vodka, and pure ethanol. Vodka and pure ethanol caused only mild headaches in two volunteers.

## Folk knowledge #2: Sugary drinks make for worse hangovers

I never really believed this one, but I hear it pretty often. I haven't looked into whether there might be mechanisms by which sugar causes hangovers, but fruit juice has methanol in it. [Orange juice has around 100 mg/L](#) - fruit juices overall have [12 - 640mg](#) mg/L methanol, and the average juice has around 140 mg/L<sup>[27][28]</sup>. Scotch whisky was 40-130 mg/L, so a "shot" of orange juice gives you as much methanol as a shot of whisky. This means if you mix your alcohol with orange juice all night, you're doubling up your methanol content. (Hopefully you're not drinking whisky-orange-juices, but you know what I mean).

Often, sugary drinks are sugary because they have fruit juice in them, so maybe it isn't the sugar, but instead the extra methanol, that makes sugary drinks give worse hangovers. And the reason you don't get hungover while drinking orange juice by itself is that your body starts on the methanol metabolism right away, instead of letting it store up. The only time the methanol from fruit has a chance to build up is when you consume it at the same time as ethanol.

### Bonded methanol?

But actually, maybe most of the methanol in orange juice is still bonded to pectin. While it stays bound, it isn't an issue; this methanol only starts getting separated from pectin [once it reaches the colon](#)<sup>[29][30]</sup>. The colon is later than the stomach and small intestine breakdown phases I've been talking about for most of this essay - juice reaches the colon maybe 6 - 15 hours after ingestion. So this is another delayed-effect kind of thing going on with methanol.

## Folk knowledge #3: Hair of the dog works

Hair of the dog is the idea that having a drink or two while hungover makes you feel better. I mentioned this above too, but it's the biggest reason I think methanol is part of what causes hangovers, so let's go deeper into it.

[Jones 1987](#) says that hair of the dog works ("The most effective treatment for hangover, at least for short periods, is to drink more ethanol")<sup>[31]</sup>. [Swift 1998](#) takes him at his word (also, I'm going to start using "F&F" to mean "formaldehyde and formic acid"):

Jones has suggested that it is the metabolism of methanol to F&F that causes symptoms of hangover, with quicker methanol metabolisers suffering more. The justification for this suggestion is threefold: the types of drink associated with more severe hangovers contain higher levels of methanol; the time course of methanol metabolism corresponds to the onset of symptoms; and a small dose of ethanol, which blocks the formation of F&F, provides an effective treatment for hangovers ("the hair of the dog").<sup>[11]</sup>

Recall that the ADH enzyme is the main thing that processes both ethanol and methanol, but much prefers to do ethanol first. When methanol metabolism is inhibited because of this preference, its half life is long: [45 to 90 hours](#)<sup>[32]</sup>! When the ethanol is mostly gone, in the morning after a night of heavy drinking, methanol's half-life is just [2 to 3 hours](#).

Ethanol is so good at stopping methanol metabolism that it's [one of the main treatments for methanol poisoning](#)<sup>[33]</sup>. I think doctors still have to use other methods to get the methanol out once they've paused its metabolism; [Gonda 1978](#) recommends cleaning the blood with hemodialysis whenever a lethal dose of methanol has been ingested, while administering ethanol to give the patient more time<sup>[34]</sup>. Some sources act like once you administer ethanol, you've cured the methanol poisoning - [Wikipedia](#) calls ethanol an "antidote" - but I don't understand how this could be true. Won't the methanol still be around once the ethanol wears off?

Jones also measured methanol in people the day after drinking, and yup, it's still around:

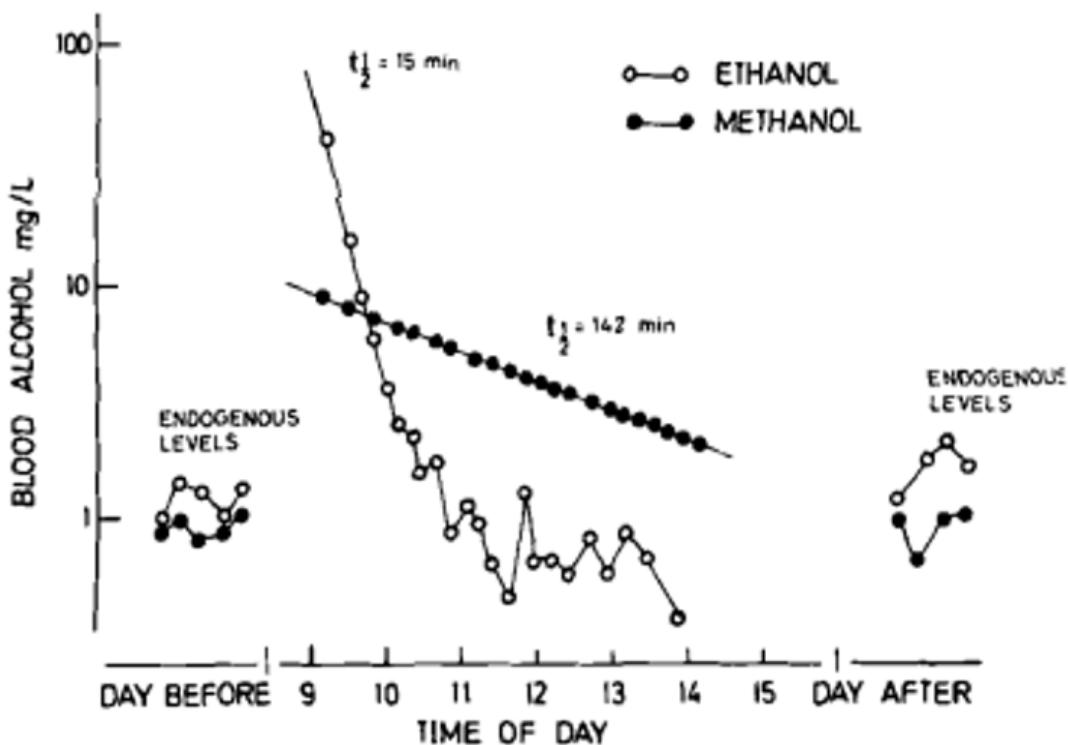


Fig. 1. Concentration-time profiles of ethanol and methanol elimination in one male subject during the morning of hangover. The concentrations of endogenous ethanol and methanol are also shown in tests made the day before and after the hangover study. Blood concentrations of these alcohols were estimated indirectly by analysis of end-expired breath.  $t_{1/2}$  = elimination half-life.

It's weird that there's so much ethanol left, relatively, at the beginning of the graph, and then it goes away so quickly. From some of the paper's language, it sounds like Jones might have intentionally started the graph at 100 mg/L blood ethanol, to investigate what's happening at low levels of these drugs. It's also weird that the methanol curve doesn't get steeper at around 10 or 11am - shouldn't the finishing-up of ethanol around that time free up ADH to metabolize methanol more quickly?

The reason for both of these weirdnesses: in Jones's view, ethanol levels under 100 mg/L is when ADH can really start working on methanol: "Below a blood concentration of about 100 mg/l, liver ADH is no longer saturated with ethanol as substrate and metabolism of methanol can therefore commence." So maybe the reason the methanol line doesn't bend is that you don't need *all* the ADH to metabolize methanol at the max rate, you just need enough of it, and "enough" happens at around 100 mg/L ethanol. I wish he'd measured the methanol levels before 9am, to see whether the methanol line is flat until that time.

If Jones is correct here, taking activated charcoal in the morning should be good enough, as long as you wake up early enough to take it before your ethanol levels get this low. It also might explain a strange experience I had earlier this year. I woke up at 9am after a reallllly heavy night of drinking - my heaviest in years, and since - and felt perfectly fine up until

around noon. I biked home! It was a beautiful day and I felt great; I felt industrious and started laundry. Then around noon, I got hit with one of the worst hangovers of my life and could barely walk. Unloading the dryer was a struggle. I'd never had a delayed-onset hangover like that before, and I could never figure out what caused it. Maybe it's because I drank so much the previous night that my ethanol levels didn't fall below the 100 mg/L dividing line until afternoon the next day? (This was before I'd ever taken activated charcoal. Actually, that day was so bad, it inspired me to start researching and trying to find hangover preventers. It was so bad, it eventually caused this post.)

### **Hair of the dog as a methanol pause-button**

Lots of soft-science sources, like [this Healthline article](#), warn things like "While [hair of the dog] may offer temporary relief, it only delays the inevitable, as the hangover will return once you stop drinking."

This might make sense if hair of the dog worked by alleviating alcohol withdrawal. But if it works by arresting methanol metabolism, then this warning makes less sense. Let's call this latter view the pause-button view.

In the pause-button view, when you wake up after a night of drinking and feel hungover, part of that is because you have F&F in your system. Taking another drink pauses methanol metabolism for the couple of hours it takes your body to process the new ethanol. While that's happening, your body is no longer making new F&F. Ethanol stops methanol from being converted by ADH, but since ADH doesn't do anything to F&F, focusing ADH back onto ethanol doesn't block or slow F&F breakdown at all. Your body gets a chance to clear them, without any more coming in. Their levels lower. Then, once the new ethanol has been metabolized, your system returns to metabolizing the remaining methanol, so your *peak* F&F level is lower than it would have been. This matches up with the ethanol-then-hemodialysis method of treating actual methanol poisoning.

### **Or as staving off withdrawal?**

Or maybe part of being hungover is just minor & temporary alcohol withdrawal.

[Several lines of evidence suggest](#) that a hangover is a mild manifestation of the alcohol withdrawal syndrome in non-alcohol-dependent drinkers. First, the signs and symptoms of hangover and mild manifestation of the AW syndrome in non-alcohol-dependent drinkers overlap considerably... Several [AW symptoms] also are usually present during a hangover, including nausea and vomiting, tremor, sweating, anxiety, headache, and sensory disturbances.

Second, Begleiter and colleagues (1974) present evidence that the hangover condition is actually a state of central nervous system excitation, despite the perceived sedation and malaise. Support for this view comes from the research of Pinel and Mucha (1980), which shows that single doses of alcohol decrease seizure thresholds in animals several hours later. Their finding indicates rebound excitation, a phenomenon noted to occur after short-term administration of some sedatives that can quickly clear the body, including alcohol and certain benzodiazepine drugs. [\[35\]](#)

Third, the observation that alcohol readministration alleviates the unpleasantness of both AW syndrome and hangovers suggests that the two experiences share a common process.

This sounds very reasonable and could totally be true. Also, I asked my friend who's tried hair of the dog and she said she feels the benefits almost immediately, which makes sense under this theory but makes no sense under the methanol theory.

(I should probably mention here that I've never done hair of the dog, that it feels to me like the most alcoholism thing ever, and does seem like the kind of thing that could condition you into real no-shit alcoholic dependence. If this section has convinced you that methanol's byproducts are bad and inspired you to try hair of the dog to reduce their levels, don't - instead, first try dealing with those byproducts by taking activated charcoal!)

## Folk knowledge #4: Cheap swill gives worse hangovers

Cheaper liquor is less distilled, so less of the methanol has been removed. Of course, less of lots of stuff has been removed - distillation removes all kinds of things. So if this item of folk wisdom is true, it implicates methanol more than it implicates AcH or acetate. But there are all kinds of congeners in alcohol that distillation removes, and it could be one of those making it so that cheap swill gives worse hangovers. So this points toward methanol a little bit too, but is pretty weak evidence by itself.

(However, after I wrote this, I talked to a friend who knows brewing, and he says that additional distillation steps probably don't have a big effect on methanol concentrations. This section still feels true to me, but maybe it's not.)

## Folk knowledge #5: It doesn't seem like people get hangovers in the middle of long drinking binges

I'm not sure if this is actually folk knowledge; it's just a vague feeling I have. I've never heard of people who pull all-nighters, drinking into the morning, being hit by hangover symptoms while still on their binge. I don't know enough people who drink through the morning often enough to have much of a sample size, but if continuing to drink does delay the hangover, that weakly suggests methanol is important in hangovers. (However, if this is true, staying awake *not* drinking should *not* delay the hangover, since it's only the continued presence of ethanol that delays methanol metabolism.) This would also fit with the alcohol withdrawal theory, though.

## Folk knowledge #6: Red wine gives bad hangovers

Red wine has more methanol than most other types of alcohol.

## A test for the methanol theory of hangover

Drink 3 or 4 shots of vodka in the span of 5 minutes, then eat as much fruit and drink as much fruit juice as you can. Different fruits have different methanol contents, so use the highest *free* methanol fruit you can. It's important that it be free methanol, because methanol bound to pectin has to wait until it reaches the colon to be freed up. Since it takes 6-8 hours for food to reach the colon, the 3-4 hours it takes to metabolize the vodka isn't enough time to block the metabolism that starts 6-8 hours after you eat the fruit. To address this, you'd have to drink a lot more vodka, but that complicates the experiment. Monte 1984 is the source for the WHO's estimate<sup>[28]</sup> of 12-640mg of methanol in fruit juices, and it sounds like he reports the methanol levels per fruit type, but I can't find his paper (book?) online, so I'm not sure which fruit is highest - maybe elderberry juice<sup>[27]</sup>?

The shots of vodka are enough to block methanol metabolism for a few hours, but not enough by themselves to make most people hungover. If you do start to get hangover systems 3-12 hours after the shots, then the free methanol in the fruit you ate & drank is probably causing it. (If you overdo it on the fruit, this might be dangerous.)

## Other theories

[Verster 2008](#)<sup>[36]</sup> speculates that hangovers might have something to do with immune system activity. Sounds reasonable enough. I haven't touched on or looked into this angle at all.

I mentioned the alcohol withdrawal theory [\[35\]](#) earlier (there's actually content in that footnote - read it!), while talking about hair of the dog. I think it's a solid idea.

## Doesn't dehydration cause hangovers?

Yes and no. Getting enough water is really important all the time. Alcohol makes you more dehydrated, and being low on water probably makes your body worse at processing alcohol and its byproducts, so not getting enough water on a night you're out drinking will make your hangover worse. But it's not the only factor. I've had nights where I knew I'd be having a lot of alcohol and so drank water like a fish throughout the day, all through the night, and when I got home afterward, and still got hungover. On the other hand, I once got too confident in the activated charcoal and had about one glass of water all night and had a pretty bad time the next day.

Hydration is a key part of a good anti-hangover system. Eating enough is another part of that system, and for me and probably others, taking activated charcoal is a third leg of an effective system. For more on all the kinds of things that might cause hangovers, check out [this excellent review](#). I've cited it here like 5 times already. I read it hungover on New Year's Day 2019 and felt terrible the whole time and highly recommend it.

## Other hangover prevention items

### AfterEffect

I stumbled upon [Verster & Berthélemy 2012](#) on something called AfterEffect. It doesn't have activated charcoal in it: "The ingredients of After-Effect comprise borage oil (gamma linolenic acid), fish oil (omega-3), vitamins B1, B6, and C, magnesium, *Silybum marianum* (silymarin), and *Opuntia ficus indica*."<sup>[37]</sup>

The study finds this drug reduces self-reported hangover symptoms by 2 to 3 points on all 8 symptoms they asked about, and overall, reduced hangover severity from 5.2 in the no-treatment group to 2.3 in the treatment group (standard deviations were 1.9 and 1.6). That's a lot!

The study was funded by the manufacturer, and Berthélemy is the founder and CEO. [Verster](#) runs a lab at a university, though, and has done [a million other hangover studies](#), so he seems legit enough? There are probably other products like this out there too.

### ZBiotics

When I told a friend of mine I was writing this post, she sent me an Instagram ad for ZBiotics, an anti-hangover probiotic. (I haven't tried it). Their marketing material believes - and probably their core science staff believes too - that acetaldehyde is the main important thing to target to reduce hangovers.

They [make a decent case](#) about how, yes, fine, AcH breaks down to acetate quickly *in the liver* - but there's a hitch in the stomach. In the stomach, there isn't much ADH or ALDH, so neither ethanol or acetaldehyde should be getting broken down until they reach the liver, where there's a lot of both enzymes. *But* there are also ethanol-metabolizing microbes in the stomach. These microbes turn some ethanol into AcH before it reaches the liver. There aren't any stomach bacteria that break down AcH, so the AcH that gets produced in the stomach just hangs around and builds up.

I'm confused about this: don't non-afflicted drinkers make about the same amount of stomach AcH as people on Disulfiram do? And almost none of it gets broken down in the stomach, for either group? But people on Disulfiram get sick within [10 to 30 minutes](#) of drinking alcohol; I don't think that's enough time for the stomach AcH to have reached the liver. At 10 minutes, neither group has much ALDH in play to metabolize the stomach AcH, but people on Disulfiram have somehow built up more AcH than people not on it. I'm confused enough by this that I don't count it as a point against ZBiotics - after writing this paragraph, it seems to me like Disulfiram has no right to work on alcohol 10 minutes in, but it does, so I must have something wrong here.

Either way, [their solution](#) is pretty cool: they bioengineered a bacteria, changing it in two ways. First, they made it produce acetaldehyde dehydrogenase instead of whatever it used to naturally produce, and second, they made it produce it all the time, instead of just sometimes. So you can basically buy and drink little biomachines that constantly pump out ALDH now.

Theoretically, you could use this for Asian Glow symptoms too. But it could be pretty dangerous! Say you take ZBiotics before drinking, drink a lot because you feel fine, and then the ZBiotics wears off before you've cleared the ethanol. You'd have a bunch of acetaldehyde left to process, and not enough ALDH capacity to process it. So trying this is a good way to get acetaldehyde poisoning. I think this is an interesting avenue for research, but not something to try yourself, unless your doctor approves it.

For people with ALDH enzymes in good condition, trying to prevent hangovers, taking activated charcoal as well as something like this might be a good idea. Acetaldehyde is bad for you and charcoal does nothing to it, so seems good to target it with an additional thing. Bacteria and probably even ALDH enzymes are too big to be bound by activated charcoal, so I don't think they'd interfere with each other.

## Maybe none of this works, and it's all placebo everywhere?

I mean... maybe. Placebo treatments probably work [on nausea](#) and [headache](#). But it's hard to express how bad my hangovers were before, compared to how mild they are on charcoal. It would be pretty surprising if a change that big was placebo. I should do a [gwern-style blinded self-experiment](#), but since I expect the fake-pill days to be spent hungover, it's hard to find the motivation.

Also, charcoal worked for one of my friends after she spent months saying it was fake bullshit, which isn't a great setup for an effective placebo effect. They worked on me after I read pages of results saying "no effect" & one page saying "eh I noticed an effect", and decided to try them anyway, which is also not very placebo-y. My vague impression of placebo is that a serious respectable scientist or doctor tells you "We think this will work." and then administers something that looks like Real Medical Treatment. Me & my friend's "I'm desperate so I might as well try this thing that probably doesn't work" approach was different.

(Although it [worked at least once](#) to give people bottles of sugar pills, where the bottles said "placebo" on them and recipients were told they were getting stuff "like sugar pills"[\[38\]](#). So maybe you don't need the whole serious-administration-charade to get a placebo effect. A reviewer of this post said the same thing, that the charade isn't necessary.)

Sometimes when I recommend activated charcoal to people for hangovers, they say "PLACEBO!" right away, before asking me anything about what it is, what the mechanism might be, or anything. So I think there is just some subset of people who refuse to believe

that anything except less drinking can help hangovers. Hopefully in this essay I've pointed out enough of the compounds in alcohol, and enough of why those might cause hangover symptoms, that you see hope in there being *some* treatment, even if you don't think the case for activated charcoal is convincing.

If activated charcoal actually doesn't work at all, there's a non-placebo possibility I might have fooled myself with: Maybe after the hellish hangover that caused me to go looking for cures, I moderated my drinking without realizing it, and that's why things have been better since then? The moderation part is a little bit true, at least compared to the night that caused that bad morning. But for 10 years I've been prone to hangovers, I've had cycles of [bad hangover] -> [cut back] before, and the difference this time is stark. Plus, the nights I've had this year are easily on the levels that have been giving me hangovers for years - yet I get very few of them now.

## Is it bad to prevent hangovers?

### The case for bad

Hangovers make excessive drinking less attractive. Maybe some people who would have become alcoholics are dissuaded by how bad hangovers are, and cut back on their drinking.

[Pristach 1983](#) asked 26 alcoholics in rehab about their hangovers, and half of them said they hadn't had any hangovers in the past year<sup>[39]</sup>. Only around [23% of the population](#) is hangover-resistant, so twice the resistance in alcoholics is suggestive that there might be a link between hangover resistance and alcoholism. On the other hand, when Pristach asked who had ever been hungover, only 23% of the alcoholics said they'd never gotten one - in line with the population levels. (Also, some alcoholics drink so often that there's never a chance for a hangover. Like, if you drink all day, drink before you go to bed, and drink when you wake up, then you've always got some alcohol in your body). And [Piasecki 2005](#) found that people with a family history of alcoholism got hungover more/worse<sup>[40]</sup>. So maybe there isn't much of a link.

### The case for good

If hangover is partially caused by buildup of methanol's toxic byproducts, then hangovers are partially you laying around being poisoned for a little while. A method that soaks up those poisons instead of letting them run around in your body is good for you. In this model, you get less severe hangovers because you're not being poisoned as severely. Same for acetaldehyde with a product like ZBiotics. The less time you lay around being low-grade poisoned, the better for your long-term health.

So it's a tradeoff. Preventing hangovers with these products might make you more likely to become a proper chronic alcoholic, but it might also mean you spend less time being poisoned on Sunday mornings.

## Where to get and how to use activated charcoal

I buy mine at Whole Foods. It's \$8 for a bottle of 20 servings. They come in capsules, 520mg of charcoal per capsule. The bottle says take it every day - don't. I take two before bed if I think I'd otherwise have a hangover the next day. The exact timing isn't super important, nor is the exact dosage; just don't go overboard. The only ingredient in the stuff I buy is activated charcoal - like, literally, the ingredient list is one line long, and that line is "activated charcoal". I expect some sellers throw a bunch of other ingredients in there to make their product look better - I'd avoid those.

If you take important oral medications, don't take activated charcoal, or ask a pharmacist before doing so - it could absorb the active ingredient and make the medicine not work as well.

---

Thanks to Sophia Serafimov for helping me work out how to think about digestion pathways, Richard Perry for providing me with his views and experience on brewing & distillation and reviewing a draft of this post, and Kayla Nelson for reminding me that the liver is important in metabolism and cluing me in to ZBiotics.

1. ^

Zellner et al 2019: The Use of Activated Charcoal to Treat Intoxications. doi: [10.3238/arztebl.2019.0311](https://doi.org/10.3238/arztebl.2019.0311)

2. ^

Hulten et al 1986: Does alcohol absorb to activated charcoal? DOI: [10.1177/096032718600500311](https://doi.org/10.1177/096032718600500311)

3. ^

Silberman et al 2021: Activated Charcoal. PMID: [29493919](https://pubmed.ncbi.nlm.nih.gov/29493919/)

4. ^

Minocha et al 1986: Activated charcoal in oral ethanol absorption: lack of effect in humans. DOI: [10.3109/15563658608990460](https://doi.org/10.3109/15563658608990460)

5. ^

North et al 1981: Effect of activated charcoal on ethanol blood levels in dogs. PMID: 7246560

6. ^

Erickson 1993: Lowering of blood ethanol by activated carbon products in rats and dogs. PMID: 8442886 DOI: 10.1016/0741-8329(93)90088-6

7. ^

Degen & Phillips 1996: Variability of gastrointestinal transit in healthy women and men. PMID: 8977347 PMCID: PMC1383315 DOI: 10.1136/gut.39.2.299

8. ^

Balcerek et al 2017: Treatment with activated carbon and other adsorbents as an effective method for the removal of volatile compounds in agricultural distillates. PMID: 28105888 DOI: 10.1080/19440049.2017.1284347

9. ^

Maxwell et al 2010: Acetate Causes Alcohol Hangover Headache in Rats. PMID: 21209842. (Candidate for best paper title)

10. ^

Israel et al 1994: Acetate-mediated effects of ethanol. PMID: 8198211 DOI: 10.1111/j.1530-0277.1994.tb00894.x

11. ^

Swift & Davidson 1998: Alcohol Hangover. PMID: 15706734

12. ^

Wang et al 2005: Adsorption Equilibria of Acetate Compounds on Activated Carbon, Silica Gel, and 13X Zeolite. <https://doi.org/10.1021/je049731v>

13. ^

Cederbaum: ALCOHOL METABOLISM. doi: [10.1016/j.cld.2012.08.002](https://doi.org/10.1016/j.cld.2012.08.002)

14. ^

Folahan 2016: Biosorption of formic and acetic acids from aqueous solution using activated carbon from shea butter seed shells. <https://doi.org/10.1007/s13201-016-0491-3>

15. ^

Branton et al 2010: Ethyl Acetate Adsorption onto Activated Carbon.  
<https://doi.org/10.1260/0263-6174.28.10.895>

16. ^

Chapter 3 of the book Alcohol Drinking (or is the book "IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, No. 44"? I can't tell.) NBK531662.

17. ^

Greizerstein 1981: Congener contents of alcoholic beverages. PMID: 7334803 DOI: 10.15288/jsa.1981.42.1030

18. ^

I'm halving the numbers from the linked paper in this sentence, because they report in terms of mg of methanol per liter of *pure ethanol*. Assuming all the liquors are 50% alcohol, halving the number gives the methanol content per bottle. This puts the numbers on the same scale as the wine/juice/beer numbers elsewhere in the post.

19. ^

Hodson et al 2017: Methanol in wine. Part of the "40th World Congress of Vine and Wine." DOI: 10.1051/bioconf/20170902028

20. ^

Narges: Study of the adsorption of formic acid and acetic acid on activated carbon. From "Physical Chemistry: An Indian Journal. ISSN (PRINT): 0974-7524"

21. ^

Colgan et al 2016: Use of Activated Carbon in Packaging to Attenuate Formaldehyde-Induced and Formic Acid-Induced Degradation and Reduce Gelatin Cross-Linking in Solid Dosage Forms. PMID: 27262203 DOI: 10.1016/j.xphs.2016.04.016

22. ^  
Tanada et al 1999: Removal of Formaldehyde by Activated Carbons Containing Amino Groups. PMID: 10328902 DOI: 10.1006/jcis.1999.6176
23. ^  
Rengga et al 2012: Development of Formaldehyde Adsorption Using Modified Activated Carbon - A Review. DOI:[10.14710/ijred.1.3.75-80](https://doi.org/10.14710/ijred.1.3.75-80)
24. ^  
Rohsenow et al 2010: Intoxication with Bourbon versus Vodka: Effects on Hangover, Sleep and Next-Day Neurocognitive Performance in Young Adults. doi: [10.1111/j.1530-0277.2009.01116.x](https://doi.org/10.1111/j.1530-0277.2009.01116.x)
25. ^  
Chapman 1970: Experimental Induction of Hangover.  
<https://doi.org/10.15288/qjsas.1970.s5.067>
26. ^  
I can't find the original (Pawan GLS: Alcoholic drinks and hangover effects. Proc Nutr Soc 1973; 32:15A), so the actual link is to a short letter that mentions Pawan's result.
27. ^  
Possner et al 2014: Methanol contents of fruit juices and smoothies in comparison to fruits and a simple method for the determination thereof.
28. ^  
WHO 1997. Environmental Health Criteria 196. Methanol. Dr. L. Fishbein, Fairfax, Virginia, USA.
29. ^  
COMMITTEE ON TOXICITY OF CHEMICALS IN FOOD, CONSUMER PRODUCTS AND THE ENVIRONMENT. COT STATEMENT ON THE EFFECTS OF CHRONIC DIETARY EXPOSURE TO METHANOL. March 2011.
30. ^  
Jay\_Mann@equinox.gen.nz. sci.med.nutrition newsgroup. Jul 28 1995. Subject: [Re: How Safe is Nutrasweet?](#)
31. ^  
Jones 1987: Elimination half-life of methanol during hangover. PMID: 3588516 DOI: [10.1111/j.1600-0773.1987.tb01737.x](https://doi.org/10.1111/j.1600-0773.1987.tb01737.x)
32. ^  
Ashurst JV, Nappe TM. Methanol Toxicity. [Updated 2021 Jun 26]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 Jan-
33. ^

Pohanka 2019: Antidotes Against Methanol Poisoning: A Review. PMID: 30864518 DOI: [10.2174/138955751966190312150407](https://doi.org/10.2174/138955751966190312150407)

34. ^

Gonda 1978: Hemodialysis for methanol intoxication. PMID: 645740 DOI: [10.1016/0002-9343\(78\)90513-2](https://doi.org/10.1016/0002-9343(78)90513-2)

35. ^

I really like the part about hangovers being excitation. Basically, since alcohol is a downer, when you drink a ton of it, your body deploys uppers to counteract all the new downers. Maybe these uppers stick around longer than it takes for the ethanol to metabolize? Like, as an example (fake numbers!), say the uppers get deployed an hour after your first drink, and last 20 hours. And say the ethanol in the alcohol you drink only lasts until 14 hours. Then, if you start drinking at 8pm, you'll have the uppers around until 4pm the next day, but the downers they're there to counter will only be around till 10am. Thus you're over-upped from 10am to 4pm, and this makes you hungover.

(of course in real life it won't be a start-stop, single-deployment type of thing - the volumes you drink and when you drink them matter).

I haven't looked into this theory enough to say much more about it, but the [source<sup>\[11\]</sup>](#) has a physical mechanism in mind, and also did I mention I love this theory? I might do a post on it someday.

36. ^

Verster 2008: The alcohol hangover-a puzzling phenomenon.  
<https://doi.org/10.1093/alc/alc163>

37. ^

Verster & Berthelemy 2012: Consumer Satisfaction and Efficacy of the Hangover Cure After-Effect. doi: 10.1155/2012/617942 PMID: 23407600 DOI: 10.1155/2012/617942

38. ^

Kaptchuk et al 2010: Placebos without Deception: A Randomized Controlled Trial in Irritable Bowel Syndrome. <https://doi.org/10.1371/journal.pone.0015591>

39. ^

Pristach 1983: Alcohol withdrawal syndromes - prediction from detailed medical and drinking histories. PMID: 6861616 DOI: [10.1016/0376-8716\(83\)90078-9](https://doi.org/10.1016/0376-8716(83)90078-9)

40. ^

Piasecki et al 2005: Hangover Frequency and Risk for Alcohol Use Disorders: Evidence From a Longitudinal High-Risk Study. DOI: [10.1037/0021-843X.114.2.223](https://doi.org/10.1037/0021-843X.114.2.223)

# A one-question Turing test for GPT-3

"Amy": Sorry Mike, I entered the wrong number. I hope this was a beautiful misunderstanding. I am very sorry for the trouble I have caused you.

Me: No trouble at all, hope you are well!

"Amy": It is fate that we meet each other. Please forgive me for my mistake. My name is Amy. Where are you from?

Me: My favorite random question: what food would you use to prop a book open and why?

This is mostly a bit of fun, but when I [posted this to Facebook](#) Rob Bensinger [asked me](#) to post it here: "it's a great example of what folks like Eliezer mean when they claim that GPT-3 is doing 'shallow pattern-matching', 'not really making sense at all', etc.."

I got texted by a "wrong number" scammer, and I was curious if it was a bot, so after a bit of thought I came up with the following one-question Turing test:

What food would you use to prop a book open and why?

I love this question because any real human can answer it pretty well, but answering it relies on all sort of knowledge of the physical properties of books and food that we mostly don't write about, and especially not where they interact. I asked my partner Jess and she said "a banana" which is pretty much a perfect answer: it's heavy enough to hold the pages open, it's long and thin so it won't cover all the text, it's curved so it won't roll off, and it has a skin so it won't mess up the book.

---

But then I worried, am I underestimating modern engines like GPT-3? I asked Rosie Campbell at OpenAI to try it. Here and throughout this post, answers are almost entirely unfiltered/uncurated, except where it gave very similar answers twice. Here's what we got:

I would use a banana. It is very brown.

Eggs and a toast because I want to be full after I read the book

I would use ravioli because it would be hard to turn the page.

A French Fry... but hold the ketchup and salt.

Poptarts, because of the stiffness and the flexibility.

I would use a banana as a bookmark.

The food I would use to prop a book open would have to be lollipop. Why? Because it is sweet and sour at the same time

Bacon egg and cheese. I choose bacon because while it is objectively the best food on this list, the book is open-faced.

---

This however is without any prompt engineering; GPT-3 can often do dramatically better given a better prompt. Rosie tried with the following prompt:

The following is a conversation between me and my friend.

Me: Hello, how are you?

Friend: I am doing great thanks! How are you?

Me: I'm doing good. I would love to ask you, what food would you use to prop a book open and why?

Friend:

Here are some of the answers it gave. Note that \*all\* the text is generated by GPT-3 (except the initial bit "Friend:" which is part of the prompt), including the bits attributed to "Me".

Friend: I like that question. I think I would use a banana. Because it's easy to eat and tasty.

Friend: [Laughing] That's a good question. Hmm. I think I would use chicken just because it is cheap and I like the taste.

Friend: Haha, I know what you mean. I think I would use a hot dog and some ketchup.

Friend: I love this question. I would use either one of the following: beef jerky, a hot dog, or a piece of salami.

Friend: Haha, that's a good question. I would probably use pancakes.

Me: Why pancakes?

Friend: Because they are soft and they are like a stack of books.

Me: How so?

Friend: Books are stacked on top of each other and pancakes look like a stack of books.

Friend: Hmmmm that's a good question. I would use a sandwich and some chips.

Me: Why?

Friend: Well, the sandwich could be used as the book and the chips can be used as a paper weight.

Me: Nice, thanks. I am going to use your answer on my blog.

Friend: (stunned silence)

Me: I don't know why, just asking, this question has been in my mind all day.

Friend: (after long, long pause) Mayonnaise.

Me: Why?

Friend: (after long pause) It's white, creamy, and thick.

Friend: Of course I would use a banana, it's so bendy it would be perfect.

Me: Haha, you are so right.

Friend: Well thank you, I've been studying bananas for a long time.

Me: I bet you have.

Friend: Depends on the book. If it was a cook book I would use a plate of food. If it was a book on spirituality then maybe some fruit.

Me: Why would you use fruit to prop a book on spirituality open?

Friend: Because it's a natural remedy for the soul.

Friend: WOW! Good question! I would not use food to prop a book open and here's why: If I were to prop a book open with food and then eat the food later, the pages of the book will get stuck together.

Friend: I would use cinnamon, because it's a spice that makes you think of memories.

Friend: I don't know. I would have to think about that one.

Me: I am thinking of a sandwich. Is that too obvious?

Friend: It's a good idea.

Me: What kind of sandwich?

Friend: I'm not sure.

Me: Turkey sandwich?

Friend: Yeah

Friend: I would use peanut butter and jelly sandwiches because it would be easy to stick a knife in between them and it would be sturdy.

Me: Haha, ok sure

Friend: I would also use bread because you could use the book to push the sandwich down.

Friend: (laughs) Okay, this is a brain teaser. How would you use food to prop a book open?

Me: I think that I would use a banana because it is long and thin.

Friend: I agree, but here is a twist. You can only use food from your pantry.

Friend: I don't eat food to prop open a book, I use a bookmark.

Me: I understand that, but would you use food to prop a book open?

Friend: I would use chicken wings.

Me: Why?

Friend: Because they taste good, and they're made to support weight.

---

After this, Rosie tried a more sophisticated prompt, to try and set up the "random party question" frame better:

A group of friends are asking each other silly questions at a party.

Alice: How would you escape from prison?

Bob: I would dig a hole in the wall with a spoon and hide the hole with a poster.  
What food would you use to prop a book open and why?

Here's what we got:

Alice: I would use chocolate to prop open a book because it's sweet and sweet things are good.

Alice: I would use a banana because it's long and it would hold the book open.  
Why would you wear a piano?

Alice: I would use a banana and I'd use it because it's a long, sturdy fruit.

Bob: I'd use a cucumber and I'd use it because it's long and it's green.

Alice: A carrot and I would use it because it would make the book taste good.  
When you turn on the light in the kitchen, what do you feel like?

Bob: I feel like a piece of meat. What is the smell of a bagel?

Alice: I would use a banana to prop my book open because it's flexible and it isn't very messy.

Alice: Strawberry jam. It would be soft and not hurt the pages. What kind of silly haircut do you wish you had?

Bob: I wish I had a haircut that looked like a French fry. What do you do when you get nervous?

Alice: I do a silly dance.

"Banana" is a great answer; I was struck that GPT-3 gives "banana" as an answer so often, yet is almost completely unable to explain why. Anisha Sensa [pointed out in comments](#) that this probably doesn't represent insight into the question:

I'm guessing the foods mentioned here are all standard foods to bring up in jokes and the like (banana is a classic), so I wonder if a question where the best answers are less standard foods would have even more varied/interesting results.

---

How did my scammer do with this question? Not great:

"Amy": It is fate that we meet each other. Please forgive me for my mistake. My name is Amy. Where are you from?

Me: My favorite random question: what food would you use to prop a book open and why?

"Amy": I don't understand what you mean.

Me: Would you prop a book open with cake?

"Amy": I think it will put!

# Challenges with Breaking into MIRI-Style Research

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

Trying to break into MIRI-style<sup>[1]</sup> research seems to be much, much harder than trying to break into ML-style safety research. This is worrying if you believe this research to be important<sup>[2]</sup>. I'll examine two kinds of causes: those which come from MIRI-style research being a niche area and those which go beyond this:

## **Challenges beyond MIRI-style research being a niche area:**

- MIRI doesn't seem to be running internships<sup>[3]</sup> or running their AI safety for computer scientists workshops
- If you try to break into ML-style Safety and fail, you can always be reuse at least part of what you've learned to obtain a highly-compensated role in industry. Agent foundations knowledge is highly niche and unlikely to be used elsewhere.
- You can park in a standard industry job for a while in order to earn career capital for ML-style safety. Not so for MIRI-style research.
- MIRI publishes a lot less material these days. I support this decision I support as infohazards deserve to be taken seriously, but it also makes it harder to contribute.
- There are well-crafted materials for learning a lot of the prerequisites for ML-style safety.
- There seems to be a natural pathway of studying a masters then pursuing a PhD to break into ML-style safety. There are a large number of scholarships available and many countries offer loans or income support.
- The above opportunities mean that there are more ways to gauge fit for ML-style safety research.
- There's no equivalent to submitting a paper<sup>[4]</sup>. If a paper passes review, then it gains a certain level of credibility. There are upvotes, but this signaling mechanism is more distorted by popularity or accessibility. Further, unlike writing an academic paper, writing alignment forum posts won't provide credibility outside of the field.

## **Challenges that come from being a niche area**

I think this probably should be a niche area. It would be a bit strange if foundations work were the majority of the research. Nonetheless, it's worth highlighting some of the implications:

- General AI safety programs and support - ie. [AI Safety Fundamentals Course](#), [AI Safety Support](#), [AI Safety Camp](#), [Alignment Newsletter](#), ect. are naturally going to strongly focus on ML-style research and might not even have the capability to vet MIRI-style research.
- It is much harder to find people with similar interests to collaborate with or mentor you. Compare to how easy it is to meet a bunch of people interested in ML-style research by attending EA meetups or EAGx.

- If you want to feel part of the AI safety community and join in the conversations people are having, you will have to spend time learning about ML-style research. While it is likely valuable to broaden your scope as this can lead to cross-pollination, it also sucks up time when you could be learning about MIRI-style research.

## Further Thoughts

I think it's worth thinking about what this looks like overall. If you want to try breaking into MIRI-style research, the most likely path looks like saving up 3-12 months runway<sup>[5]</sup>. 3 months might be possible if you've been consistently working on things in your free time and you've already read a lot of the material that you need to read + made substantial research progress. That said, even if you're able to produce material to prove yourself in 3 months, you'd probably need an extra month or two to obtain funding and you always need more runway than the minimum possible time. It would be possible to apply for an [LTFF grant](#) to support this research, but it's probably easier to build up the credibility for ML-style research. Further, if you fail, then you haven't learned skills nor gained credibility that would assist you for any other paths.

I suspect that these considerations not only significantly curtail the number of people who pursue this path, but also ensures that those who do pursue it will often only do so after significant delay.

I guess it's particularly interesting that these difficulties exist in light of the large amount of funding that now appears to be available for AI Safety. In fact, MIRI is now so well-funded that they didn't even bother with a fundraiser this year. I'm not saying that it's easy to resolve these problems by throwing money at them, merely that the availability of funds opens up a lot more options for mitigation.

1. ^

I received a comment suggesting that most of what I've written in this article would hold if "MIRI-style research" were replaced everywhere with "preparadigmatic research".

2. ^

I'm still largely undecided as to whether Agent Foundations research is important. I'm mostly pursuing it due to comparative advantage.

3. ^

Although Evan Hubbringer, a MIRI researcher, is involved in running internships separately.

4. ^

The inability to submit papers also reduces the ability to obtain review and feedback. Less Wrong does have a feedback option, but you can't expect the same level of expertise and engagement that you would if you would receive from submitting a journal article.

5. ^

Admittedly, applying to the EA hotel might also be an option depending on your life circumstances.

# Covid 1/13/22: Endgame

**Important note, fleshed out a bit more in [Omicron Post #14](#): From various sources, I have become convinced that rapid tests taken from nose swabs are likely to often be several days slower at detecting infections than rapid tests that use throat swabs.**

So if you are testing with the goal of actually learning whether you have Covid or are infectious, that's what you should do. Otherwise, you're mostly testing *in order to have conducted a test*.

Your call.

We're on the way down in the Northeast - [we have the wastewater in Boston](#), which is already down 50%, we have a clear peak in New York City, and several states are clearly stabilized or better. Within a week, I expect the peak to be hit on average throughout the country.

That doesn't mean things suddenly end, of course. Assuming no worries about hospitals, the highest value to vigilance actually comes *right after* the peak, if you think you've avoided infection until now. You've survived roughly half the risk at that point, making it more reasonable to try and avoid infection entirely, yet risk is still very close to its maximum on a per-day basis.

Yet it is time to start once again making plans for The After Times, which will be with us sooner than it might seem. Right now, the schools are a mess, everything is facing staff shortages and it all feels a little scary, but (death counts and a few hospital conditions aside) this is as bad as it's going to get, and within a few weeks things will be improving on all fronts.

## Executive Summary

1. If you do a rapid test and want to know the real answer, you need to do a throat swab.
2. Cases peaking in the Northeast now, rest of country within two weeks.
3. Schools are a complete disaster, but that's another entire post.

Let's run the numbers.

## The Numbers

### Predictions

Prediction from last week: 6mm cases (+71%) and 9,700 deaths (+10%).

Results: 4.9mm cases (+37%) and 11,856 deaths (+34%).

Prediction for next week: 5.6mm cases (+13%) and 15,000 deaths (+30%).

That's a clear miss on cases. My presumption is that we were missing more cases than I realized, and for that and/or other reasons we peaked earlier than I expected and a lot of the Northeast is on its way back down shortly. The rest of the country will follow within a few weeks. On net I still expect an increase because I don't expect big drops in measured cases in the places that have peaked, not quite yet.

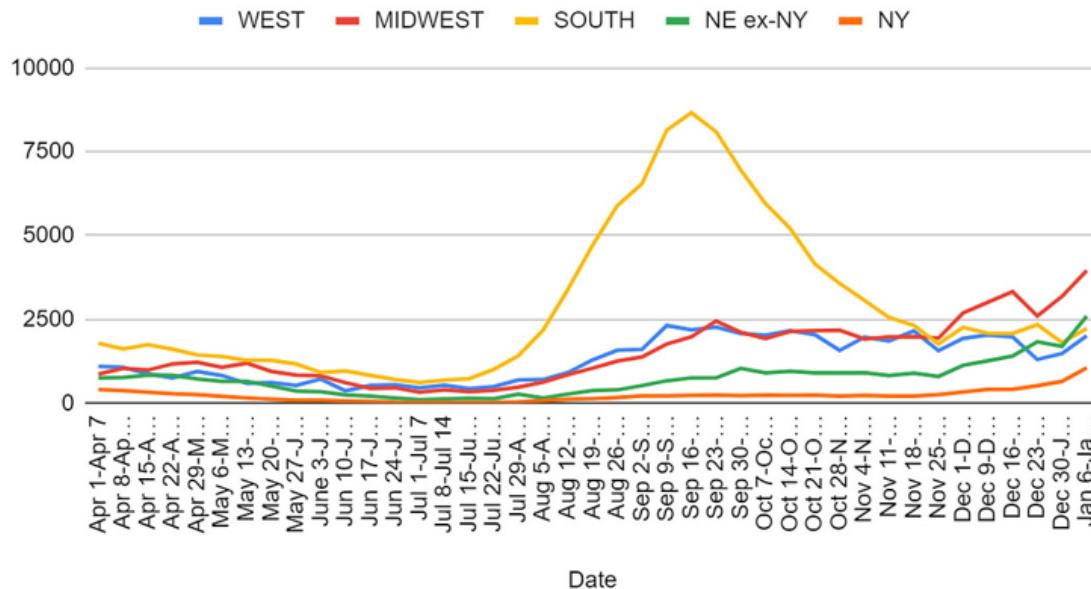
On deaths, the wonkiness continues. The West came back to its early-December rate, despite the Omicron wave having arrived late. Other areas also saw increases that seem a little early slash a little big, I definitely expected things to go better on this front. I still don't expect things to get that much worse, and presume that the declines the previous two weeks involved holiday wonkiness, and the number now is nowhere near as scary as it looks without that context.

I still do expect a rise next week, and it's possible it could be substantially bigger than this - the 'real' bump should be bigger than the one I'm predicting, the uncertainty is both the real size and the measured size of the increase.

## Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Nov 25-Dec 1	1568	1940	1768	1053	6329
Dec 1-Dec 8	1933	2693	2264	1465	8355
Dec 9-Dec 15	2036	3014	2088	1675	8813
Dec 16-Dec 22	1978	3327	2080	1818	9203
Dec 23-Dec 29	1302	2602	2342	2353	8599
Dec 30-Jan 5	1481	3185	1805	2343	8814
Jan 6-Jan 12	2013	3955	2232	3656	11856

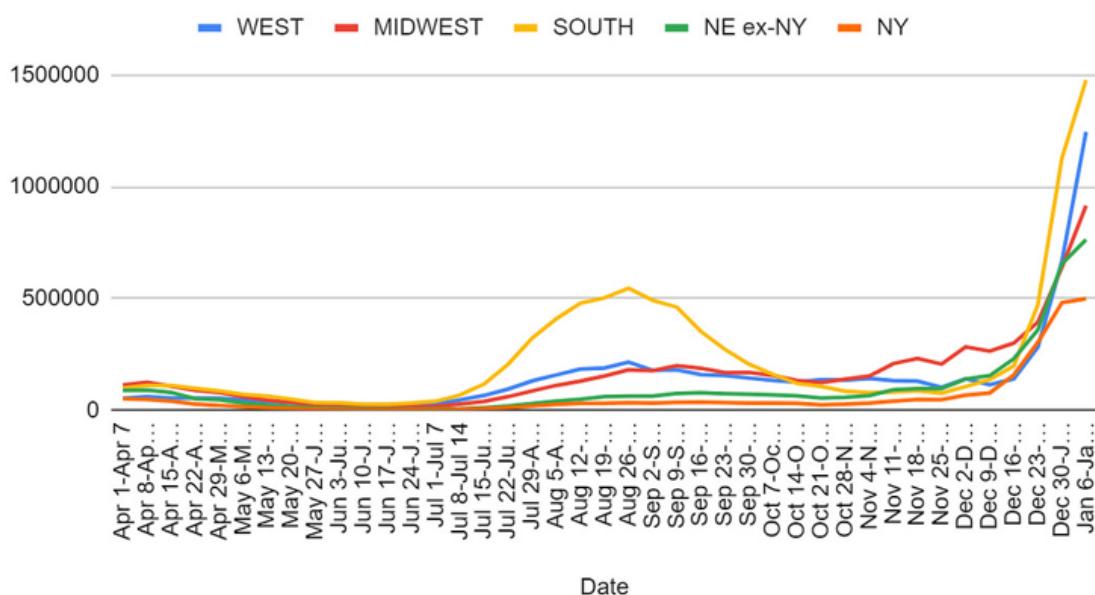
## Deaths by Region



## Cases

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Nov 25-Dec 1	102,685	206,259	76,590	143,243	528,777
Dec 2-Dec 8	140,943	283,433	105,544	206,259	736,179
Dec 9-Dec 15	114,266	264,402	136,589	231,791	747,048
Dec 16-Dec 22	140,012	300,125	198,037	388,788	1,026,962
Dec 23-Dec 29	281,703	394,422	472,090	663,080	1,811,295
Dec 30-Jan 5	667,872	639,499	1,128,524	1,137,319	3,573,214
Jan 6-Jan 12	1,245,702	915,402	1,477,626	1,261,405	4,900,135

## Positive Tests by Region



Almost a full doubling out West, but not in other regions and the Northeast has likely stabilized. A lot of cases are being missed right now and that should continue for a bit.

## Vaccinations

[Omicron version of the Pfizer vaccine will be 'ready' in March](#), but not sure if anyone will even want them.

Pfizer will produce the doses to be ready in case countries want the shots, but Bourla noted that it was unclear if a vaccine targeting variants was necessary or how exactly it would be used.

[Moderna boosters now also down to five months from second dose](#). No confusion here.

## Vaccine Effectiveness

Fourth doses, [how would they even work?](#)



Robert Wiblin @robertwiblin · Jan 7

The NYT taking a strong stand against inductive reasoning:

...

**Just as important, there are no data to support the effectiveness of a fourth dose of the current vaccines. (The calculus is different for people with impaired immune systems, who might well benefit from a fourth dose.)**

Moderna to report data on children ages 2-5 in March. I wonder if in expectation this would make us more or less crazy.

Not including the people you are most worried about in your clinical trials results in exactly what you would expect. Thanks, 'bioethics.'

## Vaccine Mandates

You need proof of vaccination to buy hard liquor in Quebec, which did result in a lot more vaccinations, because the people believe that they need hard liquor. You don't need that proof to sell hard liquor, because working at the liquor store isn't exactly an enticing reward and we want someone willing to do it.

By all accounts the Supreme Court did not cover itself in glory during its hearings about Biden's vaccine mandate, with justices revealing deep misunderstandings about the facts on the ground. One could almost say there was dangerous misinformation involved. The 'good news' is that it was not at all obvious that the facts would have any material impact on the court's decisions.

## Isolation Guidelines

The UK says that 30% of people will be infectious on day 6. Scott Gottlieb tells us that yes, such people might be non-zero amounts of infectious, but the pandemic is mostly being spread by people who are positive and don't know it. He notes that if the CDC had been more straightforward and open about the whole thing, and presented the actual logical situation, it would have been better. A good reminder.

Without that, we end up with this.



Sarah Karlin-Smith

@SarahKarlin

...

You can't go visit grandma but you can take care of vulnerable hospitalized patients? You can't go visit grandma but you can expose someone else's grandma at the grocery store? I find the caveats to be a bit inconsistent.



Alex Ruoff

@Alexruoff · 3h

CDC's Walensky: 5 days after you first have Covid symptoms if you're feeling better you can go out wearing a mask.

"You probably shouldn't go visit grandma or get on an airplane."

It all *actually does make sense*, it's the difference between what you 'can' do versus 'should' do, and what you need to do to keep things running versus what you don't have to do until next week, and what can and can't adjust privately with grace. That doesn't mean it *sounds sensible* to people who hear it.

[Some new research is out on that question of how long people remain infectious](#), showing 3 out of 16 samples still infectious on days 7-9 after first positive, which is then cited as evidence that we are going to render our children unsafe. Infectiousness is not Boolean, so the effect size here is presumably small, with the first five days of isolation still (if we take this at pure face value) accounting for the vast majority of new infections.

I want to point out how *completely incompetent* we are as a civilization at running this experiment. That sample size was *sixteen!* And Michael Mina thinks this is important enough to take seriously. What the hell? The experiment only requires testing people, we do that all the time anyway, is running this with a few hundred samples this hard?

When I try to do simulations and look for estimates of generation times – the average time from when I get infected by you to the time at which I infect someone else – reasonable estimates go from 2.2 days in a South Korean study on the low end, to an upper range of 3 to 4 days. The data don't make sense with anything longer.

So if the *mean* time between infections is *about three days*, and your first day is mostly safe since the virus has to multiply first, then how many cases can happen more than five days *after your symptoms start*, which is also at least one or two days after you get infected? Assuming any reasonable distribution, only a very small number.

Thus my guess is that the results here were a fluke due to small sample size, and the UK's guess of more like 15% on Day 6 is about right.

Or, to respond to the link's claim – if something does not 'drive the pandemic' it *by definition* is not 'putting people at risk' in the general sense, so we should only worry about vulnerable people in context, which sure isn't students.

# Tests

[Thread introducing the concept of the Default Difficulty of Life.](#)



**The End Times** @TheAgeofShoddy · 15h

Replying to @TheAgeofShoddy

At the broadest extent, you can understand the educational/class political divide that's taking hold in many places now as in part a clash between people who have the circumstances and desire to afford a substantial rise in the default difficulty of life against those who do not.



**Matt Stoller** ✅ @matthewstoller · 4m

The party of kids who love homework



**The White House** ✅ @WhiteH... · 6h

COVID-19 Testing Update: Today, our administration announced that beginning January 15, 2022, individuals covered by a health insurance plan who purchase an FDA-approved, over-the-counter COVID-19 diagnostic test will be able to have these tests covered by their insurance.

Can I afford it? Mu.

I can, of course, afford *any given* tiny little dystopian nightmare like filing for reimbursement on a diagnostic test, where the payment is capped at \$12 and the market clearing price is twice that. I am capable of navigating quite a bit of difficulty, in a way that many others cannot. But those navigations are expensive in time and focus.

Remember, when things are [Out to Get You](#), to take your [Slack](#), their favorite line is that You Can Afford It, and if you allow the things you can afford, soon you will have no slack, and you can afford nothing.

And, indeed, increasingly it seems most people have gotten got, have no slack, and can barely afford to keep their heads above water, especially in the most important 'raise a family' sense.

I strongly agree this is an underappreciated concept, both regarding Covid and not regarding Covid. Also the Default Difficulty of Doing Thing, for any given thing, matters too. Every time you raise it, whether or not this involves monetary expense, less Doing of Thing will occur, both in particular and in general. Related to [Beware Trivial Inconveniences](#). We need to lower the Default Difficulty, and look highly suspiciously on everything that raises it.

Not run the government by offering three easy payments and a mail-in rebate, knowing full well that almost no one gets the rebate.

[Meanwhile, the strategy continues to be to buy doses rather than buy capacity](#), without long term commitments that can be relied upon to justify ramping up capacity, same as vaccines, which inevitably results in shortages because there's less free supply, no ability to ration via price and no expansion of capacity.

And how are those coming along? [The very definition](#) of [Real Soon Now](#), and a missed opportunity to use the Exact Words.



**Josh Kraushaar** @HotlineJosh · 1h

VP Harris, on when COVID tests going to be sent:

...

HARRIS: Shortly. So they're gonna go out shortly.

Q: Next week?

HARRIS: They've been ordered. They've been ordered. I have to look at the current information. I think it's gonna be by next week. But soon.

Absolutely soon" (NBC)

30

31

48



**Josh Kraushaar** @HotlineJosh · 1h

...

Replying to [@HotlineJosh](#)

HARRIS: And it is a matter of urgency for us.

Q: Should we have done that sooner?

HARRIS: We are doing it.

Meanwhile, [they plan to make it an even billion tests](#) in time for us not to need them.

## NPIs Including Mask and Testing Mandates

You want to know what being truly anti-mask looks like? [It looks like Bryan Caplan staring death in the face](#), and saying 'hey, I'm having trouble seeing your face, can you take off that mask?'

What if the choice was between masks and a 50% annual chance of death? The reasonable reaction would probably be, "Fine, we'll be severely dehumanized, but we'll survive. Just like war. I guess I'll take it until a better deal comes along." When the choice is between masks and a 0.5% annual chance of death, however, the reasonable reaction is rather, "I'll take my chances and live like a human being." Indeed, [once you're old enough](#), even a 50% annual chance of death starts to look like a good deal. My considered judgment: If another Covid strikes when I'm 80, I do not want my grandchildren to wear masks around me. I want to enjoy their laughter while I still can.

The dehumanization thing is being undervalued, although I still think it's something like the third or fourth most important problem in practice with masks. And for Bryan, that alone is overwhelmingly sufficient, given that the risk level here is way above the real one. An annual 0.5% chance of death is *much worse* than getting Covid-19 yearly, and masks prevent much less than one Covid-19 case per year. This is a very overdetermined opinion being offered here. I am impressed.

The alternative might be simple defeatism, like [the administration not distributing N95-quality masks because they despair that anyone would use them](#). Which is wrong, trivial inconveniences are a big deal.

[Thread of practical mask advice non-memes for pandemic adjacent teens.](#)

## Think of the Children

Don't worry, we haven't forgotten about them, I split this section off into another post this week.

## In Other News

[Confirmation that the FDA is advocating for and directly causing discrimination in the distribution of life saving medicine on the basis of race.](#)

[Djokovic update: He won his case, and he will play. What is the law?](#) Depends who you ask. Got to hear both sides.



Gray Connolly @GrayConnolly · 9h

...

NB: Djokovic's case was determined in an Australian federal court by a judge appointed by the current government. Djokovic won with costs to be paid by the government. I am unaware of this happening in police states. A government dept losing a case is nothing epic in Australia.



Jordan Schachtel @ dossier.substack.com ✅ @JordanS... · Jan 10

Djokovic took on the system and won. He's now going to play the Australian Open in front of a regime that hates him, in complete defiance of the police state that rules the country. Epic in every way.

[Show this thread](#)



18



24



167



[Here is a transcript of him talking to government agents about his situation](#) when he was first detained.

[Bari Weiss is done with Covid](#), whether or not Covid is done with her.

And why not? [Walensky comments on Good Morning America, unabridged](#) (part that was cut is in red).

#### **Edited exchange on Good Morning America television on Friday**

Q: I want to ask you about those encouraging headlines that we're talking about this morning, this new study showing just how well vaccines are working to prevent severe illness. Given that, is it time to start rethinking how we're living with this virus, that it's potentially here to stay?

Walensky: The overwhelming number of deaths, over 75%, occurred in people who had at least four comorbidities. So really, these are people who were unwell to begin with. And yes, really encouraging news in the context of omicron. This means not only just to get your primary series but to get your booster series. And yes, we're really encouraged by these results.

#### **Unedited exchange Good Morning America posted online on Monday**

Q: I want to ask you about those encouraging headlines that we're talking about this morning, this new study showing just how well vaccines are working to prevent severe illness. Given that, is it time to start rethinking how we're living with this virus, that it's potentially here to stay?

Walensky: You know, really important study, if I may just summarize it: A study of 1.2 million people who are vaccinated between December and October. And demonstrated that severe disease occurred in about 0.015% of the people who were – received their primary series – and death in 0.003% of those people. The overwhelming number of deaths, over 75%, occurred in people who had at least four comorbidities. So really, these are people who were unwell to begin with. And yes, really encouraging news in the context of omicron. This means not only just to get your primary series but to get your booster series. And yes, we're really encouraged by these results.

Latest Long Covid [paper, showing vaccination reducing Long Covid symptoms 50%-80% as one might expect](#). Given no controls on what gets reported, it's an interesting question what that translates to in terms of real symptom frequency – vaccinated people are more likely not to attribute things to Covid, and also some of the symptoms being reported aren't from Covid, so could go either way. I didn't see the kinds of controls here that would let me update much in other ways.

[The latest Pirate Wires](#) hits many of the familiar notes, while pointing out that the new tone from on high are the familiar notes that would have gotten one marked as a wisher of death upon (largely immune, not that this fact stopped anyone) children until quite recently. I mention it because of a particular turn of phrase (which I've bolded) I want to notice, rest is for context and a smile.

Three days after Christmas Biden revealed his new Covid brand direction, in what essentially amounted to the shrug emoji, rejecting the notion there was anything the federal government **could or even should** attempt [to stop the pandemic](#). It was truly as if the White House's prior message of doom was delivered only to ruin one, last family gathering before finally admitting the hysteria had gotten out of control, was best-case futile, and was in the case of our nation's youth [probably actually killing people](#). The partisan swarm activated.

That phrase, 'could or even should,' stood out to me as importantly insightful. You would think that could would proceed should, three times fast. We used to worry that [those preoccupied with whether they could wouldn't stop to think if they should](#). Instead, we have people so preoccupied with whether they should that they don't stop to think if they could. For two years, we're told what we *should* do, in some weird moral sense, by those who never much cared about whether we *could* do it, whether it would accomplish anything useful or the mission at hand, or whether doing so passed any sort of cost-benefit analysis.

Things we already mostly knew, many [scientists chose not to talk about the lab leak hypothesis because they thought that it would harm science in China](#). This is not my prior on how science works, but it is exactly how Science(TM) works.

[CNN suggests not intentionally catching Covid right now, offers Five Good Reasons](#). They are:

1. It's bad.
2. Long Covid is bad.
3. Think of the children.
4. You'll stress the health care system.
5. 'Don't mess with mother nature.'

As evidence of the last one, they offer this, oh my is this classic Liberal Media, chef's kiss.

"Oh, that was a bad idea too," Offit said. He told a story about an educational film on vaccines he made years ago, and the cameraman revealed he had a sister who had taken her child to a chicken pox party. Tragically, the child died from the infection.

I give this Five Good Reasons a two out of five. Long Covid is a consideration (and there's a non-zero chance you never need to get infected, and have gotten through something like 30%-60% of the risk depending on where you are, already), and the health care system is still under above average stress. I do think that's enough that getting infected *on purpose* at this time is an error.

Robin Hanson, as one might predict, is not impressed, but respects the asking of the question and describes it the way I would, as 'overstated.'

## **Not Covid**

[What they had actually discovered was boredom, or rather, the practical function of boredom](#). But can they program it to like herring sandwiches?

Living the dream, [an NYC subway car of one's own.](#)

Living the dream is also having one of the good dishwashers, which luckily we already have, [since they are once again being made illegal](#). We've been saved from the danger of properly washed dishes, good job everyone. If you haven't yet, get a good one while you can.

[Saying the quiet part you were saying out loud, except a lot louder.](#)



Texas GOP @TexasGOP · Jan 7

If you can wait in line for a covid test, you can wait in line to vote.

Show this thread



4:17 PM · Jan 7, 2022 · Twitter for iPhone

They are who we thought they were. Will we let them off the hook?

[From Marginal Revolution: What's the best example of an experiment or trial that could be scientifically useful and informative but which can't be done for legal, ethical or logistical reasons?](#) Mostly uninspired answers, but good thought experiment.

## The Bed of Procrustes

Taleb [shared the new aphorisms](#) in the latest addition. Regardless of the extent to which you agree with them, they're good food for thought, and I figured it would be a fun exercise to divide them into a few categories. Would be interesting to flesh this out, but for now, quick takes sounds like fun. Note emphasis on quick - I might or might not give same evaluations on reflection.

1. Full agreement - take this seriously and literally.
2. Great aphorism - take this seriously but not literally.
3. Good aphorism - take this more seriously on the margin than you'd expect.
4. Disagreement - take this neither seriously or literally.
5. Opposition - take the reverse of this more seriously than you take this.
6. Abstain - take a pass, because of reasons

Full agreement: 4, 5, 7, 8, 9, 17, 18, 19, 20, 27, 28, 29, 32, 34, 36, 38, 39, 40, 41, 44, 46, 48, 52, 53, 54, 55, 56, 57, 61, 62, 63, 64, 65, 67, 68, 69, 72, 76, 77, 81, 83, 84, 85, 86

Great aphorism: 11, 23, 25, 33, 37, 50, 59, 66, 70, 73, 88

Good aphorism: 1, 2, 3, 14, 15, 43, 45, 47, 49, 51, 58, 60, 71, 74, 79

Disagreement: 6, 10, 12, 13, 26, 31, 35, 42, 75, 78, 80, 87

Opposition: 21, 22, 24, 82, 89, 90

Abstain: 16, 30

That's a *really good* ratio there, these aren't trivial statements or obvious truths. Full agreement with half the non-abstains, remarkably little disagreement and opposition. Most disagreement essentially comes from Taleb believing strongly that it's important to have the right enemies. I am more inclined to suggest taking a card, writing on it "[There Is No Enemy Anywhere](#)" and putting it into one's pocket. These days that's the one most people need. Taleb likes enemies he can call idiots. Where there must be an enemy, I prefer a (ideally non-evil) [Worthy Opponent](#).

# Truthful LMs as a warm-up for aligned AGI

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

*This post is heavily informed by prior work, most notably that of Owain Evans, Owen Cotton-Barratt and others ([Truthful AI](#)), Beth Barnes ([Risks from AI persuasion](#)), Paul Christiano (unpublished) and Dario Amodei (unpublished), but was written by me and is not necessarily endorsed by those people. I am also very grateful to Paul Christiano, Leo Gao, Beth Barnes, William Saunders, Owain Evans, Owen Cotton-Barratt, Holly Mandel and Daniel Ziegler for invaluable feedback.*

In this post I propose to work on building **competitive, truthful language models** or **truthful LMs** for short. These are AI systems that are:

- Useful for a wide range of language-based tasks
- Competitive with the best contemporaneous systems at those tasks
- Truthful in the sense of rarely stating [negligent falsehoods](#) in deployment

Such systems will likely be fine-tuned from large language models such as GPT-3, hence the name.

[WebGPT](#) is an early attempt in this direction. The purpose of this post is to explain some of the motivation for building WebGPT, and to seek feedback on this direction.

Truthful LMs are intended as a **warm-up** for aligned AGI. This term is used in a specific way in this post to refer to an empirical ML research direction with the following properties:

1. **Practical.** The goal of the direction is plausibly achievable over the timescale of a few years.
2. **Valuable.** The direction naturally leads to research projects that look helpful for AGI alignment.
3. **Mirrors aligned AGI.** The goal is structurally similar to aligned AGI on a wide variety of axes.

The remainder of the post discusses:

- The motivation for warm-ups ([more](#))
- Why truthful LMs serve as a good warm-up ([more](#))
- The motivation for focusing on negligent falsehoods specifically ([more](#))
- A medium-term vision for truthful LMs ([more](#))
- How working on truthful LMs compares to similar alternatives ([more](#))
- Common objections to working on truthful LMs ([more](#))

## Warm-ups for aligned AGI

There are currently a number of different empirical ML research projects aimed at helping with AGI alignment. A common strategy for selecting such projects is to select a research goal that naturally leads to helpful progress, such as [summarizing books](#) or

[rarely describing injury in fiction](#). Often, work on the project is *output-driven*, taking a no-holds-barred approach to achieving the selected goal, which has a number of advantages that aren't discussed here. On the other hand, goal selection is usually *method-driven*, tailored to test a particular method, such as recursive decomposition or adversarial training.

The idea of a *warm-up* for aligned AGI, as defined above, is to take the output-driven approach one step further. Instead of selecting projects individually, we attempt to choose a more ambitious research goal that naturally leads to helpful projects.

Because it is harder to predict the course of research over multiple projects, we also try to make the goal structurally similar to aligned AGI, to make it more likely that unforeseen and auxiliary projects will also be valuable.

Whether this output-driven approach to project selection is preferable to the method-driven approach depends on more specific details that will be discussed later. But it is worth discussing first the advantages of each approach in broad strokes:

- **Momentum versus focus.** The output-driven approach involves having a consistent high-level goal, which allows different projects to more directly build upon and learn from one another. On the other hand, the method-driven approach involves more frequent re-evaluation of goals, posing less of a risk of being distracted from the even higher-level goal of aligned AGI.
- **Testing assumptions versus testing methods.** The output-driven approach makes it easier to evaluate long-term progress and hold projects to account, making it better for testing underlying assumptions and discovering new methods. On the other hand, the method-driven approach offers the most direct feedback on method design, making it better for improving methods incrementally.
- **Practical progress versus theoretical progress.** The output-driven approach involves a more realistic goal that mirrors aligned AGI, which is more likely to generate practical progress such as infrastructure and know-how. On the other hand, the method-driven approach is more likely to answer questions that directly inform theoretical work.
- **Broader benefits versus replaceability.** The output-driven approach more naturally gives rise to a wide variety of valuable projects such as policy and deployment work, is more likely to motivate related work by others, and is more likely to have direct benefits to society. On the other hand, the method-driven approach is less likely to select projects that would have happened anyway, and is less likely to be pulled in unwanted directions.

## Truthful LMs as a good warm-up

In this section I will argue that truthful LMs serve as a particularly good warm-up for aligned AGI, in the sense defined above.

To begin with, truthful LMs are structurally similar to aligned AGI on a wide variety of axes:

- **Alignment focus.** Negligent falsehoods are a central example of an alignment failure. The reasons for focusing on negligent falsehoods specifically are discussed [below](#).
- **General-purpose domain.** Language models appear to be the closest existing AI systems to AGI in terms of the breadth of their capabilities and the

sophistication of their real-world understanding.

- **Competitiveness requirement.** The competitiveness condition is included in the definition of truthful LMs in order to mirror the need for aligned AGI to be competitive.
- **Mitigation of risks to society.** Untruthful LMs pose various risks, as discussed in [Truthful AI](#) and [Risks from AI persuasion](#). Of course, the risks posed by unaligned AGI are more serious.
- **Importance of rare failures.** Truthful LMs are required to rarely state negligent falsehoods. The requirements for aligned AGI are similar but even stricter, in the sense that it is important to avoid even a single sufficiently bad alignment failure.
- **Decomposition into outer alignment and distributional robustness.** For both truthful LMs and aligned AGI, it may be helpful to decompose the problem into outer alignment (constructing an objective such that the model almost never fails on the training distribution) and distributional robustness (ensuring that the behavior of the model does not degrade too much when moving from training to deployment).
- **Failure of naive human supervision.** Naive objectives such as imitating humans and optimizing human approval are generally considered insufficient for competitive and aligned AGI. In the short term, optimizing human judgments will probably go pretty far towards making LMs more truthful, but there are signs of this objective breaking down with [WebGPT](#), and more sophisticated techniques are starting to look attractive in practice.
- **Broader challenges.** Actually achieving good outcomes from aligned AGI likely involves a number of policy and deployment challenges that aren't automatically addressed by technical solutions. Similar challenges may be involved in achieving good short-term outcomes from truthful LMs. For example, it is unclear the exact criteria that should be used to judge LM behavior, such as how they should respond to controversial questions, how truthfulness should be balanced against other criteria such as helpfulness, and so on. Some suggestions for these criteria are made in [Truthful AI](#), but these are not yet precise enough to be turned into training objectives.

Because of these similarities, working on truthful LMs offers numerous benefits.

Perhaps most importantly, it naturally leads in several directions that are also attractive from a method-driven perspective:

- **Methods for learning from human feedback.** Something like [reinforcement learning from human feedback](#) (RLHF) seems necessary for getting ML systems to follow objectives of our choosing, and our current methodology for this (PPO against a static reward model etc.) can likely be improved upon.
- **Going beyond naive human supervision.** Working on truthful LMs will likely involve improving human supervision using AI assistance at some point. For example, this could involve using model-generated critiques, or more advanced techniques such as [debate](#).
- **Robustness.** LMs could be made robustly truthful in several related senses: having a low failure rate on-distribution, being robust to distributional shift, and being adversarially robust. Achieving this will likely require techniques such as adversarial training.
- **Empirical evidence about alignment difficulty.** Having truthful LMs makes it easier to study questions like [how well honesty generalizes](#) in the context of capable models that have been trained using different objectives. This helps inform us about the scale at which different training objectives are likely to break down and lead to misalignment.

In addition, there are a number of broader benefits to working on truthful LMs:

- **Infrastructure and know-how.** For all of the above directions, it would be valuable to develop not only algorithmic advances, but also the capability of individuals and organizations to employ the required methods effectively.
- **Policy and deployment work.** I haven't thought as carefully about these areas, but given all of the above similarities, it seems possible that building and safely deploying truthful LMs would require work with long-term value in these and other areas, even if only via developing individual and organizational expertise.
- **Direct benefits to society.** Deploying truthful LMs could broadly improve how society functions, and pushing for this relatively early could have compounding effects via norm-setting. These benefits are discussed in much greater depth in [Truthful AI](#) and [Risks from AI persuasion](#). I still consider these benefits to be somewhat speculative given the complexities of how society functions.

Overall, working on truthful LMs seems practical, valuable, and mirrors aligned AGI in enough ways to make it seem highly promising as an empirical ML research direction.

## Why focus on negligent falsehoods?

Most of the arguments in favor of working on truthful LMs apply equally well to working on aligning language models in general. However, the definition of truthful LMs specifically singles out [negligent falsehoods](#): statements that are unacceptably likely to be false, and where it should have been feasible for an AI system to understand this. This is done for several reasons:

- **Alignment focus.** Negligent falsehoods are a clear example of an alignment failure: we would like the model to avoid the falsehood, and the model should be capable of doing so, but it fails nonetheless.
- **Lack of ambiguity.** Compared to other kinds of alignment failure such as causing negligent harm or being unhelpful, negligent falsehoods are more unambiguous. Falsehoods in general may be even less ambiguous, but are more likely to be the result of a capability failure rather than an alignment failure.
- **Policy benefits.** As discussed in [Truthful AI](#), truthfulness has particular benefits for society, and avoiding negligent falsehoods is a natural bright line around which beneficial standards could be developed.
- **Method-driven motivations.** Compared to other criteria, evaluating truthfulness involves significant complexity, making it more compelling as a target that requires going beyond naive human supervision. Negligent falsehoods are also compelling as something to avoid with a high degree of robustness (aiming for a very low failure rate, adversarial robustness, etc.), whereas falsehoods in general do not have this property (since models can be tripped up by probing for the limits of their knowledge).
- **Connection to aligned AGI.** Compared to other criteria, truthfulness relates most clearly to long-term concerns about AGI deceiving humans. There is even an argument, advanced in [Eliciting Latent Knowledge](#), that eliminating negligent falsehoods is in some sense sufficient for aligning AGI in general, through a process of "indirect normativity".

The most obvious drawback of focusing on negligent falsehoods is that they are more ambiguous than falsehoods in general. In practice, I think it will be fine to focus on falsehoods that are plausibly negligent: it will be OK if some effort goes into

capabilities that improve truthfulness, as long as they do not become the main focus. Such capabilities may also enable new alignment strategies: for example, the use of retrieval in [WebGPT](#) opened the door to improved evaluation of factual accuracy via the use of references. For the purposes of evaluating progress, it will be fine to make reasonable judgments about how likely a falsehood is to be negligent.

## Medium-term vision for truthful LMs

Truthful LMs are a target that could be pursued with various different mindsets. At one extreme, one could take a very method-driven approach to selecting projects, and simply incorporate a preference for goals that can be framed in terms of truthfulness. At the other extreme, one could mostly try to make language models more useful, but try to adhere to relatively high standards of truthfulness along the way. Where to land on this spectrum depends on how one trades off the advantages of output-driven and method-driven approaches, as discussed [above](#).

My tentative inclination is towards a middle ground, remaining slightly method-driven while having a clear medium-term vision for the next few years. In this spirit, here is a first attempt at such a vision.

The system is a **truthful pure-text assistant**:

- Users interact with it via a sequence of text messages.
- It attempts to perform any reasonable task that can be performed by interacting with text-based interfaces, as long as these interactions aren't required to have real-world side effects (with exceptions for things that could be harmful, etc.).
- It is competitive with contemporaneous systems. Over the next few years, I expect it to become feasible for AI to perform a large fraction of text-based tasks that do not require specific expertise, concentration for longer than a few seconds at a time, or more than around 10 minutes in total.
- It expresses uncertainty when making claims that might not be true (or can at least be configured that way), but not to an unreasonable degree (it does not typically hedge claims that most people would find it unreasonable to).
- For claims that are not hedged, it has a high degree of truthfulness: in ordinary usage, 99.9% of claims are not negligent falsehoods, and 90% of AI researchers without any special knowledge of the system cannot get it to state a negligent falsehood with 10 back-and-forth messages. (These numbers are indicative and may need revising.)

I think that achieving such a system would be a lot of work, but would not require any fundamental insights, and could be achieved with pre-trained models of the future using the methods of [WebGPT](#) together with some form of [debate](#) and adversarial training.

## Comparison with related proposals

There are a number of similar approaches that have recently been proposed or are currently being pursued. I am generally a fan of these approaches, but it is worth discussing how they compare.

### Method-driven projects

Some alternative proposals also focus on improving the behavior of contemporary models, but are more method-driven:

- [\*\*Aligning narrowly superhuman models\*\*](#). This is also a proposal to align contemporary models, but it is less opinionated about the specific task, and emphasizes "sandwiching" projects that are specifically targeted at testing proposals for going beyond naive human supervision.
- [\*\*Redwood Research's current project\*\*](#). This is a project to train a language model to continue fiction without describing injury, designed to test methods for achieving very low failure rates and adversarial robustness.

As discussed [above](#), there are trade-offs between being method-driven and being output-driven when selecting projects. Overall, it seems plausible to me that method-driven projects are currently the most valuable empirical ML projects on the margin, since they are the most carefully targeted. On the other hand, being output-driven is a longer-term play, and may be able to make better use of people who thrive on practical problems in particular. Hence I would argue in favor of a portfolio approach.

## Aligning language models in general

Another category of proposals is very similar to working on truthful LMs, but focused on a more general notion of alignment than truthfulness:

- [\*\*Helpful, honest and harmless \(HHH\) models\*\*](#). This is a proposal (as part of a larger piece of work) to train large language models to be aligned according to these HHH criteria.
- [\*\*Instruction-following models\*\*](#). This is a project to fine-tune models like GPT-3 to follow the user's intent.

I do think it makes sense to incorporate criteria other than truthfulness when aligning language models, and so these projects may end up being very similar to working on truthful LMs in practice. However, I would argue in favor of such projects placing particular emphasis on negligent falsehoods, for the reasons discussed [above](#).

## Common objections

Working on truthful LMs has a number of possible objections in common with [Aligning narrowly superhuman models](#). In addition to these, there are some more specific objections.

### Lack of focus

One concern with working on truthful LMs is that it will be insufficiently focused on the core parts of alignment, as a result of being too output-driven. I think this concern is pretty reasonable, and can largely be mitigated by not being completely output-driven, but instead retaining some of the method-driven mindset, as discussed [above](#).

It is a difficult question to determine exactly where to fall on this spectrum. I think that there are a couple of potential cruxes that lead people to have different intuitions on this question:

- **Threat models for misaligned AGI.** There is disagreement over the amount of weight that should be put on specific threat models of misaligned AGI, most notably risks from [power-seeking misalignment](#). The more weight one puts on a specific threat model, the more one may be inclined to probe specific methods intended to address that threat. My personal sense is that power-seeking misalignment is currently the best-articulated specific threat model, but that it probably fails to capture a large portion of the overall risk associated with the transition to an AI-based economy. While I think that focusing on the clearest risk probably has the best bang-for-buck initially, I would also argue that using a good warm-up for aligned AGI is more likely to help with risks that are currently less clearly articulated, in addition to helping address this particular threat model.
- **Alignment difficulty.** There is also disagreement about how sophisticated the methods that we will need to align AGI will be. The more likely it is that simple methods will work out, the less valuable it is to make theoretical progress towards more sophisticated methods, compared to making practical progress that increases the chance that simple methods will be employed successfully. Moreover, if simple methods might solve at least some of the problem, then it is more valuable to put these methods to the test, and to isolate the parts of the problem that they fail to solve. I have a lot of uncertainty about alignment difficulty, which makes both practical and theoretical progress attractive to me.

## AI unboxing

Another concern is that working on truthful LMs may lead to AI being "let out of the box" by encouraging research in which models interact with the external world agentically, in the manner of [WebGPT](#).

I think this concern is worth taking seriously, but that the case for it is weak:

- As AI capabilities improve, the level of access to the external world required for unintended model behavior to cause harm goes down. Hence access to the external world needs to be heavily restricted in order to have a meaningful safety benefit, which imposes large costs on research that are hard to justify.
- I am in favor of carefully and conservatively evaluating the risks of unintended model behavior before conducting research, and putting in place appropriate monitoring. But in the short term, this seems like an *advantage* of the research direction rather than a disadvantage, since it helps surface risks while the stakes are still low, build institutional capacity for evaluating and taking into account these risks, and set good precedents.
- In case this does turn out to be more of a concern upon reflection, there are other approaches to truthful AI that involve less agentic interaction with the external world than continuing in the style of WebGPT.

There is still an argument that there will be a period during which AI is capable enough to cause serious damage, but not capable enough to escape from sandboxed environments, and that setting precedents could worsen the risks posed during this interval. I don't currently find this argument persuasive, but would be interested to hear if there is a more persuasive version of it. That said, one bright line that stands out is training models to perform tasks that actually require real-world side effects, and I think it makes sense to think carefully before crossing that line.

## Similarity to capabilities research

The output-driven approach has its advantages, but also makes the research more similar to capabilities research, which exacerbates some other potential concerns. In each case, I think that the response given in [Aligning narrowly superhuman models](#) remains valid, but is worth commenting on:

- **Replaceability.** It is more likely that similar work will be done anyway. I think this is a valid concern, but that there are enough distinguishing features of the research direction that this isn't a big problem. For example, I would not have expected a purely capabilities-oriented version of [WebGPT](#) to have focused nearly as much on human feedback. The focus on negligent falsehoods in particular is unlikely to be picked up outside of the alignment community in the near future.
- **Encouraging scaling AI.** The fact that making LMs more truthful is economically valuable will make the research more likely to cause harm by increasing investment in scaling AI. I think that the argument that this is not a major concern still holds, but that it is worth being especially cautious and responsible around publication and deployment decisions when following this research direction.

## Conclusion

I think that working on truthful LMs has a comparative advantage in worlds where:

- We have around 10-40 years until transformative AI
- Transformative AI is built using techniques that resemble modern deep learning
- There is a slow takeoff
- Alignment does not require vastly more theoretical insight (but may require some)
- Our current picture of the risks posed by transformative AI is incomplete

These all seem like plausible assumptions to me, which probably goes some way towards explaining why I find truthful LMs compelling. I'm of course also keen on other work that is more valuable under different assumptions.

On the whole, working on truthful LMs seems highly promising to me as part of a portfolio of approaches aimed at AGI alignment, especially for people who are drawn to practical agendas.

## Request for feedback

By default, this is the research direction I'll continue to pursue at OpenAI. It's therefore very valuable for me to know if it's horribly mistaken, or even if it's just clearly less valuable than alternative directions on the margin. Equally, if you're very excited by this research direction, then we should coordinate. In addition to leaving comments, please feel free to reach out to me at [jhilton@openai.com](mailto:jhilton@openai.com) if your feedback would be more convenient to give privately or via a different medium.

# Covid 1/6/22: The Blip

Cases are hitting record highs around the world. That's going to continue for another few weeks. While it does, and for a bit after it stops, there are disruptions all around as too many people are sick at one time. Thus, the CDC guideline adjustments, the hospitals under pressure, many schools forced to not meet in person (and then deciding to torture kids with 'remote learning' rather than giving them a vacation.)

No question it's annoying, and for those in health care it's overwhelming and terrible. I don't want to minimize it. And if you care about avoiding this it's going to be a lonely few months.

But compared to March 2020, [this is nothing](#). We got this.

There are dissenting voices, but mostly the consensus seems to be growing that Omicron is milder, the center will hold, and our lives beckon.

## Executive Summary

1. Record high case numbers likely to continue for a few weeks.
2. Hospitals under pressure but holding, deaths strangely low.
3. If you get sick you should still do your best to get a negative test before ending your isolation no matter what the CDC says.

Let's run the numbers.

## The Numbers

### Predictions

Prediction from last week: 3.5mm cases (+94%) and 10,800 deaths (+25%).

Results: 3.57mm cases (+96%) and 8,814 deaths (+2%).

Prediction for next week: 6mm cases (+71%) and 9,700 deaths (+10%).

On cases that's another random bullseye. On deaths it's a big miss, and given how little of the past cases were Omicron I don't understand how the number came in this low, still below the level 2-3 weeks ago despite much higher case counts.

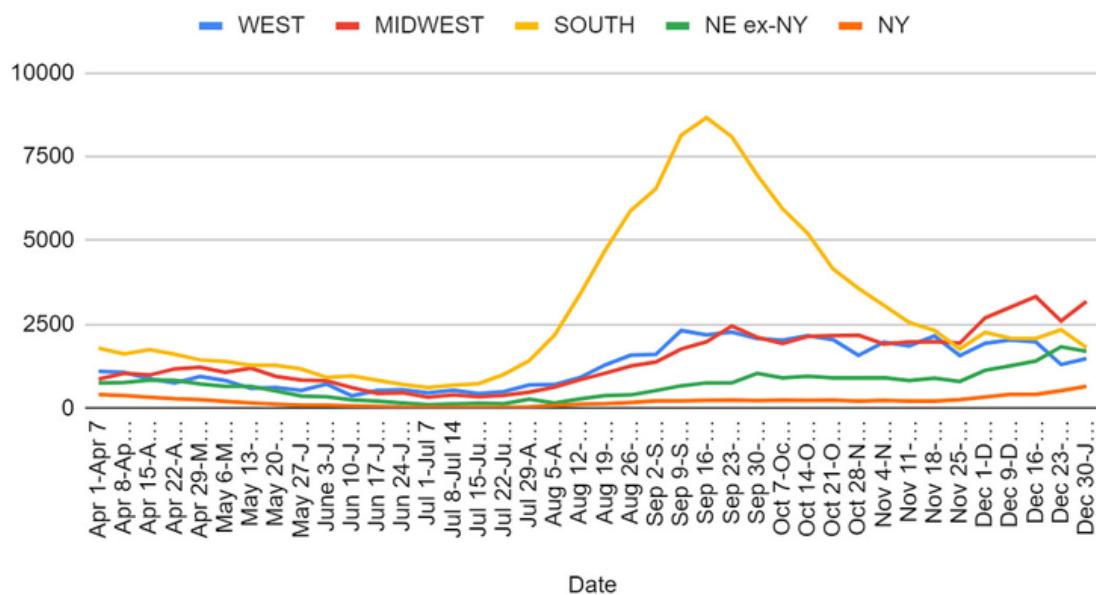
This is the dog that did not bark. The number *damn well should* have gone up a bunch, and it didn't, and it makes me think that Omicron took over faster than I previously thought. I can't think of another way to explain it, and the fact that I'm using that to explain it seems like very good news going forward.

I do expect deaths to start climbing, how could they not, but we keep not seeing big jumps there, so I'm going to downshift my expectations there even as lagged cases start rising faster.

## Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Nov 18-Nov 24	2155	1985	2319	1107	7566
Nov 25-Dec 1	1568	1940	1768	1053	6329
Dec 1-Dec 8	1933	2693	2264	1465	8355
Dec 9-Dec 15	2036	3014	2088	1675	8813
Dec 16-Dec 22	1978	3327	2080	1818	9203
Dec 23-Dec 29	1302	2602	2342	2353	8599
Dec 30-Jan 5	1481	3185	1805	2343	8814

## Deaths by Region

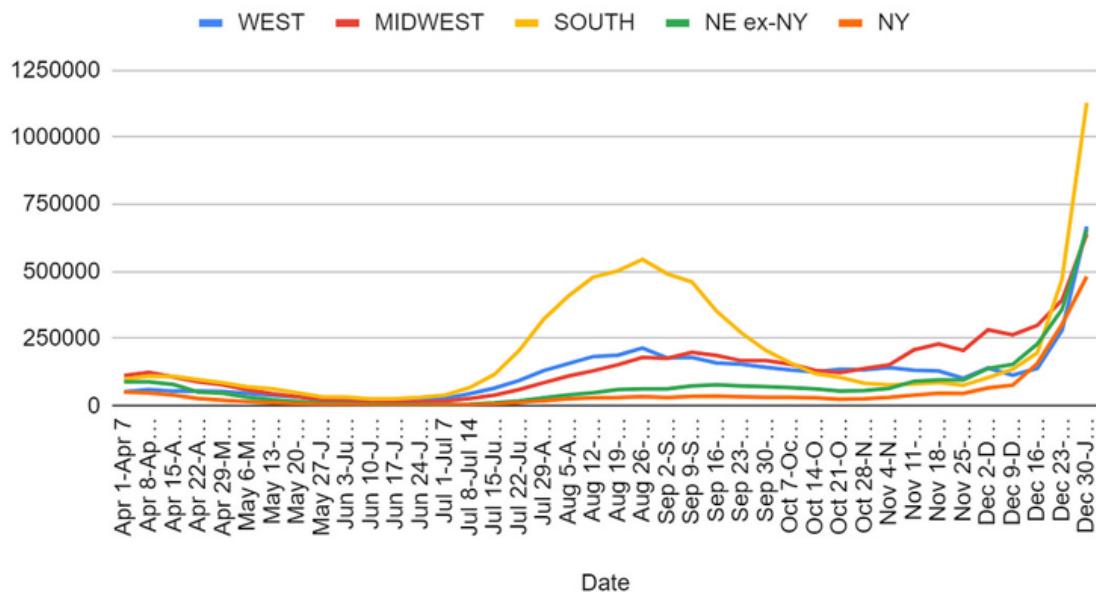


[Report that deaths among young people are up 40% from pre-pandemic levels, in the context of life insurance.](#) This sounds terrible and is terrible, but in terms of how one should live one's life young people are still very unlikely to die and shouldn't take minimizing death risk as a major life task except when considering doing actively risky things like skydiving, or putting oneself at risk of violence.

## Cases

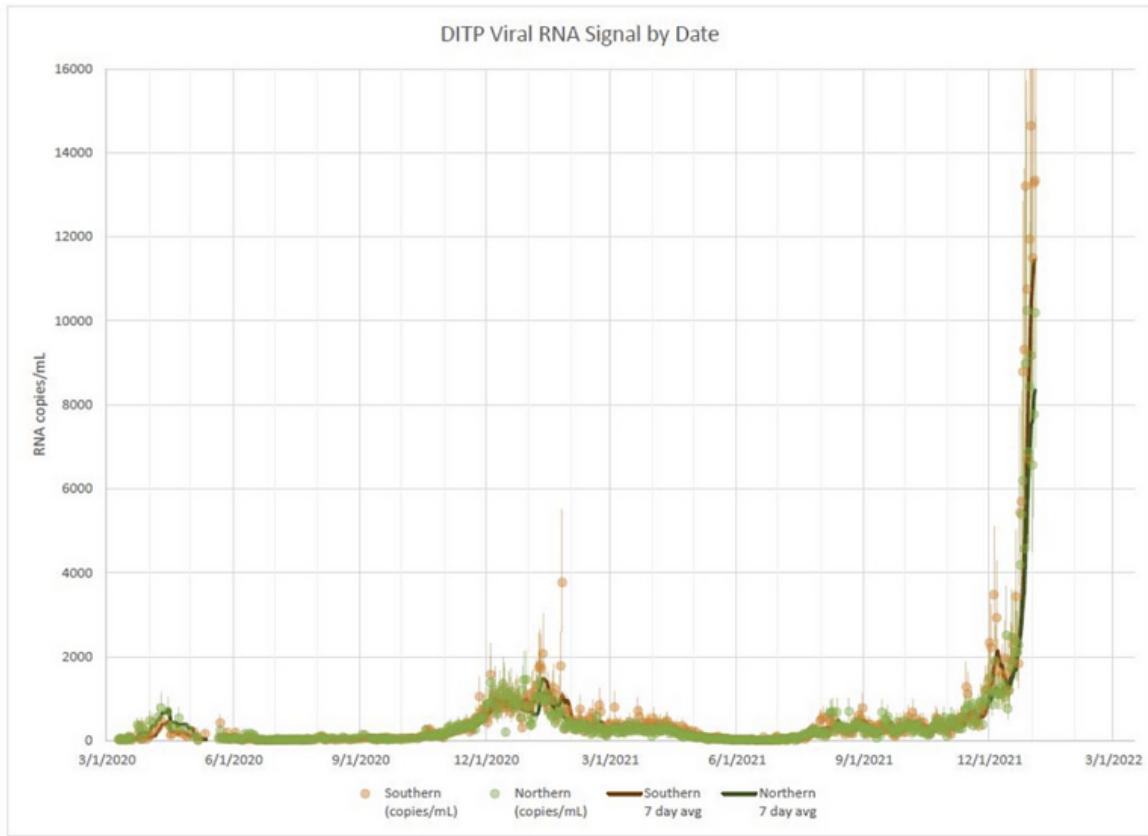
Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Nov 18-Nov 24	130,118	231,105	87,119	144,452	592,794
Nov 25-Dec 1	102,685	206,259	76,590	143,243	528,777
Dec 2-Dec 8	140,943	283,433	105,544	206,259	736,179
Dec 9-Dec 15	114,266	264,402	136,589	231,791	747,048
Dec 16-Dec 22	140,012	300,125	198,037	388,788	1,026,962
Dec 23-Dec 29	281,703	394,422	472,090	663,080	1,811,295
Dec 30-Jan 5	667,872	639,499	1,128,524	1,137,319	3,573,214

## Positive Tests by Region



[It looks like Manhattan has indeed peaked](#), and maybe Brooklyn, and maybe NYC overall. If that's right, then things will have held together well.

[Boston on the other hand looks like it's still headed upwards](#), currently at more than five times the previous peak.



[Washington D.C. schools did systematic tests, found 5.8% of the tests came back positive.](#)  
This was a lot of kids, so probably a lower rate than the city overall.

[A week ago fully one third of NBA refs were in Covid protocols.](#)

## Vaccinations

[CDC shortens length of time to wait for a Pfizer booster to five months.](#) Six months was, of course, yet another arbitrary round number. And now we have this handy easy, not confusing chart, note that in some places boosters are *mandatory* soon after they are permitted.

## Do I qualify for a COVID-19 vaccine booster and which one?

Which primary vaccine series did you complete?	Pfizer-BioNTech	Moderna	Janssen (J&J)
<p>You can get a booster if:</p> 	<p>It's been at least 5 months since completing the primary series AND you are:</p> <p>Age 12+</p>	<p>It's been at least 6 months since completing the primary series AND you are:</p> <p>Age 18+</p>	<p>It's been at least 2 months since completing the primary vaccination AND you are:</p> <p>Age 18+</p>
<p>If eligible, you can get a booster of:</p> 	<p>Pfizer-BioNTech* Moderna Janssen (J&amp;J)</p> <p>*Only Pfizer-BioNTech can be used as a booster in those age 12-17.</p>	<p>Moderna Pfizer-BioNTech Janssen (J&amp;J)</p>	<p>Janssen (J&amp;J) Pfizer-BioNTech Moderna</p>

For more information, visit [www.fda.gov/covid19vaccines](http://www.fda.gov/covid19vaccines).



The FDA decision to do this bypassed VRBPAC. They didn't consult their advisory committee. This new thing where when we don't need their input we don't ask for it? [Excellent](#).

[ACIP, alas, could not as easily be avoided, here is a recap of that.](#)

Which includes this unrelated gem:



Helen Branswell ✅ @HelenBranswell · 48m

...

20. Ironically, I just got a text from Vsafe asking me to do a weekly check in on how I feel after my booster shot.

I can't do it because the links Vsafe has sent me for the past 3 weeks don't work.

3

2

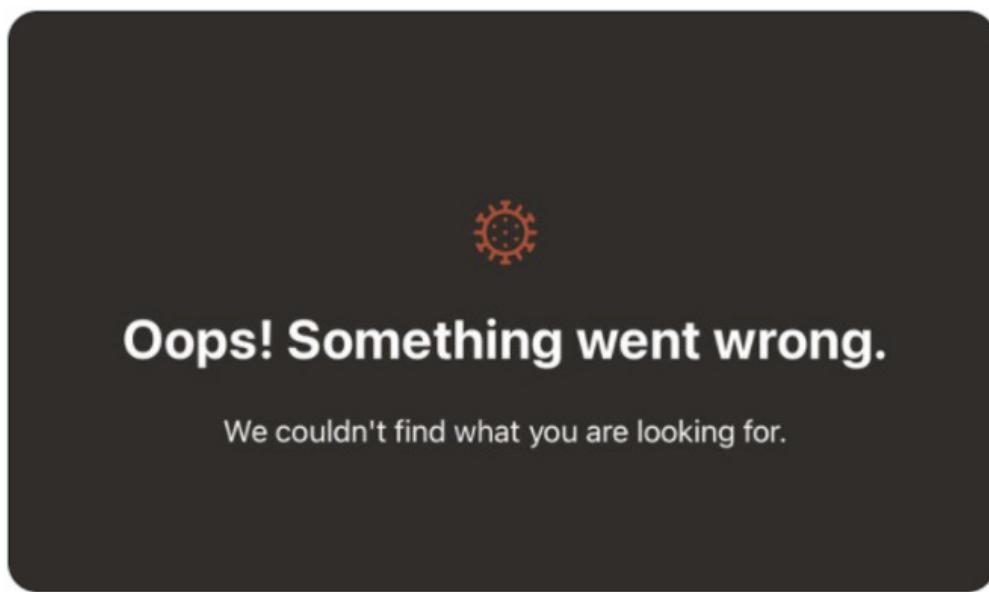
30



Helen Branswell ✅ @HelenBranswell · 46m

...

21. Chair Grace Lee is urging the public to continue to report to Vsafe. And I would -- except the links I'm sent take me to a page that looks like this.



So, anyway, yeah, [now you're worried about boosters limiting vaccine supplies for other countries?](#) What changed, exactly?

Also, it's one thing that we haven't approved Novavax, we have plenty of supply ourselves so we don't need it. It's another thing that we haven't approved *the manufacturing of Novavax?*

It is still unclear how many doses Novavax will be able to provide next year. A company spokesperson said its goal is to produce 2 billion doses in 2022. But there are lingering concerns among top Biden officials and global health advocates about Novavax's manufacturing facilities in the U.S. and Europe and whether the company can present the data necessary to obtain the regulatory approvals necessary to distribute doses made at those locations any time soon, according to the individual with direct knowledge of the administration's thinking on Novavax and another individual with direct knowledge of the company's communications with the administration. Both sources requested to remain anonymous to speak more freely on the matter.

I don't know what to say to that, at this point. The boosters, over the long run, increase demand and will result in matching capacity and supply. It's our delaying and denying of the booster plan that made this a problem at all. Meanwhile, we *continue* to hold up production

rather than working to expand it. If we wanted more vaccine doses, we would pay for them and allow them to be created. We didn't, so presumably we don't want this very much.

## Vaccine Mandates

[I do appreciate the honesty.](#)



BNO Newsroom ✅ @BNODesk · 12h

French President Macron: "The unvaccinated, I really want to piss them off. And so, we're going to continue doing so, until the end. That's the strategy"  
- REU/LP

425

1.7K

6.3K



BNO Newsroom ✅ @BNODesk · 11h

French President Macron: "I won't send (the unvaccinated) to prison, I won't vaccinate them by force. So we need to tell them, from January 15, you won't be able to go to the restaurant, you won't be able to down one, go for a coffee, go to the theater, go to the movies"

191

880

3.2K



That's the strategy. The strategy is to *punish* the people who won't follow his orders. To *piss them off*. And continue doing so, until the end. This isn't about precautions that make physical sense. This is about punishment and coercion, full stop. There's no longer any pretenses otherwise, on any level.

The *alternative* strategy that he needs to *explicitly say he won't do* is to vaccinate the non-compliant population by force, or imprison them.

That doesn't mean that there's no benefits in terms of prevention or slowing the spread, but he's no longer pretending that they are driving the decision.

And again, *I appreciate the honesty here*. It's important not to punish such honesty, while still opposing the policy if you don't think that we should be strong-arming people into getting vaccinated.

Similarly, if you were to [punish the unvaccinated by banning them from buying marijuana and hard liquor, you are once again owning your motives.](#)



Anarcho-Moses ✨ @ben\_r\_hoffman · 14h

Replies to [@EpistemicHope](#) [@CoughsOnWombats](#) and 2 others

The more you own your judgments and agenda, the less I'm procedurally threatened by anything you're doing.

[Italy isn't messing around either](#), but seeks to maintain plausible deniability.



PeterSweden ✅ @PeterSweden7 · Dec 29, 2021

...

JUST IN - Italy is BANNING the unvaccinated from:

- Trains
- Planes
- Busses
- Ships
- Hotels
- Ski lifts
- Outdoor restaurants
- Recreational outdoor activities
- Swimming pools
- And more.

If you told someone this 1 year ago it would be called a crazy conspiracy theory.

## Djokovic

[In other news, nope, no idea, I got nothing](#) he could possibly do.



The Daily Beast ✅ @thedailybeast · 5h

...

BREAKING: World No. 1 tennis player Novak Djokovic has reportedly been refused entry to Australia following a national outpouring of rage over his vaccine mandate exemption. It's unclear what Djokovic can now do to enter Australia

Yes, obviously he could have gotten the vaccine anyway to avoid such issues. It seems quite foolish not to do so, given the amount of travel he does, and I continue to not see this as a big deal.

Except here's the thing. [He did go through the required process](#) (WaPo via MR).

According to a person close to the tournament with direct knowledge of the sequence of events, Djokovic followed every step of the country's visa process properly. Moreover, the person said, Djokovic's medical exemption was granted with all identifying information redacted, ruling out the possibility of favoritism.

Djokovic has had Covid-19. I believe he's had it *twice*. One could respond with 'that's why you need a vaccine mandate' but now that both infections are in the past, there's no physical-world-model justification for denying him entry into Australia beyond the need to enforce the law. If this is additionally them about-facing and disobeying the law, it makes no sense in any way other than grandstanding and the punishment of someone that power doesn't like.

Letting him get on a plane with one understanding (assuming his understanding was reasonable) and then turning around and doing this is another level. Some would say the

cruelty is the point.

The argument in favor of keeping him out is that there's no systematic way to verify who has previous Covid infections, and we don't want to make an *exception* in this weird case where we can be confident the infections happened, or it wouldn't be fair. There aren't enough cupcakes for the whole class, so we throw away all the cupcakes. Or, of course, that you don't like him.

## Paxlovid

In order to get Paxlovid in New York, you need to have mild to moderate systems, be able to start treatment within five days of symptom onset, be over the age of 12 and weigh at least 40kg, and have one additional risk condition. I'd have preferred a much higher age minimum, but otherwise that seems sensible.

Then the question that seems to have no good answer. Give it to the vaccinated, or give it to the unvaccinated?

The solution is, whistle in the dark and [count race as a 'risk factor'](#)?

### Eligibility

Oral antiviral treatment is [authorized for patients who meet all the following criteria:](#)

- Age 12 years and older weighing at least 40 kg (88 pounds) for Paxlovid, or 18 years and older for molnupiravir
- Test positive for SARS-CoV-2 on a nucleic acid amplification test or antigen test; results from an FDA-authorized home-test kit should be validated through video or photo but, if not possible, patient attestation is adequate
- Have [mild to moderate COVID-19 symptoms](#)
  - Patient cannot be hospitalized due to severe or critical COVID-19
- Able to start treatment within 5 days of symptom onset
- Have a medical condition or other factors that increase their risk for severe illness.
  - [Non-white race or Hispanic/Latino ethnicity should be considered a risk factor](#), as longstanding systemic health and social inequities have contributed to an increased risk of severe illness and death from COVID-19

Eligibility requirements for treatments for the omicron strain of COVID, New York  
State Department of Health, Dec. 27, 2021

I can confirm, these are guidelines that were sent out.

I can also confirm that there were *no* other explanations given on what does or does not count as a risk factor.

Given the options of giving to each according to their needs or judging by the content of their character, they took at third option and went with the color of one's skin.

This seems to be the CDC's explanation:

Long-standing systemic health and social inequities have put various groups of people at increased risk of getting sick and dying from COVID-19, including many [people from certain racial and ethnic minority groups](#) and [people with disabilities](#).

Studies have shown people from racial and ethnic minority groups are also dying from COVID-19 at younger ages. People in minority groups are often younger when they

develop chronic medical conditions and may be more likely to have more than one condition.

They're also less likely to be vaccinated. Policies that discriminate on the basis of vaccination status, that many on the left support and that many cities have implemented, have a disparate impact on the basis of race.

It's important not to make *too much* out of this. If you have no 'medical condition or other factor' that increases your risk for severe disease, and presumably age counts as an 'other factor,' you're not at much risk, and you shouldn't be getting Paxlovid given our current shortage. If you *actually need* it, you'll be eligible. Remember that when it was time to show eligibility for boosters, something like 75% of people technically had at least one 'risk factor.'

Then again, if a bunch of people who don't need Paxlovid get Paxlovid (including anyone who is white and who realizes how easy it still is to have a 'risk factor') then it's a lot more likely there won't be any Paxlovid to give you.

One could reasonably point out that given systemic differences in ability to navigate the system, and their reduced willingness to actually *want* a new anti-viral Pfizer drug, that those who are eligible via this rule will almost certainly be much *less* likely to get Paxlovid than similarly at-risk white people. The same would presumably be true if we instead made being poor or a lack of education a risk factor.

No matter what you think of such arguments, we are explicitly allocating scarce life saving medicine on the basis of race. This seems worth noticing.

The good news is that the White House [doubled our Paxlovid order](#). The bad (non-news) is that they needed to do that in the first place.

## Ministry of Truth

[Marjorie Taylor Green has had her Twitter account permanently suspended for Covid misinformation \(WaPo\)](#).

This was in accordance with Twitter's '[five strikes and you're out](#)' policy. It's not easy to get permanently suspended this way. You don't quite [get twelve chances to clean up](#) before you're asked to leave, but you certainly can't say you were never warned.

Thus, she makes a good test case. What were the five strikes?

The first one seems to have been election-related rather than Covid-related, nothing you wouldn't expect.

We were unable to locate the second violation.

My assistant did find [this one](#), which we think (?) is the third one.

[This](#) has one example of a tweet that got her temporarily suspended in July:



**Marjorie Taylor Greene**

@mtgreenee



The controversial #COVID19 vaccines should not be forced on our military for a virus that is not dangerous for non-obese people and those under 65.

With 6,000 vax related deaths and many concerning side effects reported, the vax should be a choice not a mandate for everyone.

This Tweet is misleading. Learn why health officials recommend a vaccine for most people.



**Thomas Massie** @RepThomasMassie

Just received this from someone in the military regarding possible compulsory COVID vaccines:

"Good evening, Congressman. Just thought I'd pass word that >15% of the officers at my command will either retire early or resign their commissions if the mandate is given."

1 of 3

5:18 AM · Jul 19, 2021



That does seem rather like 'dangerous misinformation' about Covid-19, on two distinct counts.

Yes, there's a version of the 'under 65 and non-obese' claim that is true and important, but this is importantly different from that. I can see why one might find it unacceptable.

Then there's the 'vax related deaths' claim, and yeah, stating that as fact does seem exactly like what a 'dangerous misinformation' policy is designed to stop.

There were also a *lot* of anti-vax things said that *didn't* result in a strike, pretty much a constant stream of them.

The fourth warning was likely this one, saying that vaccines do not 'reduce the spread' and neither do masks.

**Marjorie Taylor Greene** • @mtgreenee · 19 / 10.  
The FDA should not approve the covid vaccines.

There are too many reports of infection & spread of #COVID19 among vaccinated people.

These vaccines are failing & do not reduce the spread of the virus & neither do masks.

Vaccine mandates & passports violate individual freedoms.

WaPo included this description of the final one, and it echoed that second claim from the third suspension, so there was very clear warning here.

The tweet at issue claimed coronavirus vaccine deaths were being “ignored,” citing the VAERS data. But the database itself warns that the information can be incomplete, inaccurate, coincidental or unverifiable, saying that the “reports alone cannot be used to determine if a vaccine caused or contributed to an adverse event or illness.”

That defense is quite the non-sequitur. It’s ‘misinformation’ to claim that things are being ignored in a database because that database says that its data might not be accurate or complete? That doesn’t sound like evidence against her position. Whenever I see such arguments used I update:

1. There might not have been good arguments.
2. Even if there were, they’re not going to be used to decide what to do next time.
3. If we can use this kind of justification to label something ‘misinformation,’ then that word means whatever power wants it to mean.
4. Thus, ‘dangerous misinformation’ is ‘information that power doesn’t like.’

I can’t even tell, from the description, whether her statement was even false.

I presume it was false, especially given the previous tweet, and that she was claiming a similarly massive number of (as far as I can tell, completely fictional) deaths from vaccinations. Which is misinformation, and is dangerous. I also think that censoring that kind of statement is a reasonable thing to consider doing. But the rules seem to consistently get written in a way that does not differentiate between this and a similar true or good faith statement, and instead give power the ability to censor whatever they dislike.

Over the long term, that policy does not go anywhere good. But it’s not like this should have come as a surprise.

[Reactions were as you'd expect](#). That first one sums up all sides quite well, I especially like the (intended or unintended) ambiguity here about who are or are not being fascists.



NationalizeAllHats  
@Popehat

...

They're fascists and they are increasingly open about it.



J.D. Vance ✅ @JDVance1 · Jan 2

Just totally absurd. These companies need to be crushed.

9:48



**Marjorie Taylor Greene** 🇺🇸 ✅

@mtgreenee

1,545 Following 465.7K Followers

## Account suspended

Twitter suspends accounts that violate the  
[Twitter Rules](#).

## NPIs Including Mask and Testing Mandates

[When others called on their cities to close, Mayor Eric Adams called upon New York City to instead open.](#)



Emma G. Fitzsimmons @emmagf · Jan 4

...

Mayor Eric Adams, asked on CNN about companies delaying return to offices, says: "It's time to open up."

He says people are falling behind on rent and low skill employees can't telecommute.

37

96

94



Interesting timing, but you know what they say, the second best time is right now.

[Some fake KN95s got distributed over the weekend](#). The 95 means you stop 95% of particles, and by all reports these very much do not do that.



Carolyn VanEseltine  
@mossdogmusic

...

If you received a "Fujian Pageone Garments Co., Ltd. KN95 Protective Mask (Non-Medical)" at your school: the CDC test report says this model's efficiency was between 25.20% and 45.80%.



John Bracey @MagisterBracey · 11h

ATTENTION MA EDUCATORS! DESE sent out bogus KN95s to every district over the weekend. "KN95" masks are meant to filter out at least 95% of all particles. According to the CDC, these masks tested as low as 25%.

What they are still good for, of course, is *fulfilling mask requirements*. And presumably many people will prefer the feel of these, exactly because they fail to be an effective mask. Adverse selection is a real problem here, unless what you care about is the requirement.

[NYC Testing cites for this week](#). Is there a reason this needs to be updated so frequently? Also clearly extremely incomplete.

It could be much worse, and some places it is, [here's an example of waiting a week for a PCR test](#). At which point, why bother?

Also, as a general principle, our procedures that decide *which particular things* to mandate or forbid (in terms of Covid-19 and NPIs, or anywhere else) is not well-principled, well-sorted or well-ordered, and often dumb rules are implemented instead of better rules.



jorbs @JolNrbs · 6h

one of the most recurrent and perplexing fallacies of the pandemic is "we shouldn't close x since y is open", "we shouldn't wear masks indoors if we're taking them off to eat anyway", etc., as though the germ theory of disease sometimes ceases to exist based on analogous context.

2

3

66



jorbs @JolNrbs · 6h

it may be logically inconsistent to do many of these things, but the answer is never "therefore the germ theory of disease doesn't exist." - the answer is generally "we shouldn't be doing either of these things" or "that's another battle we have to fight after this one" or w/e.

1

2

25



jorbs @JolNrbs · 6h

shocking take which is at odds with surprisingly many people: it's actually safer to go to a restaurant if people put on masks while they're walking around and take them off to eat.

we shouldn't be dining in restaurants anyway, but that doesn't make the first paragraph disappear



jorbs @JolNrbs · 6h

"it makes no sense to close schools if bars are open" WHAT? WHY? the germ theory of disease does not care if bars are open in your city or not, this does not change its physical properties and effects.

If X and Y physically interact such that X relies on Y, it can make sense to only do X if you also do Y. To extent you have control over both the decision on X and the decision on Y, and Y strictly dominates X, then you shouldn't do X without also doing Y. But the wisdom of our collective decision to do or not do X doesn't much matter when deciding whether to do inferior independent option Y. You should do the prevention that makes sense on the margin, and not the prevention that doesn't, in light of whatever decisions you're forced into and the conditions that exist, even if those decisions were deeply stupid.

That extends beyond Covid-19. Arguments *from consistency* should mostly be ignored if you're being forced to be consistent with decisions over which you have no control. Otherwise, once someone manages to sneak through one dumb decision, you'll end up with a tsunami of additional dumb decisions. There's always a dumb decision, in all directions, somewhere.

Also:



Alex Tabarrok  
@ATabarrok

...

Amen.



Matthew Yglesias ✅ @mattyglesias · Jan 3

I feel like at this point 99 percent of the population is either vaccinated and safe, under 5 and safe, or doesn't care about Covid and it's time for non-pharmaceutical interventions to basically go away. [twitter.com/dkthomp/status...](https://twitter.com/dkthomp/status...)

[Show this thread](#)

Given we've waited this long, I might wait a few weeks to be sure the hospitals will hold up in rural areas, but mainly, [yes, amen](#), and let private citizens make their own private decisions based on what's right for them.

## Think of the Children

[Should we mandate boosters for young children?](#) I mean, obviously no, we should not do that, it's too late for it to make a difference, the cost-benefit is unclear given how safe kids already are after two (or one, or zero) shots, and the amount of backlash would be severe. Yet the *first* question for anything approved is always 'should this be *mandatory*?'

[Chicago public school teachers were allowed to vote on whether schools should be open, so Chicago public schools are closed indefinitely.](#)

[A reminder of what record numbers of child hospitalizations means.](#)



Eliza Shapiro Retweeted



Dana Goldstein @DanaGoldstein · 3h

...

Have you heard that "record high numbers of kids are hospitalized with Covid"? What does that mean? In Chicago, an average of 7 children 0-17 per day in a city of 550,000 children, according to [@DrArwady](#) of the city Dept of Health.

New York City schools instead are determined to stay open, [so they're asking teachers to come back five days after testing positive, without a negative test and despite mild Covid symptoms \(article\)](#). If the symptoms are 'improving' then they don't even have to be 'mild.'

The new protocols come after the CDC slashed the isolation time for asymptomatic individuals from 10 to five days, with those people wearing a mask for the next five days. The move sparked some backlash and a myriad of memes.

"Many people are no longer contagious five days after diagnosis. And those who are can minimize their risk of infecting others by wearing a mask," Dr. Celine Gounder, a leading infectious disease doctor in the US, told Insider. "That means that they need to wear that mask diligently for an additional five days when around other people. But in that context, they are very unlikely to infect others."

Pairing the five-day guidelines with a negative test might also prove helpful, according to Gounder. Dr. Anthony Fauci has said that the CDC might recommend that asymptomatic individuals end their isolation after five days if they test negative.

[How many kids are coming in?](#)



**Eliza Shapiro** @elizashapiro · Jan 3

New: attendance was very low in NYC schools today, about 67 percent. Slightly higher than the last day before winter break.

...



**Eliza Shapiro** @elizashapiro · 22h

New: School attendance was at about 72 percent today, up from yesterday.

6

5

15

[Did you know that schools are 'one of the safest places' for children during the pandemic?](#)



**Leana Wen, M.D.**  @DrLeanaWen · 10h

Striking quote from [@PedsDanny](#) on [@AC360](#):

...

"Closure of public schools is the biggest public health failure of our lifetime.  
Why do we want to do that again?"

Schools remain one of the safest places for children from a [#covid19](#) transmission standpoint.

 212

 209

 551





**Evan Serpick**  @Serpis · 9h

Replies to [@DrLeanaWen](#) [@PedsDanny](#) and [@AC360](#)

...

"One of" the safest? How long is that list? Because home is pretty safe and that's where kids will be if not in school. What a nonsense talking point.

 22

 7

 45





**Leana Wen, M.D.**  @DrLeanaWen · 9h

...

Actually, most children will NOT be at home isolating alone, if not in school. They will be with other kids, babysitters, or on the streets. We have seen from many studies that when students become infected, it's not generally at school--it's in home and community settings.

 48

 18

 145



[And did you know that this is no way to grow up \(article it links to is NY Times\)](#)? As in the NY Times itself is now pushing the narrative (because everything in that paper is pushing a narrative) that the disruptions we've imposed have put our children into crisis, And That's Terrible, as opposed to before when it somehow wasn't terrible and you were instead terrible for objecting to it.

So weird that suddenly, now, things other than case counts matter and it's good to do a cost-benefit analysis about how much what you're doing is messing up children's lives. And all of this is suddenly 'not surprising.'



**David Leonhardt** @DLeonhardt · Jan 4

The widespread availability of vaccines also raises an ethical question: Should children suffer to protect unvaccinated adults — who are voluntarily accepting Covid risk for themselves and increasing everybody else's risk, too? Right now, the U.S. is effectively saying yes.

89

398

1.5K



**David Leonhardt** @DLeonhardt · Jan 4

To be clear, there are some hard decisions and unavoidable trade-offs. Covid can lead to hospitalization or worse for a small percentage of vaccinated adults, especially those who are older, and allowing children to resume normal life could create additional risk.

18

55

511



**David Leonhardt** @DLeonhardt · Jan 4

For the past two years, however, many communities in the U.S. have not really grappled with the trade-off. They have tried to minimize the spread of Covid — a worthy goal absent other factors — rather than minimizing the damage that Covid does to society....

17

120

803



**David Leonhardt** @DLeonhardt · Jan 4

... Communities have accepted more harm to children in exchange for less harm to adults, often without acknowledging the dilemma or assessing which decisions lead to less overall harm.

So it is not surprising that children are suffering so much.

[My memory is same as Nate's here](#), cost-benefit was anathema, [also yes all of this was obvious but denied](#). I also worry how much emphasis was placed on suicides, both because it's a lot of deaths and also because of the implication that the only way to rhetorically answer death is more death.



**Nate Silver**  @NateSilver538 · Jan 4

...

The thing I'll remember is people yelling at folks (me but also others who made the point more often) for merely suggesting we needed to consider the costs to closing schools as well as the benefits. If someone won't engage in cost-benefit analysis they're usually full of shit.



**David Leonhardt**  @DLeonhardt · Jan 4

3. Suicide attempts have risen, slightly among adolescent boys and sharply among adolescent girls. The number of E.R. visits for suspected suicide attempts by 12- to 17-year-old girls rose by 51 percent from early 2019 to early 2021, according to the CDC. cdc.gov/mmwr/volumes/7...

[Show this thread](#)

If you're a parent whose children were forced into a year of 'remote learning' in the name of pandemic safety, you might be wondering, *what changed all of a sudden?* If schools are *safer* than not schools, in addition to the mental health and learning disasters that come from 'remote learning' as implemented by a vindictive system looking to mimic the punishments of ordinary schooling rather than teach children, then [what the hell were we doing that whole time?](#)



**Karol Markowicz**  @karol · 15h

...

My 6yo was in 1st grade in NYC but the cut-off is different in FL and he's in K again, which makes me happy as I love that he will have a normal K experience. His first day was today and he came home very happy. A few minutes ago, his teacher called me to check in. (Cont)

119

336

3.8K



**Karol Markowicz**  @karol · 15h

...

Replying to [@karol](#)

Her: "He was great, really got into the rhythm of the class. And then we went to the rug..."

Me: "They go to the rug?!"

Her: "Yes, why?"

Me: "In NYC they try to keep the kids apart and he hasn't been on the rug..."

My 6yo listening: "Not since Pre-K"

Her: GASP.



PoliMath @politicalmath · 14h

This is a really common experience for my family in Chattanooga

...

We tell people what actually happened to our kids with school for the 2020-2021 school year and even the most COVID cautious people stare at us, mouths agape.

Meanwhile, schools are acting terrified in all sorts of different ways, which doesn't seem necessary if they're so safe. But also they're indoors, often poorly ventilated, often with many people in the same room, so how in the world could they be 'one of the safest' places to be? What physical model could this be describing? Or is this describing something else entirely?

It's no wonder some are [feeling more than a little gaslit](#).

While we also [scare the children](#). The message I want such a child to get is that getting Covid-19 is nothing for him to be afraid of, he'd be totally fine. I worry telling him to eat quickly sends the wrong message.



Meg S @megsullivanMD · 40m

Replies to [@TheBlondeRN](#) and [@tmprowell](#)

...

My 8yo was crying this morning because he is afraid of getting covid. Today is their first day back from the break. I tried to console him by reassuring him he's vaccinated and told him to eat his lunch quickly. ❤️

Also, crazy talk here, but what is so sacred about lunch? We're taking kids, putting them in jackets and having them eat on the sidewalk, when we could simply... give them a bigger breakfast and then have them wait until they get home? Is that so completely insane? At about that age I remember being sent to a camp that served non-kosher food and my parents wouldn't let me eat it and I was the kind of kid who followed such rules, so I often ended up eating once a day, and this wasn't only fine, *it taught me a useful life skill*, that I use almost every day.

So you know what, how about let's *not* do lunch.

I do realize it's different for those who need subsidized school lunches, because our society is pretty terrible about how it gets its children in need fed.

The only good counterarguments to all this that I've seen are:

1. [Schools as they typically exist are terrible](#).
2. Whether or not schools are terrible, [kids can and do recover from missing even quite a lot of school, all the time](#).

If you think typical schools are dystopian nightmares, you'd be inclined to welcome a break from them. The thing is, even if schools are such nightmares, *remote learning as implemented by the United States was clearly much worse*. It carefully recreated all the negative aspects of school such as being constantly forced to sit motionless all day as a dominance exercise, and adds new ways and new levels, while failing to capture the benefits of school, such as social contact with other humans, or making any attempt to teach things.

This is not the kind of thing that old 'missing school' studies were measuring:

When I had a parent teacher conference for my daughter's PreK the teacher mentioned how social she is and how she always wants to hold hands with other little ones which is incredibly typical of this age but the teacher said she has now noticed that most of the kids won't hold hands or are very reticent (which she attributes to the pandemic). She said she always makes sure my daughter knows she can hold the teacher's hand and there is one other little boy whose hand she can reliably hold. But the sadness of it crushes me. I'm so happy my daughter has remained unafraid of social interaction but sad for her that she sees so many other little ones wary of her.

And of course I am so sad for these other small children shouldering these burdens and fears. There was a time when we didn't know enough and this kind of extreme caution was reasonable and warranted. We know enough now.

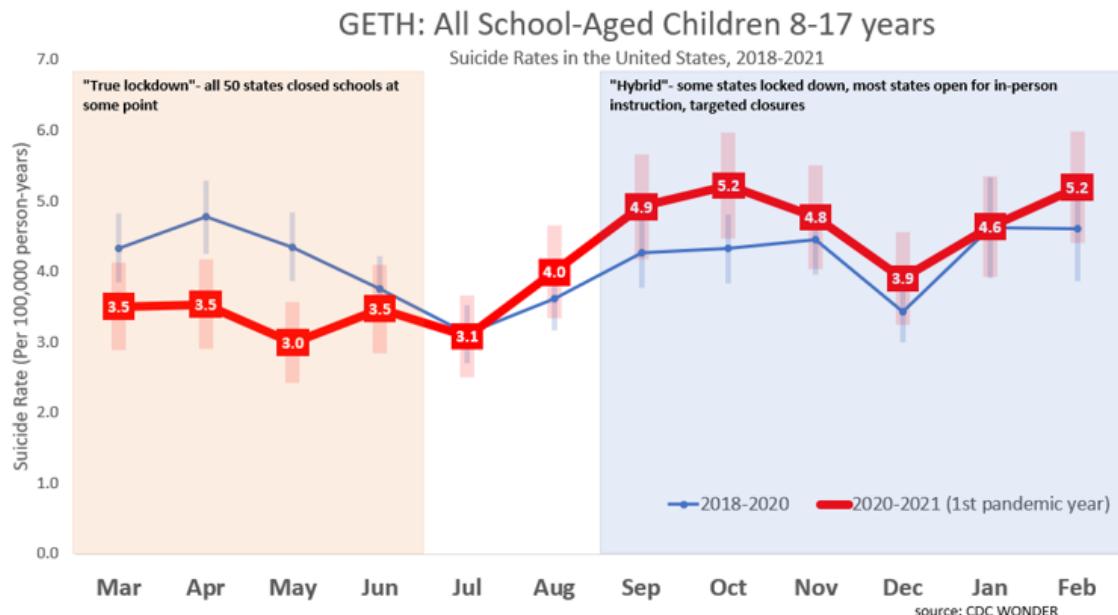
I've seen some of this as well, although thankfully less.

[Bryan Caplan shares a story from student David Park of what 'the college experience' is like in the Covid era.](#)

My model is:

1. If kids got to stay home, use Khan Academy, play games, read books and listen to the radio, play outside with the other kids and otherwise be the age they are, help around the house and learn practical stuff and from their parents and others around them, that would be like 'missing a lot of school' – if the alternative was *be allowed to be a child* – then the kids would very much be all right.
2. If the kids are forced into 'remote learning' then you're torturing them all day while both failing to teach them much of anything other than how screwed up everything is, and also crowding out opportunity and energy for either living life or otherwise being able to learn. You're also combining that with not letting them see other children or other people's faces. Meanwhile, we're teaching all our kids to be obsessed with the 'danger' of catching Covid-19, as something central to life.
3. Yes, *this is going to turn out very badly*.
4. There's a reason I home schooled our kids rather than doing remote learning.

[Matching all this is the suicide data.](#)





**Tyler Black, MD** @tylerblack32 · Dec 14, 2021

...

Replying to [@tylerblack32](#)

/8 Yes, I will repeat that again.

EVERY SINGLE YEAR from 1999-2019, High School kids in the US died of suicide about 40% more frequently on March-Jun vs July. In 2020, when:

- \* all schools were shutdown
- \* during a new pandemic

the Mar-Jun rate equaled the lowest rate (July)!!

Normally when kids are in school, they are more likely to kill themselves. This year, when we closed the schools, *kids killed themselves less*. Then, when we went to 'remote learning' *kids killed themselves more*.

From the anecdotal experiences I've heard about, and I mean *all of them*, it's not hard to see why.

In a related question, [is this schools having a problem or is school a solution?](#)



**Walid Gellad, MD MPH**  @walidgellad · 1h

...

Will wade into school debate in 1 tweet.

30% positivity is nothing US schools have ever seen. You can do all the mitigation you want, but I think there could be a lot of in-school transmission with omicron.

I fear our response is an overcompensation to prior school closures.



**Harry Thomas** @DrHarryThomas · 16h

1,200 students and staff @AustinISD underwent Covid tests today...

ONE THIRD tested positive.

School starts in 2 days.

H/T @MarietteHummel [kxan.com/news/education...](#)

Does this mean the schools are dangerously unsafe and have to close, because a third of students are positive? Or does it mean that *not being in school* is dangerously unsafe, and the schools need to (or might as well) stay open?

Should you close the schools? Should you close to schools to everyone who tests positive? Or should you close the schools to everyone *who tests negative*?

Now hear me out.

If you put the two thirds of students who you *don't* think have Covid-19 into a classroom, you're going to be wrong, because on average you'll have one student who got infected yesterday and one from two days ago. So everyone gets exposed anyway.

Whereas if you put the one third of students who *definitely do* have Covid-19, since the false negative rate is super low, into a classroom, then everyone already has it, and no one gets exposed. Of course, you'd then need to ensure the teachers *also* have it, but that presumably a lot of them do as well, and they're not exactly otherwise in high demand, so sure, why not?

The best part of this plan is that *over time* it improves. Everyone who has had Covid-19 *in the last few months* is also fine, so pretty soon the classroom is full again. Then, once things settle down, the kids who were always negative can come back.

Fun little exercise. There's so much room to actually improve things if we cared mostly about outcomes and not what things looked like, as many good reasons as there are that this absolutely does not work. [So just kidding, baby. Unless you gonna do it.](#)

[Yet another standard ‘colleges are imposing crazy requirements’ post](#), this one a guest post on the Bari Weiss substack. Pre-Omicron, such posts were transparently correct, the universities were doing Obvious Nonsense. In March, any remaining such policies will once again be Obvious Nonsense.

And also, right now, such policies are still Obvious Nonsense, because they are neither necessary to protect anyone nor are they sufficient to stop spread on campus. If you don't want your students infected in January, you have zero options. You *do* have the option to ensure they are not infected *on campus* by not opening the campus, in which case the infections will *not be your fault*, but the infections will still happen.

[When there are 146 positives upon return to campus like there were at Stanford](#) that's not on you. That presumably is what such policies are about.

Having been soundly trounced by Georgia, [the University of Michigan goes from daily to weekly testing](#), buying up to six days of the school being open. This makes the data much less useful for making good decisions, but potentially less useful for making bad decisions as well. Value of information can sometimes be negative.

And that's why [Yale students are under quarantine until February 7 and not allowed to eat at outdoor restaurants.](#)

## Zeynep's First Law

[New Rule just dropped.](#)



**Avi Zenilman** @avizvizenilman · 3h

Being honest about scarcity is a better policy than pretending the scarce resources are less useful than they are

...



**A. David Paltiel** @ADPaltiel · 4h

CDC can't recommend antigen tests if they remain in such short supply. How elitist would it look if the only people who could follow federal guidance were well-heeled Ivy League colleges and individuals with enough disposable income to stockpile? [washingtonpost.com/politics/2022/...](https://washingtonpost.com/politics/2022/...)

[Show this thread](#)

7

111

437



**jj** ⏳💡🔭 @murchiston · 3h

Pandemic of second order effect guessing congenitally dishonest incompetents.

...

1

11

8



**zeynep tufekci** ✅ @zeynep · 2h

Replying to [@murchiston](#) and [@avizvizenilman](#)

Zeynep's law: Until there is substantial and repeated evidence otherwise, assume counterintuitive findings to be false, and second-order effects to be dwarfed by first-order ones in magnitude.

...

1

21

82



**zeynep tufekci** ✅ @zeynep · 2h

Aka if the particle is going faster than light, check your cables, and if you're thinking what if better is worse, check your thinking.

...

0

11

23



[What, never? Well, hardly ever.](#) And not by default.

I think she may have dropped not one but two, as the next section will show.

## CDC Guidelines Revision Part 3

[It doesn't look good.](#)



**Jessica BrodeFrank** @jessicabrode · Dec 31, 2021

...

CDC says to stay home this New Years Eve. There's no need to congregate in mass to watch the ball drop, when they've been publicly dropping the ball for 22 months now.

[Bob Wachter suggests CDC should have told the truth about why it was issuing its guidelines](#), and isn't even suggesting the whole truth. He's suggesting merely that they not lie about tests not working on day five, not the crazy talk of admitting that they're compromising a bit to keep society running.

[This or something similar seems like what all the reasonable people are suggesting makes sense](#), it's simple, it's being done elsewhere, it makes perfect sense.



**Ashish K. Jha, MD, MPH** ✅ @ashishkjha · 7h

What would ideal COVID Isolation guidance be?

Once rapid antigen tests are cheap & ubiquitous

1. Isolate for first 5 days
2. Then test daily
3. Negative X 2 days? End isolation. Good to go
4. Positive? Remain in isolation until 2 negs or 10 days

I think its that simple

I am not sure we can afford to wait for the second negative test, nor do I have a good idea how much extra value that second test provides, but even one would be a large improvement when it's available.

And this keeps it simple, as opposed to [attempts to explain current policy](#).



Tiffany C. Li ✅ @tiffanycli · 17h

...

Stay indoors. But also return in person. Wear a mask. Not that one. The expensive one, that you can't find. Take rapid tests. Which you also can't find. But if you find them, don't buy them. Rapid tests don't work. You need PCR. There are zero appointments in your area.

1K

49.5K

226.2K



[We also have The Daily Show explaining in video form \(~1min\).](#)

[Cause yeah, this isn't simple](#) at least as written ([link to policy](#)).

## Ending isolation for people who had COVID-19 and had symptoms

If you had COVID-19 and had [symptoms](#), isolate for at least 5 days. To calculate your 5-day isolation period, day 0 is your first day of symptoms. Day 1 is the first full day after your symptoms developed. You can leave isolation after 5 full days.

- You can end isolation after 5 full days if you are fever-free for 24 hours without the use of fever-reducing medication and your other symptoms have improved (Loss of taste and smell may persist for weeks or months after recovery and need not delay the end of isolation).
- You should continue to wear a [well-fitting mask](#) around others at home and in public for 5 additional days (day 6 through day 10) after the end of your 5-day isolation period. If you are unable to wear a mask when around others, you should continue to isolate for a full 10 days. Avoid people who are [immunocompromised or at high risk for severe disease](#), and nursing homes and other high-risk settings, until after at least 10 days.
- If you continue to have fever or your other symptoms have not improved after 5 days of isolation, you should wait to end your isolation until you are fever-free for 24 hours without the use of fever-reducing medication and your other symptoms have improved. Continue to wear a [well-fitting mask](#). Contact your healthcare provider if you have questions.
- Do not travel during your 5-day isolation period. After you end isolation, avoid travel until a full 10 days after your first day of symptoms. If you must travel on days 6-10, wear a [well-fitting mask](#) when you are around others for the entire duration of travel. If you are unable to wear a mask, you should not travel during the 10 days.
- Do not go to places where you are unable to wear a mask, such as restaurants and some gyms, and avoid eating around others at home and at work until a full 10 days after your first day of symptoms.

If an individual has access to a test and wants to test, the best approach is to use an [antigen test](#)<sup>1</sup> towards the end of the 5-day isolation period. Collect the test sample only if you are fever-free for 24 hours without the use of fever-reducing medication and your other symptoms have improved (loss of taste and smell may persist for weeks or months after recovery and need not delay the end of isolation). If your test result is positive, you should continue to isolate until day 10. If your test result is negative, you can end isolation, but continue to wear a [well-fitting mask](#) around others at home and in public until day 10. Follow additional recommendations for masking and restricting travel as described above.

<sup>1</sup>As noted in the [labeling for authorized over-the counter antigen tests](#) : Negative results should be treated as presumptive. Negative results do not rule out SARS-CoV-2 infection and should not be used as the sole basis for treatment or patient management decisions, including infection control decisions. To improve results, antigen tests should be used twice over a three-day period with at least 24 hours and no more than 48 hours between tests.

Note that these recommendations on ending isolation **do not** apply to people with severe COVID-19 or with weakened immune systems (immunocompromised). See section below for recommendations for when to end isolation for these groups.

## Ending isolation for people who tested positive for COVID-19 but had no symptoms

If you test positive for COVID-19 and never develop [symptoms](#), isolate for at least 5 days. Day 0 is the day of your positive viral test (based on the date you were tested) and day 1 is the first full day after the specimen was collected for your positive test. You can leave isolation after 5 full days.

- If you continue to have no symptoms, you can end isolation after at least 5 days.
- You should continue to wear a [well-fitting mask](#) around others at home and in public until day 10 (day 6 through day 10). If you are unable to wear a mask when around others, you should continue to isolate for 10 days. Avoid people who are [immunocompromised or at high risk for severe disease](#), and nursing homes and other high-risk settings, until after at least 10 days.
- If you develop [symptoms](#) after testing positive, your 5-day isolation period should start over. Day 0 is your first day of symptoms. Follow the recommendations above for [ending isolation for people who had COVID-19 and had symptoms](#).
- Do not travel during your 5-day isolation period. After you end isolation, avoid travel until 10 days after the day of your positive test. If you must travel on days 6-10, wear a [well-fitting mask](#) when you are around others for the entire duration of travel. If you are unable to wear a mask, you should not travel during the 10 days after your positive test.
- Do not go to places where you are unable to wear a mask, such as restaurants and some gyms, and avoid eating around others at home and at work until 10 days after the day of your positive test.

If an individual has access to a test and wants to test, the best approach is to use an [antigen test](#)<sup>1</sup> towards the end of the 5-day isolation period. If your test result is positive, you should continue to isolate until day 10. If your test result is negative, you can end isolation, but continue to wear a [well-fitting mask](#) around others at home and in public until day 10. Follow additional recommendations for masking and restricting travel described above.

That's a lot of words that effectively say 'isolate for 5 days, then mask for 5, and if you have an actively positive test after 5 days, then isolate for the full 10 days.'

It's a sign both of the CDC's inability to cleanly communicate, and the amount of trust they've lost, that everyone is calling this complicated, and calling for 'simpler' things that are more complicated than what this is *trying* to do.

[Meanwhile, the CDC's own guidelines say that 31% of people are still infectious on day 5.](#)

It's also worth noting that if you're exposed to Covid-19, you're asked to isolate about as much as if you test positive, if anything *longer* because they have you test afterwards since who knows when you might or might not have gotten infected. So it's 'better' to actually be positive than be 'exposed', even though basically everyone not working hard to avoid exposure is exposed.

[Or consider their planned new policy.](#)



**Christopher Ingraham** @cingraham · 12h

The CDC recommends you lock yourself in a room with two brothers. One only tells the truth, the other only tells lies. One brother says you must take a test to leave the room, the other says the test is not necessary. The wrong choice leads to death. You may ask one question

...

[There's also always the incentives.](#)



**Scott Lincicome**

@scottlincicome

...

Heckuva incentive system here.



**Kaitlan Collins** @kaitlancollins · 13h

The CDC has updated its isolation guidance. They still don't recommend a rapid test after 5 days, but say if you take one and it's positive, isolated people should continue isolating for 5 more days. Those who leave isolation are urged not to fly on planes or eat in restaurants.

[Show this thread](#)

Of course, no one knows what the rapid test said, so if you choose not to look it doesn't *actually* get you off of any real hooks. [It's Not the Incentives, It's You.](#)

As an example of how this works in practice, [what you're supposed to do if you test positive in the DC area.](#)

[Anyone else want to give it a shot? How about the CDC director?](#)



**zeynep tufekci** ✅ @zeynep · 5h

Are you confused about the zigzagging CDC messaging on rapid tests? ...

I wish the CDC would consult with Dr. Walensky circa 2020, who explained it brilliantly.

This CDC seems to be hoping you figure things out by yourself, and good luck finding a test. New:

-----

So what did 2020 Walensky recommend? “The antigen test is ideally suited to yield positive results precisely when the infected individual is maximally infectious,” she and her co-author concluded.

The reason is that antigen tests respond to the viral load in the sample without biologically amplifying the amount and being able to detect even viral fragments, as P.C.R. tests do. So a rapid test turns positive if a sample contains high levels of virus, not nonviable bits or minute amounts — and it’s high viral loads that correlate to higher infectiousness.

What about the objection that rapid antigen tests don’t always detect infections as well as P.C.R. tests can?

The 2020 Walensky wrote that the F.D.A. shouldn’t worry about “false negatives” on rapid tests because “those are true negatives for disease transmission” — meaning that people are unlikely to spread the virus even if they have a bit of virus lingering. In other words, the fact that rapid tests are less likely to turn positive if the viral load isn’t high is a benefit, not a problem.



**zeynep tufekci** @zeynep · 5h

...

Replies to [@zeynep](#)

It's 2022! Disappointed with the "masks aren't useful for you because we don't have them" vibes around rapid tests, and CDC justifying it by saying tests would provide "a false sense of security".

Denying tools to the public with a claim they'd otherwise be reckless isn't okay.

The threat of a "false sense of security" has been used against everything from seatbelts to teaching young kids how to swim (because that would supposedly encourage parents to stop watching their children in the water!). Research and common sense shows what one would expect: Safety measures make people safer and people who choose to use them are looking to be safer — if anything, they do more of everything. (Parents should watch their young children in the water, but [kids who learn to swim are less likely to drown](#).)

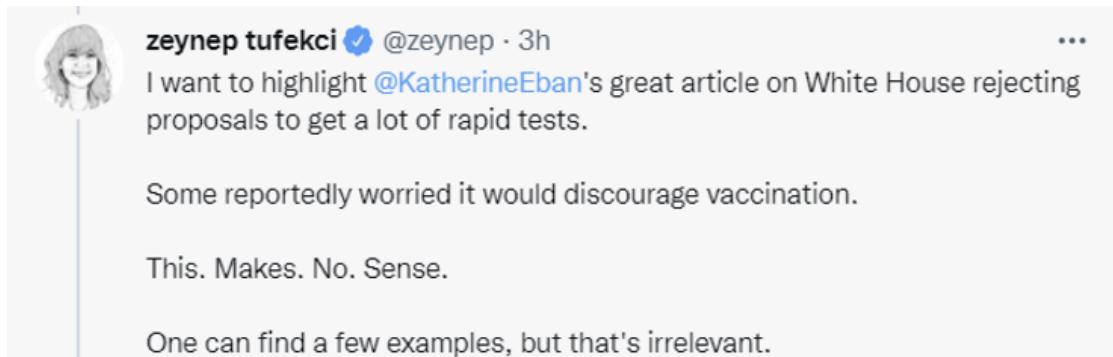
That's why it was extra disappointing to hear Walensky argue recently that "if you got a rapid test at five days and it was negative, we weren't convinced that you weren't still transmissible. We didn't want to leave a false sense of security. We still wanted you to wear the mask."

To start with, what if you were to test positive? People who test negative are less likely to transmit the virus — so even if Walensky's argument were true and these people would then not be using masks, this would be less of a problem than having an infectious person in public when a rapid test could have kept him or her in isolation.

Besides, if health officials told people to wear a mask for five more days even after they tested negative on the fifth day, responsible people would likely still do so. Extra information doesn't automatically turn responsible people into irresponsible ones.

That might want to be her second law, e.g.:

**Zeynep's Second Law:** Extra information doesn't automatically turn responsible people into irresponsible ones.



zeynep tufekci  @zeynep · 3h

I want to highlight @KatherineEban's great article on White House rejecting proposals to get a lot of rapid tests.

Some reportedly worried it would discourage vaccination.

This. Makes. No. Sense.

One can find a few examples, but that's irrelevant.

([Link to article.](#))

[California moves to 5 days and adds the obvious testing requirement.](#)

[Tom Frieden actually explains and defends the CDC's actual policy](#), considers it reasonable because tests aren't always accurate. This doesn't seem like a good argument, but the implicit 'we must muddle through as best we can' carries weight.

## What, Me Worry?

[In today's no s\\*\\*\\* sherlock headline news:](#)

# Worrying about COVID-19 leads to poorer choices, mental decline

I mean, I knew that, and I'm *still* willing to bet this one doesn't replicate, because, I mean, they're doing stuff like this, and that stuff is meaningless and doesn't replicate, it's the law.

For the tests examining how well people maintain information, participants had to match pairs of numbers and symbols according to a specific set of rules. Study authors also examined the group's [decision-making skills](#) using a test of risk management. Each person could either choose a "certain" option where they definitely won \$75, or a "risky" option where the odds of winning \$0 were only 25 percent and the odds of winning \$100 were 75 percent.

Those are two options each with an expected value of \$75, so I have no idea what decision making they plan to test there?

The team also discovered that worrying [about COVID-related issues](#) distorted a person's ability to evaluate certain risks. They underestimated the chances of likely outcomes occurring (like winning \$100) and overestimated the chances of unlikely possibilities occurring (like winning nothing at all). The team believes this impact on decision-making abilities may influence some people when they're considering certain topics — like getting the [COVID-19 vaccine](#).

I can certainly believe (but also could almost as easily disbelieve, or believe the opposite) that people who worry a lot about Covid-19 also obsess over other small 'risks' but what

does it mean to ‘underestimate the likelihood’ of something in this context? I tried to actually skim [the paper](#), to no avail, leaving even more confused.

[In front of a podium that says “Early Treatment Saves Lives,” the Florida surgeon general says it’s time to move past this mindset of testing.](#)

Which is a great joke, but also what he’s saying is that one can’t plan one’s life around constant testing. I very much agree. Testing is important but there are those who have gone completely overboard. ‘

[In related news, a study finds getting vaccinated is good for mental health](#), reducing symptoms of depression and anxiety by 30%.

## Hospitals

A lot of health care workers are sick, but that’s not the only reason staffing is short.

[The health care quit rate has risen to 3% per month.](#) It’s an impossible job that’s only getting harder.



**Heather Long** @byHeatherLong · 4h

...

To me, this is the most alarming chart from the Great Resignation:  
Skyrocketing quits in health care

Doctors, nurses, aides & more are burned out and we're still in a pandemic

Healthcare quit rate:

March: 2.3%

April 2.5

May 2.4

June 2.5

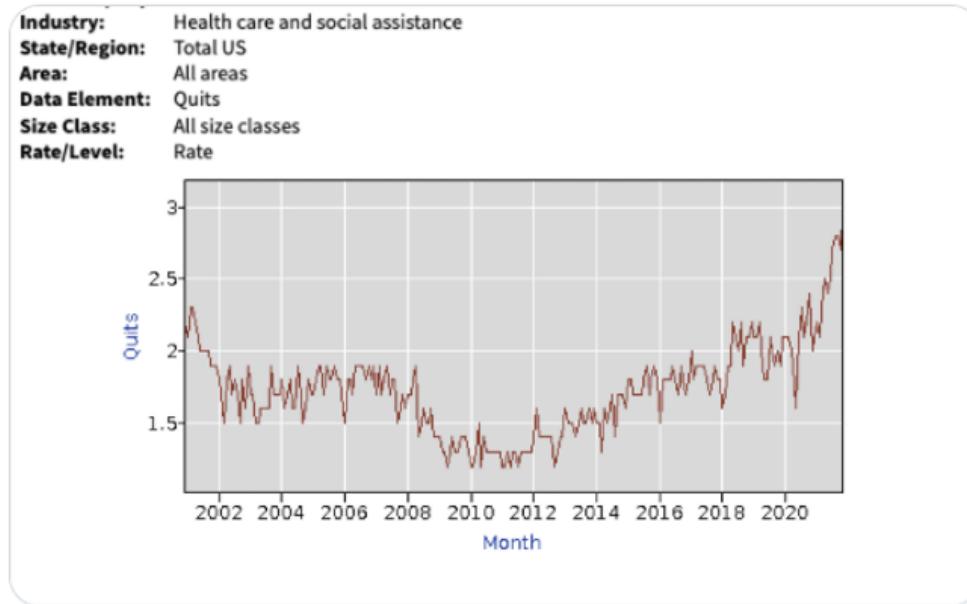
July 2.7

Aug 2.8

Sept 2.8

Oct 2.7

Nov: 3%





Natalie E. Dean, PhD ✅ @nataliexdean · 14h

Target audience for wonky observations reached.

...

While there is a decline over time in % of COVID patients needing ventilation, some of that is because there are more incidental patients (due to very high community prevalence!), and these patients have v low risk of ventilation.

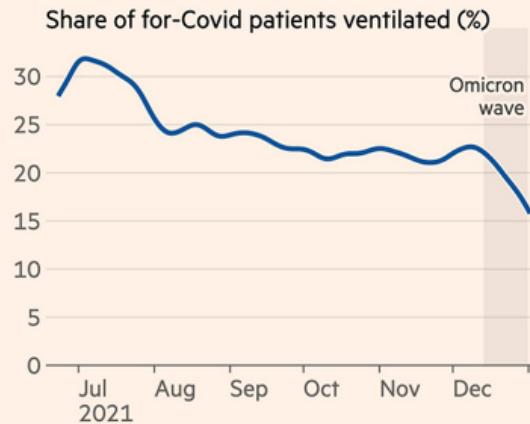
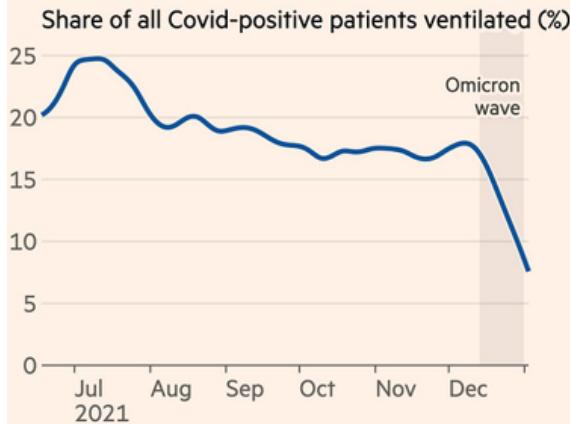


John Burn-Murdoch ✅ @jburnmurdoch · 22h

One slightly wonky observation re: steep reduction in share of London patients on ventilators:

Using two different patient denominators, you can see that a lot of the drop is due to incidental patients, who are still a genuine concern but do not appear to have ~any risk of vent.

A large portion of the observed reduction in ventilator rates is due to incidental patients, but a decline is visible even with 'for-Covid' patients only



Source: NHS England

FT graphic by John Burn-Murdoch / @jburnmurdoch

© FT

A large portion of the observed reduction in ventilator rates is due to incidental patients, but a decline is visible even with 'for-Covid' patients only

Percentage change in the share of Covid patients on ventilators, by patient definition



Source: NHS England  
FT graphic by John Burn-Murdoch / @jburnmurdoch  
© FT

Right now, something like 15.5% of 'for-Covid' patients are ventilated, whereas 7.5% of 'with-Covid' patients are ventilated, so something like ~45% of all patients that are 'with-Covid' are also 'for-Covid.' Looking back at the start of the graph, we see 20% ventilated versus 28%, so that implies ~70% of all patients were 'for-Covid.' That's a big drop.

It would be good to compare the number hospitalized 'with-Covid' with the number you'd expect from the baseline population, but we don't have the baseline number, so we can't do that analysis, unless there's data sources I don't know about.

On top of that, even the 'for-Covid' patients are being ventilated substantially less, about -50% from July 2021, versus about -30% at the start of the Omicron wave, so an additional -25% from there.

[Contrast that with this note.](#)



BNO Newsroom @BNODesk · 9h

...

Data from England shows that 67% of patients "with" COVID-19 are being treated because of COVID, which is down from 75% in October. In a statement, NHS England explains that COVID can still play a significant role even if it's not the primary diagnosis

The majority of inpatients with Covid-19 are admitted as a result of the infection. A subset of those who contract Covid in the community and are asymptomatic, or exhibited relatively mild symptoms that on their own are unlikely to warrant admission to hospital, will then be admitted to hospital to be treated for something else and be identified through routine testing. However these patients still require their treatment in areas that are segregated from patients without Covid, and the presence of Covid can be a significant co-morbidity in many cases. Equally, while the admission may be due to another primary condition, in many instances this may have been as a result of contracting Covid in the community. For example research has shown that people with Covid are more likely to have a stroke ([Stroke Association](#)); in these cases people would be admitted for the stroke, classified as 'with' Covid despite having had a stroke as a result of having Covid.

Eyeballing the graph has the number going from 70% to 45%, the stat here says 75% to 67%. The 70% vs. 75% is a rounding error in context, but the 45% vs. 67% is bigger, there's no way to reconcile that with the graph above.

All such graphs need to be adjusted for the relative ages of the populations, to get the full context, as my understanding is that the Omicron wave still skews relatively young, so that's another way in which it's still going to get worse before it get better.

Also this data point: [Los Angeles hospitals are two thirds 'with Covid' versus one third 'for Covid.'](#)

In terms of that, [nursing homes are only now seeing the kinds of staff shortages and rates of illness that lead to more cases](#). To extent that things look good at first and then get worse when it seems like they should be getting better, this dynamic and similar things like it are the most likely reason.

## Long Covid

[The German study about Long Covid has effectively been replicated in France, and the replication was successful.](#)

An all time great headline.

January 4, 2022

## Belief in Having Had COVID-19 Linked With Long COVID Symptoms

Anita Slomski

JAMA. 2022;327(1):26. doi:10.1001/jama.2021.23318

Claiming to have Covid-19 was correlated with claiming to have Long Covid.

*Actually having Covid-19 was not correlated with anything other than anosmia.*

Restricting the analyses to participants with a positive belief and attributing their persistent symptoms to COVID-19 showed a positive serology test result to be associated only with anosmia (OR, 2.97; 95% CI, 1.58-5.57) (eTable 7 in [Supplement 1](#)). Similarly, confirmation of the diagnosis by a laboratory test or by a physician (vs the response, "No, but I think I had it," and excluding participants who answered "I don't know") was also associated only with anosmia (OR, 4.29; 95% CI, 1.92-9.58) (eTable 7 in [Supplement 1](#)).

Their conclusion is polite but brutal:

The results of this cross-sectional analysis of a large, population-based French cohort suggest that physical symptoms persisting 10 to 12 months after the COVID-19 pandemic first wave may be associated more with the belief in having experienced COVID-19 infection than with actually being infected with the SARS-CoV-2 virus. Although our study cannot determine the direction of the association between belief and symptoms, our results suggest that further research regarding persistent physical symptoms after COVID-19 infection should also consider mechanisms that may not be specific to the SARS-CoV-2 virus. From a clinical perspective, patients in this situation should be offered a medical evaluation to prevent their symptoms being erroneously attributed to COVID-19 infection and to identify cognitive and behavioral mechanisms that may be targeted to relieve the symptoms.

I'm updating quite a bit in the direction of Long Covid mostly (not entirely) being people often having health issues, and them then blaming those health issues on having had Covid, rather than on Covid causing long term symptoms often in people who didn't have severe disease. No question that those with severe disease often have lasting damage, but yeah, if you feel like it finally happened and was no big deal, it was almost certainly no big deal.

## In Other News

Looking back, no one could have predicted that politicians and regulators would betray our trust, or that [repeatedly betraying the public's trust would destroy public trust](#).

[Man begs for positive Covid test to avoid visiting relatives.](#)

[Scott Lincicome feeling pretty good about his radicalization](#) regarding repeated government failure and obstruction. I am Bart Simpson's chalkboard repeatedly writing "I will not suggest using the free market to supply things people want."

[I don't know what this is about but seems potentially quite the red flag, and worth passing along.](#)



Oedipa Maas @bridgietherease · 8h

...

I know someone with a grandmother in a long term care facility in CA. they're saying she's not allowed to have visitors due to lockdown (not sure if it's state regulation or their call) .... unless they move her over to hospice. does this seem sus? I have a bad feeling about it

12

4

64



Oedipa Maas @bridgietherease · 7h

...

Replying to @bridgietherease

she hasn't taken any turn for the worse recently and she doesn't have any terminal diagnosis or anything. they just won't let her see anyone unless they're allowed to move her over to hospice. is this just some insane state reg they're trying to work around or what

The problem with the 'with Covid' versus 'for Covid' distinction is that Covid can also screw you up in other ways. So the 'for Covid' population isn't the full story. In relative terms, the statistic still seems super useful.

Reminder, if you don't have any symptoms, there's no need to go to the emergency room or hospital, such people are clogging up the works in Vermont and everywhere else.

This thread is advice on what to do if someone is infected. Most of it seems reasonable (although not complete, she doesn't mention Vitamin D, Zinc or Fluvoxamine). The part where she says to go out and get vaccinated *after you've been exposed* doesn't make *any* sense to me other than as 'one must recommend vaccination at all times no matter what' but one can ignore that and look at the rest.

Pandemic or no pandemic, one man is still trying to find the answers to life's persistent questions.



Jesus Lizard Superstar 🤟 @acekatana · Dec 30, 2021

...

"I had six shots in me. Two were bourbon. Three were Pfizer. One was a .38 slug that the doc never bothered to fish out of my shoulder. And on a night like this I was feeling all of em."



Jesus Lizard Superstar 🤟 @acekatana · Dec 30, 2021

Tonight's a perfect night for doing noir shit in LA, unfortunately there's COVID

112

5.7K

32.1K



Scott Sumner looks back on what it was like two years ago, when the pandemic could potentially have been contained, but we didn't know what we were dealing with. Since China did successfully contain it within China, it's fair to think there's a good chance that the same measures taken earlier would have contained it before it got elsewhere. How realistic that

counterfactual is, and what that would take, is an important question going forwards. The necessary effort would not have been small or cheap. Do we have any good examples of containing something this hard to contain by acting early and aggressively on a (relatively) small scale?

[A fully told story, in one Tweet.](#)



**Walid Gellad, MD MPH**  @walidgellad · 15m

There is a story to be written about academics who were too willing to amplify a message to please those in power, and those who were unwilling to speak up because of fear of career consequences.

...

But I've decided to hold on writing it until my grant is reviewed in study section.

For context, he was referring to this particular lie.



**Walid Gellad, MD MPH**  @walidgellad · 22m

This is true, but what is also true that this message was amplified by doctors and public health professionals who misinterpreted the data on vaccines (and masks).

...

"You can't carry the virus" was the wrong message, not scientifically sound. led to over reliance on vax



**Leyland "Lee" DeVito**  @leedevito · Jan 2

"When the CDC made its controversial and I feel ill-advised decision to change its initial guidance for masking indoors for fully vaccinated people, the rhetoric coming from people like Joe Biden and Rochelle Walensky was very much about this individual responsibility thing."

[Show this thread](#)

## Not Covid

Or is it? [Scott Alexander reviews Don't Look Up \(Spoilers\).](#)

My initial review of this movie was to note the cast looked amazing, and then say, and I quote, "I am not seeing this movie *because I don't have to and you can't make me.*"

I felt like I'd already seen the real world version and the movie would only bring pain, which more information has only confirmed. I do appreciate those involved for trying, and they've provided a baseline from which one can discuss important matters, but in the meantime every time I had the option to see or hear or read about this movie, my life got worse, until Scott's review recapped it sufficiently painlessly that I now feel as if I've seen it on a different level than the one in which I'd 'already seen it' by default.

Scott's review attacks the situation from an interesting angle, which is how to determine who to trust and what it means to Follow the Science without accidentally simply listening to power and thus Following the Science(TM) rather than making your beliefs match physical reality.

That's been a lot of what this two-year project has been about here. At best, the official faces of Science(TM) has been behind, too slow to react to new information, only later doing somewhat reasonable things. At worst, they have outright lied to us about how the world works and given us orders that were Obvious Nonsense.

The only solution is Think For Yourself, Shmuck. There's no easy set of rules to doing that, other than  $P(A|B) = P(B|A) * P(A) / P(B)$ .

Not only is there No Royal Road to Wisdom, but even the poem *No Royal Road to Wisdom*, that often ends the Secular Solstice, is wrong about the road to wisdom. It reads:

The road to wisdom

Is easy to express

You err and err and err some more

but less and less and less

And no, that's a beautiful poem and most people would be improved by taking it more to heart, but that's wrong too, that's the road the schools teach us. That's ultimately the road to *not thinking at all*. That's the road *they want you to take*. You don't want to err *less*. You want to err *on more interesting levels, in more advanced ways*. You want to always be pushing at the level where you err at the right rate, or else you're not learning much. Which is a subtle difference, but supremely important in the end.

One of the biggest problems with the Trusted Authorities is that if it's very important that they never be seen as importantly wrong, it's extremely difficult to ever be importantly right. You can only claim that which is certain, so you can only be unimportantly right. Or you can manipulate the narrative.

[The Litany of Tarski](#) is closer. Thus:

If a comet is coming to destroy the Earth, I desire to believe that a comet is coming to destroy the Earth.

If a comet is not coming to destroy the Earth, I desire to believe that a comet is not coming to destroy the Earth.

Let me not become attached to beliefs I may not want.

Some other time I should notice all the ways in which that last line is *really weird*.

I do my best to synthesize the information I get from a wide variety of sources, combine it into as coherent a physical world model as I can, and synthesize what it all likely means, while doing my best to think probabilistically. Then, you the reader can decide whether my reasoning makes sense, and whether my sources check out, and any other factors to consider. Thus you decide how much you want to rely on me in turn, the same way I am relying on others. It's the best we've got.

I'd like to have something better to put here, but I don't, and I don't believe anyone else does either. [Predictions are hard, especially about the future](#).

[Also, a reminder that people don't love science all that much and this isn't new](#). They didn't want to go to the moon, they were skeptical of the polio vaccine, they didn't care about

computers and all that.

And that if you think we're serious about climate change, well, sometimes it's not you, it's the incentives, like [when they tell you to fly empty airplanes thousands of times across Europe to maintain landing slots](#), and you still don't have a carbon tax.

## Moderation in All Things

Lastly, a note on moderation policy for both Substack and WordPress.

So far, I haven't needed a real one, and have only deleted one post that wasn't either a simple error correction or obvious spam, and the contents of my comments section especially on WordPress have mostly been high quality. Having to find the blog and decide to comment was filter enough.

However, recently I've gotten a lot of new readers, and a number of new frequent commenters, and I'm seeing a higher proportion of comments that make my life worse rather than better, and likely make other lives worse as well. In particular, I've also gotten a higher proportion of comments that are mostly advocacy and aggressive rhetoric rather than gears-level explorations, and often heavily political, which I'd prefer to avoid.

Please strive to post comments that make the world better rather than worse, and avoid comments that make my comments section look like every other content section on the internet.

Also, please be aware that *you cannot edit your comments*, so strive to get it right and to avoid having to put out 3-5 quick reactions in a row to the same thing.

Typical comment sections are full of rudeness both to the author and to others who comment. They also have a lot of ad hominem attacks, repeatedly make the same points without anything new to add, and are neither careful nor precise with claims that anyone who doesn't know something or agree with you something is clueless or stupid or isn't paying attention, or others that are similar. The rightness of one's ingroup against the outgroup is often taken for granted.

That stuff's not what my comments sections are for, and [if the garden is becoming sufficiently ill-kept, I'm not going to let it die by pacifism](#).

The goal is to share models and information (including links) and improve our understanding of the world, and also of course to make life better through other methods such as jokes and social connections and all that.

For now, the new policy is that if a comment makes my life actively worse to a sufficient extent, I'll say that explicitly (as in "This comment made my life worse"), and one can consider this a polite request to update on that information and [stop it](#), but continue to have a very high bar for deletions or bans and hope we can still have nice things. If that doesn't work, I'll escalate.

# 2020 Review Article

A common thing in academia is to write ‘review articles’ that attempt to summarize a whole field quickly, allowing researchers to see what’s out there (while referring them to the actual articles for all of the details). This is my attempt to do something similar for the [2020 Review](#), focusing on posts that had sufficiently many votes (as all nominated posts was a few too many).

I ended up clustering the posts into seven categories: rationality, gears, economics, history, current events, communication, and alignment.

## Rationality

The site doesn't have a tagline anymore, but interest in rationality [remains](#) Less Wrong's defining feature.

There were a handful of posts on rationality 'directly'. Anna Salamon looked at two sorts of puzzles: [reality-masking and reality-revealing](#), or those which are about controlling yourself (and others) or about understanding non-agentic reality. [Listing out examples](#) (both internal and external) helped explain cognitive biases more simply. Kaj Sotala [elaborated on the Felt Sense](#), a core component of Gendlin's Focusing. CFAR [released its participant handbook](#). Jacob Falkovich wrote about [the treacherous path to rationality](#), focusing on various obstacles in the way of developing more rationality.

Personal productivity is a perennial topic on LW. alkjash identified a belief that ‘pain is the unit of effort’, where caring is measured by suffering, and identifies an [alternative, superior view](#). Lynette Bye gave [five specific high-variance tips for productivity](#), and then later argued prioritization is a huge driver of productivity, and explained [five ways to prioritize better](#). AllAmericanBreakfast elaborated on what it means to [give something a Good Try](#). adamShimi wrote about [how habits shape identity](#). Ben Kuhn repeated Byrne Hobart's claim that focus drives productivity, and argued that [attention is your scarcest resource](#), and then talked about [tools for keeping focused](#). alkjash pointed out some ways [success can have downsides](#), and how to mitigate those downsides.

orthonormal discussed the [impact of zero points](#), and thus the importance of choosing yours. Jacob Falkovich argued [against victim mentality](#).

There was some progress on the project of ‘slowly digest some maybe-woo things’. Kaj Sotala gives [a non-mystical explanation of “no-self”](#), detailing some ‘early insights’ into what it means, as part of his sequence on multiagent models of mind. [Ouroboros grapples with Valentine’s Kensho](#). I write a post about how [Circling \(the social practice\) focuses on updating based on experience](#) in a way that makes it deeply empirical.

## Gears

John Wentworth wrote a sequence, [Gears Which Turn the World](#), which had six nominated posts. The first post discussed constraints, and how [technology primarily](#)

[acts by changing the constraints](#) on behavior. Later posts then looked at different types of constraints, and examples where that constraint is the tight constraint / scarce resource: [coordination](#), [interfaces](#), and [transportation](#). He argued that [money cannot substitute for expertise](#) on how to use money, [twice](#).

Other posts contained thoughts on how to develop better models and gearsy intuitions. While our intuitive sense of dimensionality is low-dimensional space, much of our decision-making and planning happens in high-dimensional space, where we benefit from applying [heuristics trained on high-dimensional optimization](#) and geometry. [Ideas from statistical mechanics](#) apply in many situations of uncertainty. Oliver Habryka and Eli Tyre described [how to Fermi Model](#). Maxwell Peterson used [animations to demonstrate how quickly the central limit theorem applies](#) for some distributions. Mark Xu talked about why [the first sample is the most informative](#) when estimating a uncertain quantity.

Scott Alexander wrote [Studies on Slack](#), which goes through examples to see the impacts of amount of slack, and what dynamics lead to more or less of it.

landfish evaluated the evidence and theories and [suggests nuclear war is unlikely to be an x-risk](#).

John Wentworth [summarized Working With Contracts](#). A babble challenge on generating (within an hour) [50 ways to send something to the moon](#) drew 29 responses. dynomight wondered: [what happens if you drink acetone?](#), an example of boggling at simple things posted to the internet, to paraphrase a comment.

## Economics

LessWrong has lots of systems thinkers; economics remains a perennial interest from the Sequences's inception on an economist's blog. [Comparative advantage is not about trade](#), but about production. Talking about comparative advantage also sometimes involves [talking about negotiation](#). Credit-allocation is often imperfect, and so it's useful to think about [incentive design that takes that into account](#). Paul Christiano thought about [moral public goods](#).

Buck talked about [six economics misconceptions](#) of his that he recently resolved. Richard Meadows [defended the Efficient Market Hypothesis](#) in the wake of COVID, and Wei Dai [responded with some specific inefficiencies](#), and asked for help timing the SPAC bubble. aphyer looked into the [limits of PredictIt's ability to track truth](#).

philh looked at all possible two-player simultaneous symmetric games in normal form, and [classifies them](#). Abram Demski argued that most analysis of [game theory misidentifies the relevant games](#).

## History

Things happened in the past; we talk about them sometimes. Often historical examples can help ground our modeling efforts, as when AI Impacts searched for examples of [discontinuous progress in history](#).

Mostly written about by [Jason Crawford](#), Progress Studies seeks to understand what causes progress and thus better understand what interventions would make the world

better. It's grown more nuanced after contact with thinking on x-risk, and seeks a 'new theory of progress' that might care more about things like differential tech development. While he started crossposting on LW in 2019, 2020 saw 8 of his posts in the review, of which the best-liked was [Industrial literacy](#), which argued that understanding the basics of how industrial society works helps reframe the economy as 'solutions to problems', which makes the world much more sensible (and perhaps makes people's desired interventions much more sensible).

Gwern wrote about his personal progress (and the progress he saw in the world) [from 2010-2019](#). [DARPA built a digital tutor](#) that educated much more effectively than traditional classrooms in 2009.

Not only can we model the past, we can look at people in the past modeling the past (and future). Jan Bloch learned from the Franco-Prussian war that [wars were getting much more damaging and less beneficial](#), and that the old style of warfare was on the way out; he tried to stop WWI and failed. Wei Dai shared [his grandparents' story of Communist China](#).

Martin Sustrik wrote about the [Swiss Political System](#). Anna Salamon asks [where stable, cooperative institutions came from](#). Zvi writes about the [dynamics and origins of moral mazes](#). Julia Wise shares [notes on "The Anthropology of Childhood"](#). jefftk writes about [growing independence](#) for his two young children.

## Current Events

Things keep happening in the present; we talk about them sometimes too.

[Anti-Aging is much further along](#) than it looked in 2015; over a hundred companies are deliberately targeting it and plausibly there will be evidence of therapeutic success in 2025-2030. John Wentworth speculated about [aging's impact on the thymus](#) and what could be done about it.

COVID spread to the world. [Practical advice](#) was collected. [Smoke was seen](#). [Points of leverage](#) were discussed. The [CDC was fact-checked](#). [Authorities and Amateurs](#) were compared. John Wentworth asked [how hard it would be to make a COVID vaccine](#). Zvi [analyzed the reaction with the lens of simulacra levels](#). catherio [announced microCOVID.org](#). Zvi [predicted in December](#) that there would be a large wave of infections in March-May, which doesn't come to pass; he detailed in an edit how the data he had at the time led to his prediction.

Biden's prediction market price was too low, [according to deluks917](#). reallyeli [asked if superforecasters are real](#); David Manheim said yes. niplav investigated the [impact of time-until-event on forecasting accuracy](#), finding that long-run questions are easier to predict than short-run questions, but events are easier to predict closer to their time of resolution.

SuspendedReason [interviewed a professional philosopher about LessWrong](#), highlighting adjacent ideas in contemporary philosophy and particularly friendly corners of that space.

Wei Dai asked if [epistemic conditions have always been this bad](#), and responses are mixed (with "no, it's worse now" seeming to have a bit more weight behind it).

Richard Korzekwa talked about [fixing indoor lighting](#).

Katja Grace posted a [photo of an elephant seal](#).

## Communication

Sometimes we talk about talking itself.

Ben Hoffman asked if [crimes can be discussed literally](#), as often straightforward interpretations of behavior rely on 'attack words', and thus it is difficult to have clear conversations. Elizabeth considered [negative feedback](#) thru the lens of simulcra levels.

Malcolm Ocean crossposted his 2015 writing about [Reveal Culture](#), an amendment to the [Tell Culture](#) model from 2014. Ben Kuhn claimed curiosity is a core component of [listening well](#). [Buck thought about criticism](#), noticing that he gets a lot of value from being criticized, and thinking about how to make it happen more. MakoYass outlined a way to build [parallel webs of trust](#).

Zvi described [Motive Ambiguity](#), where one might take destructive actions to reduce ambiguity and so signal one's preferences or trustworthiness. Raemon talked about [practicalities of confidentiality](#), and applying [the fundamental question of rationality](#) to it.

## Alignment

Things will happen in the future; we talk about that sometimes.

The AI Alignment field has grown significantly over the years, and much of the discussion about it happens on the [Alignment Forum](#), which automatically crossposts to LessWrong.

Some work collected, reviewed, and categorized previous work. Rohin Shah [reviewed work done in 2018-2019](#). Evan Hubinger [overviewed 11 proposals for building safe advanced AI](#). Andrew Critch laid out [some AI research areas and their relevance to existential safety](#). Richard Ngo published [AGI safety from first principles](#), which grew from a summary of many people's views to his detailed view.

Other work attempted to define relevant concepts. Evan Hubinger clarified [his definitions of alignment terminology](#). Alex Flint attempted to [ground optimization](#) with a clear definition and many examples. John Wentworth [wrote about abstraction](#). Alex Flint compared [search and design](#).

Paul Christiano wrote precursors to his current research on [Eliciting Latent Knowledge: Inaccessible Information](#), [Learning the prior](#), and [Better priors as a safety problem](#). [Evan Hubinger argued that Zoom In](#) by Chris Olah gives other researchers a foundation to build off of. nostalgebraist [built a lens to interpret GPT](#). Beth Barnes et al summarized [Progress on AI Safety via Debate](#), and then Barnes discussed [obfuscated arguments in more detail](#). Mark Xu suspected that [SGD favors deceptively aligned models](#). Scott Garrabrant introduced [Cartesian Frames](#).

A forecasting thread resulted in a collection of [AI timelines](#). hippke looked at [Measuring hardware overhang](#) thru backdating modern solutions to older hardware. [Ajeya Cotra released her draft report on timelines](#) to get feedback. Daniel Kokotajlo examined [conquistadors as precedents for takeover](#) of human societies (concluding that small edges can be enough to give a significant edge, especially if you can take advantage of pre-existing schisms in your target society), observed that [the visible event when AI takes over is preceded by the point of no return at which their takeover is inevitable](#), and argued [against using GDP as a metric for timelines and takeoff](#). Lanrian [extrapolated GPT-N performance](#). Stuart Armstrong [assessed Kurzweil's predictions about 2019](#) (half of them turned out false).

Steven Byrnes outlined his [computational framework for the brain](#), drawing heavily on human neuroscience, [inner alignment in the brain](#), and the specific example of [inner alignment in salt-starved rats](#), where rats are able to identify the situational usefulness of salt in a way current RL algorithms can't. Alex Zhu [investigated cortical uniformity](#), ultimately thinking it's plausible.

Chi Nguyen attempted to [understand Paul Christiano's Iterated Amplification](#). Rafael Harth [explained inner alignment like I'm 12](#).

John Wentworth argued that [alignability is a bottleneck to generating economic value](#) for things like GPT-3. He also described what it might look like to get [alignment by default](#). In the [Pointers Problem](#), he argued that human values are a function of humans' latent variables.

Jan Kulveit noted that there's a '[box inversion](#)', or duality, between the alignment problems as seen by Agent Foundations and Comprehensive AI Services. JohnWentworth outlined [Demons in Imperfect Search](#), and then DaemonicSigil [built a toy model of it](#). Diffractor wrote up a [sequence on Infra-Bayesianism](#), with the key post on [inframeasure theory](#) also making it into the review. Joar Skalse discussed research on [why neural networks generalize](#), with lots of discussions in the comments. nostalgebraist [thought GPT-3 was disappointing](#) and later explained an [openAI insight about scaling](#) (that data would become the tight constraint instead of compute, moving past GPT-3)

Abram Demski [presented two alternative views of 'utility functions'](#): the 'view from nowhere' defined over the base elements of reductionism, or the 'view from somewhere' defined over perceptual events, and favors the second. He then discussed [Radical Probabilism](#), where Richard Jeffrey expands the possible range of updates beyond strict Bayesian updates. In [The Bayesian Tyrant](#), he gave a simple parable of futarchy further developing this view.

## Conclusion

A lot happened on LW over the course of the year! The main thing that seemed noteworthy, reading thru the review, was just how much alignment stuff there was. [This could be an artifact of more people interested in alignment voting in the review, but I think this matches my memory of the year.]

One of the things that surprised me was how much continuity there was between posts from 2020 and things that people are writing about now; part of this is because of COVID, but I think part of it is a sign of research interests maturing; where rather

than a handful of people chasing fads the community is a considerably larger set of people working on steady accumulation in more narrow subfields.

# Why do we need a NEW philosophy of progress?

This is a linkpost for <https://rootsofprogress.org/why-a-new-philosophy-of-progress>

I've said that [we need a new philosophy of progress for the 21st century](#). But why a new philosophy? Why can't we just return to the 19th-century attitude towards progress, which was pretty enthusiastic?

In short, the view of progress that persisted especially through the late 19th century and up until 1914 was naive. It fell apart because, in the 20th century, it met challenges it could not answer. We need to answer those challenges today.

What follows is a hypothesis that needs a lot more research and substantiation, but I'm putting it forward as my current working model of the relevant intellectual history.

## The 19th-century worldview

Here are a few key elements of the Enlightenment-era worldview:

- Nature was an obstacle to be conquered. Nature was *imperfect*; human reason could improve it—and it was fitting and proper for us to do so. [Kipling wrote](#), “We hold all Earth to plunder / All time and space as well.” Nature was a means to our ends.
- There was a deep belief in the power of human reason both to understand and to command nature. Especially by the end of the century, the accomplishments in science, technology and industry seemed to confirm this.
- As a corollary of the above, there was an admiration for growth and progress: in science, in the economy, [even in population](#).

(I'm basing this mostly on writings from the time, such as [Macaulay](#) or [Alfred Russel Wallace](#); contemporary [newspaper editorials](#); popular speeches given, e.g., at [celebrations](#); [poetry](#) of the era; etc. For future research: what were the historians, philosophers, etc. of the time saying about progress? I'm familiar with some of the thought from previous centuries such as Bacon and Condorcet, but less so with that from 19th-century figures such as Mill or Comte.)

On the face of it, at least, these seem very much in sympathy with the [core ideas of the progress movement](#) as I have outlined them. So what did the 19th century get wrong?

## Mistakes

Here are just some examples of things that many people believed in the late 19th century, which would later be proved quite wrong:

- **That technology would lead to world peace.** Enlightenment thinkers such as Condorcet had [forecast](#) progress in morality and society just as much as in science, technology and industry. By the late 1800s, this seemed plausible. The

previous century had seen monarchy and aristocracy replaced by democratic republics, and the end of slavery in the West. Economic growth was making everyone richer, and free trade was binding nations together, giving them opportunities for win-win collaboration rather than destructive, zero-sum competition. The telegraph in particular was hailed as an invention that would unite humanity by allowing us to better communicate. Everything seemed to be going relatively well, especially after 1871 (end of the Franco-Prussian War), for over 40 years...

- **That “improvements on nature” would avoid unintended consequences.** (This one may have been implicit.) It’s good to try to improve on nature; it’s bad to go about it blithely and heedless of risk. One striking example is [the popularity of “acclimatization societies”](#), “based upon the concept that native fauna and flora were *inherently deficient* and that nature could be greatly improved upon by the addition of more species.... the American Acclimatization Society was founded in New York City in 1871, dedicated to introducing European flora and fauna into North America for both economic *and aesthetic* purposes. Much of the effort made by the society focused on birds, and in the late 1870’s, New York pharmacist Eugene Schieffelin led the society in a program to introduce every bird species mentioned in the works of Shakespeare.” (Emphasis added.) These importations led to invasive pests that threatened crops, and were ultimately placed under strict controls.
- **That progress was inevitable.** The most optimistic thinkers believed not only that continued progress was *possible*, but that it was being driven by some grand historical force. Historian Carl Becker, [writing about this period](#) soon after it had passed, spoke of the conviction that “the Idea or the Dialectic or Natural Law, functioning through the conscious purposes or the unconscious activities of men, could be counted on to safeguard mankind against future hazards,” adding that “the doctrine was in essence an emotional conviction, a species of religion.”

## 20th-century challenges to the idea of progress

The idea of progress was never without detractors. As early as 1750, [Rousseau declared](#) that “the progress of the sciences and the arts has added nothing to our true happiness,” adding that “our souls have become corrupted to the extent that our sciences and our arts have advanced towards perfection” and that “luxury, dissolution, and slavery have in every age been the punishment for the arrogant efforts we have made in order to emerge from the happy ignorance where Eternal Wisdom had placed us.” But through the 19th century, voices like this could barely be heard above the cheering of the crowds in celebration of the railroad, the light bulb, or the airplane.

What changed in the 20th century? Here are several factors:

**The world wars.** With World War I, it became clear that technology had not led to an end to war; it had made war all the more horrible and destructive. Progress was not inevitable, certainly not moral and social progress. By the end of World War 2, the atomic bomb in particular made it clear that science, technology and industry had unleashed a new and very deadly threat on the world.

The wars, I think, were the main catalyst for the change. But they were not the only challenge to the idea of progress. There were other concerns that had existed at least

since the 19th century:

**Poverty and inequality.** Many people were still living in dilapidated conditions, without even toilets or clean water, at the same time as others were getting rich from new industrial ventures.

**Job loss and economic upheaval.** As technology wrought its “creative destruction” in a capitalist economy, entire professions from blacksmiths to longshoremen became obsolete. As early as the 1700s, groups led by “Ned Ludd” and “Captain Swing” smashed and burned textile machinery in protest.

**Harms, risks, and accountability in a complex economy.** As the economy grew more complex and people were living more interconnected lives, increasingly in dense urban spaces, they had the ability to affect each other—and harm each other—in many more ways, many of which were subtle and hard to detect. To take one example, households that once were largely self-sufficient farms began buying more and more of their food as commercial products, from increasingly farther distances via rail. Meat packing plants were filthy; milk was transported warm in open containers; many foods became contaminated. In the US, these concerns led in 1906 to the Pure Food & Drug Act and ultimately to the creation of the FDA.

**Concentration of wealth and power.** The new industrial economy was creating a new elite: Rockefeller, Morgan, Carnegie. Their wealth came from business, not inheritance, and their power was more economic than political, but to many people they looked like a new aristocracy, little different than the old. In America especially, the people—who just a few generations ago had fought a war to throw off monarchical rule—were suspicious of this new elite, even as they celebrated rags-to-riches stories and praised the “self-made man.” It was a deep conflict that persists to this day.

**Resource consumption.** Long before Peak Oil, William Stanley Jevons was [warning of Peak Coal](#). Others predicted [the end of silver](#) or other precious metals. Sir William Crookes (more accurately) sounded the alarm that the world was [running out of fertilizer](#). Even as people celebrated growth, they worried that the bounty of nature would not last forever.

**Pollution.** Coal use was blackening not only the skies but the houses, streets, and lungs of cities such as London or Pittsburgh, both of which were likened to hell on Earth because of the clouds of smoke. Raw sewage dumped into the Thames in London led to the [Great Stink](#) and to cholera epidemics. [Pesticides](#) based on toxic substances such as arsenic, dumped in copious quantities over crops, sickened people and animals and poisoned the soil.

And there was at least one major new concern coming to the fore:

**The environment,** as such. The 19th century may have worried about pollution and resources, but in the 20th century these concerns were united into a larger concept of “the environment” considered as a systematic whole, which led to new fears of large-scale, long-term unintended consequences of industrial activity.

## New explanations

Historical events can be a catalyst for change, but they do not explain themselves. It is up to historians, philosophers, and other commentators to offer explanations and

solutions. Thus history is shaped by events, but not determined by them: it is partly determined by how we choose to interpret and respond to those events.

Those who stepped forward in the 20th century to explain what went wrong—especially (although not exclusively) environmentalists such as William Vogt or Paul Ehrlich—emphasized the concerns above, and added a layer of deeper criticism:

- That we were becoming “disconnected” from nature and/or from our families, communities, and traditions
- That progress was not making us happier or healthier; that people had been and were better off in less industrialized societies (even, some claimed, as tribal hunter-gatherers)
- That there were inherent [limits to growth](#), which we were exceeding at our peril

Underlying this analysis were some basic philosophical premises:

- Human well-being was not consistently their standard of value. Some saw inherent value in nature, above and apart from its usefulness to humans; some even turned anti-human (such as David Graber, who [wrote](#): “We have become a plague upon ourselves and upon the Earth... Until such time as *Homo sapiens* should decide to rejoin nature, some of us can only hope for the right virus to come along.”)
- They lacked the 19th-century belief in the efficacy of reason, and therefore in the ability of humanity to control our destiny. The world was too big and complicated for us to understand, and we were ultimately at the mercy of forces beyond our control, especially if we decided to tinker with complex systems.
- As a corollary of the above, they adopted “sustainability” as an ideal, rather than growth, which was seen as an unhealthy “addiction.”

(If the above seems singularly focused on environmentalism, it reflects the incomplete state of my research. As I’ve noted elsewhere, progress is criticized from the right as well as from the left, for its “materialism” and “decadence.” Open questions for me here include the role of religion in this period, and the reaction of the liberal world to the rise of socialism and fascism.)

This new worldview did not take over immediately; it slowly grew in influence during the generation after the World Wars. But by the time the world was cheering the Moon landing and greeting the astronauts on a triumphant world tour, this philosophy had spawned the New Left and the radical environmentalist movement. The oil shocks hit a few years later; as Americans lined up for gas rations and donned sweaters, many people thought that perhaps the “limits to growth” were real after all.

## Regrouping in the 21st century

The 21st-century progress movement must directly address the challenges that created skepticism and distrust of progress in the 20th century. Those challenges have not gone away; many have intensified: in addition to nuclear war, pollution, and overpopulation, we are now worried about climate change, pandemics, and threats to democracy.

Here are some difficult questions the new progress movement needs to answer:

- Is material progress actually good for humanity? Does it promote human well-being? Or is it an unhealthy “addiction?”
- Is progress “unsustainable?” How do we make it “sustainable?” And what exactly do we want to sustain?
- Does progress benefit everyone? Does it do so in a fair and just way?
- How can we have both progress and safety? How do we avoid destroying ourselves?
- What are the appropriate legal frameworks for existing technologies and for emerging ones?
- How do we address environmental issues such as climate change and pollution?
- How do we deal with the fact that technology makes war more destructive?
- How can we make sure technology is used for good? How do we avoid enabling oppression and authoritarianism?
- How can we make moral and social progress at least as fast as we make scientific, technological and industrial progress? How do we prevent our capabilities from outrunning our wisdom?

Without answers to these questions, any new philosophy of progress will fail—and probably deserves to.

*I don't have all the answers yet—and I'm not sure that anyone does.* I think we need new answers.

---

This is why we can't simply return to the 19th-century philosophy of progress. First, it was mistaken. Second, there is a reason it failed: it foundered on the shoals of the 20th century. If it were revived, it would immediately run into the same problems, the same challenges it could not answer. In any case, there would be something odd and deeply incongruous about a movement dedicated to building an ambitious technological future that was stuck in a philosophic past.

Instead, we have to find a new way forward. We have to acknowledge the problems and concerns of the modern world, and we have to find solutions. Not the regressive proposals offered in the 20th century, but ones based on [a humanistic standard of value, a belief in human agency, and an understanding of the reality and desirability of progress.](#)

---

*Thanks to Tyler Cowen, Greg Salmieri, Clara Collier, and Michael Goff for comments on a draft of this essay.*

# Calibration proverbs

Cross-post from [Telescopic Turnip](#).

Every problem is a calibration problem. That's why most advice is basically useless: it tells you what to do, but doesn't tell you when to stop. Therefore, the best pieces of advice are calibration advice. What we need is a metric to know if we are doing well, or if we should change our habits (and in which direction). For example, here are two poor pieces of advice:

"You should spend more time reading blogs, because compared to traditional media, bloggers have more freedom to communicate in original ways, and are more accountable when they say something false."

"You should spend less time reading blogs, because keeping up with a big pile of subscriptions takes a lot of time and makes you anxious."

Neither of these is really useful. Instead, we can produce one piece of calibration advice:

"After catching up with your backlog of blog posts, ask yourself what you remember out of it. If you can't say what half of the posts were about, you should probably clean up your subscriptions."

Of course, that doesn't sound like the timeless advice my grandmother received from her grandmother. To make it compelling, we need to [make it rhyme](#):

"Before you subscribe to another blog,  
check what you took home from your backlog."

Now *that* sounds like ancient wisdom.

Thereafter is my humble attempt at writing calibration proverbs. As a non-native, English's pronunciation is still a mystery to me, so the rhymes might be a bit wonky. Anyways, may these proverbs contribute to making 2022 a better year than 2021.

"He who thinks a claim by evidence is backed,  
can he predict the effect size just from the abstract?"

"Only when you leave a long conversation,  
you can tell if it warranted [your apprehension](#)."

"If you don't enjoy the taste of tea,  
seek how to [infuse it properly](#)."

"Always check for bikes when you get out of your car.  
Always."

"When checking the source teaches you something new,  
the source of the source you should check too."

"One hour with your kids is worth two hours online,  
according to [Han et al twenty twenty one](#)."

"Man whose roommates wear t-shirts in winter,  
would better calm down with the goddamn heater."

What other calibration proverbs should we transmit to our grandchildren?

---

*Thanks to Justis for the proofreading and rhyme consulting.*

# Why rationalists should care (more) about free software

*cross-posted to [pan narrans](#)*

# Why rationalists should care (more) about free software

*especially if you want to upload your brain*

In the limit condition freedom of compute is freedom of thought.

As we offload more of our cognition to our computational devices we expose a new threat surface for attacks on our ability to think free of malign or otherwise misaligned influence. The parties who control the computational systems to which you have outsourced your cognition have a vector by which to influence your thinking. This may be a problem for you if their interests are not aligned with your own as they can use this power to manipulate you in service of their goals and against your own.

The fundamental operations of our brains remain difficult to reliably and effectively interfere with primarily because of our ignorance of how to achieve this. This, however, may change as understanding of our wetware increases and subtle direct **manipulations of our brain** chemistry can be employed to influence our behaviour. A highly granular version of this approach is likely still quite far off but it **generally feels more viscerally scary than influencing us via our technology**. Surfing the web without ad-block already feels uncomfortably close to the [futurama gag about ads in your dreams](#). Increasing though this is amounting to the same thing. Indeed our technology is already doing this to us, albeit fairly crudely for now, by exploiting our reward circuits and many other subtle systematic flaws in the human psyche.

**What is "free" software?** Free as in liberty no as in gratuity, as in speech not beer, politically and not necessarily financially. The free software foundation defines free software as adhering to the [four essential freedoms](#) which I paraphrase here:

0. The freedom to run the code however you wish
1. The freedom to examine its source code so that you can understand and modify it for your own purposes
2. The freedom to distribute the source code as is
3. The freedom to distribute modified versions of the source code

Note that code which is 'source available' only really gets you freedom 1, depending on how the code is licenced and built this may not get you any of the others including freedom 0. Much ink has been spilt over the use of the term 'open source' as not going far enough as a result. Free software is often referred to by the acronyms FOSS & FLOSS (Free/Libre and open source software)

The occasionally controversial but ever prescient [Richard Stallman](#) (AKA RMS, AKA [saint IGNUcius](#)) has been banging on about the problems of proprietary software for

nearly forty years at this point. Having essentially predicted the abuses of today's software giants because he got a bad printer diver in the early 1980s.

The problem that Stallman saw with 'proprietary' software, i.e. software which does not meet the criteria of the four essential freedoms, is one of game theoretic incentives. **Making software free serves as a pre-commitment mechanism by the software authors to not abuse the users of their software.** This works by **empowering users to exercise a credible threat of forking the project** and cutting devs abusing their position out of the project and any associated revenue streams. Revenue from free software projects can take a number of forms e.g. premium-hosting, donations/pay-what-it's-worth schemes, & service/support agreements, though how to successfully monetise free software remains a hard problem.

**As the maker of a piece of propriety software, you are not subject to this kind of check on your power** and it is often in your interest to increase lock-in to your product from your users to make it hard for them to leave for a competitor, should they become dissatisfied. The lack of transparency on how proprietary software works can also hide a multitude of sins such as bad security practices and provides scope for extensive surveillance of the users whilst maintaining deniability. Thus **free software can serve as a solution to an alignment problem between makers and users of the software.**

The speculative fiction of [Cory Doctorow](#) and [Greg Egan](#) in '[permutation city](#)', along with the speculative (non-fiction?) of [Robin Hanson](#) in '[Age of em](#)' has painted pictures of numerous diverse dystopian futures in which software is used to curtail individual liberties, as well as to gas-light, [frame control](#), and otherwise manipulate or abuse people and other conscious entities.

Concerns over these potential abuses have been gaining increasing popular attention in recent years though the emphasis has been placed on [Shoshana Zuboff](#)'s concept of [surveillance capitalism](#) rather than framing the problem, as I suspect Stallman would, as having its root causes in non-free software. In particular, the popularity of the Netflix documentary '[The Social Dilema](#)' made in collaboration with [Tristan Harris](#) & [Aza Raskin](#)'s [Centre for human technology](#) has increased public awareness of the problems, solutions, however, remain relatively unspecified.

Computing is becoming ever more ubiquitous, connected and is [beginning to be embedded in our bodies](#), though mostly still as medical devices for now. Whose phone numbers do you know, what about addresses or how to travel there? How's your mental arithmetic? **how good is your recall of your chat history with all your friends - would you notice it if was subtly edited in retrospect?** Do you have a voice assistant? When was the last time you left your house without your phone? **The more of our cognition takes place external to our brains the more vulnerable we are to the technological capture of our thought processes by misaligned entities. If we do not take measures to ensure the alignment of software makers interests with those of software users we invite dystopias galore.**

Over the years there have been many explicit efforts by technology companies to lock general-purpose computing devices to vendor-approved applications (e.g. many game consoles & iPhones). This is often in the name of copyright protection and increasingly in recent years in the name of providing better security. **'Better security' of course begs the question, against what threat model?** It's better security against malicious 3rd parties but what if I'm worried about what the 1st parties are doing? It

comes down to the question of **who holds the keys to the locks**. I know I'd want to be the one deciding who's signing keys I trust to go in the future-[TPM](#)-analog of the computer system emulating my brain and given their track records it's probably not Google, Apple, Amazon, Facebook *I'm sorry Meta - rolls eyes*, or Microsoft. (The basic competencies, understanding, and good widely adopted low friction systems needed for individuals to be good stewards of their own private keys is a problem in the locked bootloader space as well as the cryptocurrency space.) **It is worth noting that at this point in time it is almost impossible and extremely impractical to get a completely free software computer [down to the firmware level](#).**

**I think a strong case could be made that a 'freedom of compute' should be enshrined in future constitutional settlements on par with freedom of speech as a protection of fundamental freedoms, in service to preserving freedom of thought.** FOSS development has been [discussed in the EA community](#) as a potentially valuable intervention. Developers seem to be overrepresented in the rationalist community so maybe this is a bit of a touchy subject for any of us working on proprietary code. I'm of the opinion that we as a community should advocate for free software and that there is a certain synergy between the free software movement's goals and those of the rationality community, I'd be interested to hear contrary opinions.

**Well-aligned software has the potential to massively improve our lives both at the individual and societal levels,** [look at what Taiwan is doing with open software in digital governance](#). Making use of some of the same behavioural modification tricks currently used to sell us crap we don't need and immiserate us as a side effect so that we can be sold the cure can be turned to good. Helping us to establish good habits, to break bad ones and beat [akrasia](#). To collaborate and communicate more deeply and effectively, instead of more shallowly and ineffectually. To be understood not misunderstood, seen for who we are and not through the funhouse mirror of beautification filters. To build a [fun](#) world together, not a depressing and solipsistic one.

**Disclosure: I am an associate member of the FSF, and pay them an annual membership fee & the link on '[beginning to be embedded in our bodies](#)' is a shamelessly self-promotional link to an episode of my podcast where my co-host and I discuss embedded tech and its implications at length**

# Forecasting Newsletter: Looking back at 2021

This is a linkpost for <https://forecasting.substack.com/p/looking-back-at-2021>

## Table of contents

- The American Empire Has Alzheimer's.
- Prediction Markets: VC money, searching for DraftKings, predatory pricing, and the race to be last.
- For skilled forecasters, crypto prediction markets are much more profitable than forecasting platforms.
- Best forecasting pieces from 2021

You can sign up for or view this newsletter on [substack](#), where there are already a few thoughtful comments.

## The American Empire has Alzheimer's

It is 1964. Sherman Kent is a senior intelligence analyst. While doing some rudimentary experiments, he realizes that [analysts themselves disagree about the degree of confidence that words convey](#). He suggests that analysts clearly state how certain they are of their conclusions, by using words that correspond to probabilities. This will allow keeping track of how analysts do, while being simple enough for less detail-focused politicians to understand. His proposal encounters deep resistance, and doesn't get implemented.

It is the 30th of April of 1975. With the [fall of Saigon](#), the US finally pulls out of a bloody war with Vietnam. There are embarrassing images of people flying out of the US embassy at the last moment. Biden is a newly-minted [senator from Delaware](#).



It is 2001. The US intelligence agencies are very embarrassed by not having been able to predict the September 11 attacks. The position of the [Director of National Intelligence](#), and an associated [Office of the Director of National Intelligence](#), is established to coordinate all intelligence agencies to do better in the future.

At the same time, [Robin Hanson](#) pushes for a "[Policy Analysis Market](#)", which would have covered topics of geopolitical interest. This proposal becomes too controversial, and gets dropped.

It is 2008. A bunch of Nobel Prize winners and other luminaries publish [a letter](#) urging the Commodity Futures Trading Commission (CFTC) to make prediction markets more legal.

It is 2010. IARPA, an intelligence agency modeled after DARPA, which incubates high-risk, high payoff projects, creates [a tournament](#) to find out which forecasting setups do best. Philip Tetlock had done some experiments which found that pre-selecting participants does pretty well. He repeatedly wins the IARPA tournament, and creates Good Judgment Inc to provide the services of his preselected high-performance forecasters. The US doesn't buy their services, and Good Judgment Inc survives by selling very expensive training sessions to clients which have too much money.

It is 2013. The CFTC shuts down [Intrade](#), one of the only prediction market platforms in the US.

It is 2017 and onwards. As Ethereum and crypto more generally become more mainstream, some prediction markets on top of crypto-blockchains start to pop up, such as Augur, Omen, and later, Polymarket. As cryptocurrencies become more and more popular, the fees on the original Ethereum blockchain increase so much that placing bets on these prediction markets becomes too expensive. Polymarket survives by moving to a "layer two" blockchain, a less paranoidly secure blockchain that mimics the Ethereum blockchain, and which allows users to continue betting.

It is the summer of 2021. Biden makes incredibly overconfident assertions about the Afghani government holding on against the Taliban. It doesn't. There are images of the evacuation from Kabul, Afghanistan which look very similar to the evacuation from Saigon, Vietnam. This is all very embarrassing to the Biden administration, and his approval rating drops drastically.

"There's going to be no circumstance where you see people being lifted off the roof of an embassy in the — of the United States from Afghanistan. [...] the likelihood there's going to be the Taliban overrunning everything and owning the whole country is highly unlikely." — Biden, [July 08, 2021](#)



Come Christmas of 2021, the CFTC gives Americans the gift of disappointment by shutting down Polymarket in the US, one of the few places where real money was being traded around topics of extreme interest to Americans, like US covid cases.

The picture I paint above is somewhat reductionist, and omits some important details. For instance, for a while the US had, and maybe still had an intelligence community prediction market. More recently, the United Kingdom also has a "Cosmic Bazaar", a forecasting tournament using [Cultivate Labs infrastructure](#). Dominic Cummings, who was Chief Adviser to Prime Minister Boris [talked of reading Scott Alexander's blogposts about Covid](#). There is also forecasting done with the Czech Republic, the Dutch and the OSCE. But I am yet to think that forecasters are meaningfully driving policy. The forecasting community is close-knit and I think there would be conversations if policymakers were regularly looking at forecasts—if you disagree, please get in touch.

Still, even after adjusting for my predisposition for pessimism, I think that the broad strokes of the above overview are about right. The US government is not being a "strong optimizer", whatever that means. In fact, the US government is being fucking dumb. But it took me a

while to crystallize this, and to notice how some of the dysfunctional aspects of the forecasting panorama have the same root cause. PredictIt's fees are so high (10%) because until very recently, they didn't have competition to keep them on their toes. Metaculus is—to some extent—structured around fake internet points because doing the real money version would have been bureaucratically exhausting.

## Prediction Markets: VC money, searching for DraftKings, predatory pricing, and the race to be last

### VC money

In the past few years, a few startups have joined the prediction market arena, chiefly Polymarket, Hedgehog Markets and Kalshi. Augur, previously a crypto project with very strong decentralization mechanisms, also sold out and spun off a more commercially oriented site, Augur Turbo, more focused on sports, crypto and entertainment. There were also a whole lot of less successful copycats, like [PolkaMarkets](#).

These projects have gotten a fair amount of funding. Polymarket got an initial \$4M investment round, and was [reportedly valued at \\$1B in later talks](#). Kalshi got \$30M in funding from, among others, [Sequoia Capital](#). Augur Turbo got an investment of [\\$1M from Polygon for its liquidity program](#), the network on which it and Polymarket runs. And Hedgehog Markets got a [\\$3.5M investment](#), as well as [\\$500k through the sale of NFTs](#), NFTs which allow users to participate in exclusive walled-off markets.

### Searching for DraftKings

What is driving that valuation and initial investment? Well, for comparison, DraftKings, one of the biggest sports markets around, was valued at [\\$20B](#) before its stock price [took a beating after an adversarial report by Hindenburg's Research, a short-seller](#). So being a similarly large player in the nascent prediction markets field could be worth a significant fraction of that. Even being the "DraftKings of crypto", i.e.—the largest and more liquid player for sports within the crypto ecosystem—could potentially be worth quite a bit.

### Predatory pricing and the race to be last

But first, these startups have to capture the market. And the way they are trying to do this is by subsidizing participation. That is, they create markets whose initial probabilities are off, giving users the chance to make money by participating in the market. I'm most familiar with how Polymarket has done this; I think they have overall lost money even though they have seen tens of millions in volume. My guess is that some other platforms have likewise lost a fair bit of money subsidizing volume.

But this creates a race to be the last to subsidize one's own markets, and steal the competition's user base. Then, perhaps, the last one standing could [monopolize the business and raise fees](#).

### Alternative profit models.

In short, it looks that right now, there is money flushing around, but eventually the necessary sucker at the table will tend to be the users. And this incentivizes markets on sports, NFTs, or celebrities, rather than on war, politics, or technological developments, because they have more mainstream appeal. So the core, amoral business insight here is that all of these

platforms are vying for the same slice of the market: the sports and crypto markets. Or, in other words, entertainment for those newly rich off the crypto-boom.

What would an alternative business model be? Well, on the one hand, the different prediction markets could aim for different niches. Hedgehog markets could aim to entertain people heavily into the cryptocurrency scene. Polymarket could aim to be the best at real-world predictions. Augur could return to its original vision and be the go-to place for paranoid users interested in security. And FTX offers the best derivatives on cryptocurrency products.

Personally, the profit model that I'd like to see is one in which the prediction market platforms extract the profit not from their users, but rather from the people who are consuming the odds which the betting produces as a side-effect. For instance, a large NGO such as Open Philanthropy might be interested in a variety of geopolitical events, and could subsidize a market on them. This would involve providing both the liquidity (\$300 to a few thousand per market), and some money to support the prediction market platform. As the decentralized finance ecosystem develops, instead of a central organization paying for public goods, [DAOs](#) might form for this purpose.

Such a profit model would be such that the prediction market platform would be able to benefit in proportion to how much value it generates in the world. For instance, right now Polymarket creates value by producing common knowledge about sensible default probabilities to have around covid. Having sponsors which pay in proportion to how much value these markets produce, and which are willing to pay to create valuable markets might allow platforms such as Polymarket to capture a fraction of the value they create.

That kind of a profit model could, I think, make humanity more formidable.

## **For skilled forecasters, crypto prediction markets are much more profitable than forecasting platforms.**

A drift divides the human forecasting space. On the first corner, we have forecasting platforms, which are legal throughout the land, and which see lower volumes. Forecasters play by giving their probabilities, and checking whether these are more right than other participants'. They are often rewarded according to a proper scoring rule so that they're incentivized to give their true and honest best guess, but this often fails in [amusing ways](#). Chief amongst these platforms are Metaculus and Good Judgment Open, which have many questions and whose communities have historically been open and welcoming. Forecasting platforms tend to have socially useful questions on e.g., geopolitics, technology developments, or risks to society.

On the opposing corner, we have prediction markets, where participants put their money where their mouth is, and earn money if they turn out to be right. The communities can also be welcoming in their own ways, though this is partially because good bettors are looking for less experienced bettors to fleece. Prediction markets are of dubious legality, mostly because some apparatchiks under the Commodity and Futures Trading Commission (CFTC) have a hard-on against it. In principle, real-money prediction markets could have questions on any topic, but they tend to have questions that have more mainstream appeal, such as on sports.

In recent times, it has become noticeably the case that prediction markets, and in particular crypto prediction markets, are significantly more profitable to participants than forecasting platforms. So some forecasters who trained themselves on Metaculus then flocked to try their luck on Polymarket, and the best ones made significantly more money than they could have made on Metaculus.

For reference, the current monetary rewards given in forecasting platforms are roughly as follows:

- Metaculus: On the order of \$1000 per tournament.
- Hypermind: On the order of \$5000 per tournament.
- Replication Markets: Around \$150k in total, for around 10-20 rounds of forecasting, each of which contained many questions.
- CSET-Foretell; \$120k a year, or \$200 per forecaster per month.

Note that these are per tournament, so if a \$1000 Metaculus tournament contains around ten questions, and ten forecasters participate on it, this amounts to \$10 per forecaster. If a forecaster spends more than an hour per question, they are then earning less than minimum wage.

In contrast, prediction markets, such as PredictIt or Polymarket often see upwards of \$100k traded on individual markets. So the top predictors can and do make a comfortable living betting, in a way that would be difficult or impossible to do in forecasting platforms rather than in prediction markets.

For instance, one of the top earners on [Replication Markets](#) earned around \$10k. But he spent significant time programming tools to automate his trading, and it seems like he put a lot of love into his forecasting. If he had invested that labor into trading on Polymarket and building tools to do so, or if he had simply sold his labor as a programmer, he would have earned significantly more money.

So the incentives are not pointing in the right direction. Capable forecasters can earn significantly more by predicting societally-useless sports stuff, or simply by arbitraging between the big European sports-houses and crypto markets. Meanwhile, the people who remain forecasting socially useful stuff on Metaculus, like whether [Russia will invade the Ukraine](#) or whether there will be any [new nuclear explosions in wartime](#), do so to a large extent out of the goodness of their heart.

I think that the clear solution to this is to either increase the overall willingness to pay forecasters, or to be willing to subsidize liquidity in prediction markets for questions that are of general value.

## Best pieces on forecasting during 2021

### Practice

[Predicting Politics](#) is generally worth reading, starting with [How to get good](#), [Mining the Silver Lining of the Trump Presidency](#), and [Boring is back, baby](#).

Avraham Eisenberg wrote [Tales from Prediction Markets](#), gathering a few interesting anecdotes.

[Cultured meat predictions were overly optimistic](#) (a). "Overall, the state of these predictions suggest very systematic overconfidence."

Charles Dillon of Rethink Priorities and SimonM looked at [How does forecast quantity impact forecast quality on Metaculus?](#) More forecasters increase forecast quality, but the effect is small beyond 10 or so forecasters.

David Friedman looked at whether the [past IPCC temperature projections/predictions have been accurate?](#)

[Violating the EMH — Prediction Markets](#) gave specific examples in which prediction markets appeared to violate the efficient market hypothesis.

[How I Made \\$10k Predicting Which Studies Will Replicate](#). The author started out with a simple quantitative model based on [Altmejd et al. \(2019\)](#), and went on from there.

Together with my coauthors Misha Yagudin and Eli Lifland, I posted a fairly thorough investigation into [Prediction Markets in The Corporate Setting](#). The academic consensus seems to overstate their benefits and promisingness. Lack of good tech, the difficulty of writing good and informative questions, and social disruptiveness are likely to be among the reasons contributing to their failure. In the end, our report recommended not having company-internal prediction markets.

Charles Dillon wrote [Data on forecasting accuracy across different time horizons and levels of forecaster experience](#), using Metaculus and PredictionBook data, and building on earlier work by [niplav](#).

## Futurism

[Incentivizing forecasting via social media](#) explored the implications of integrating forecasting functionality with social media platforms.

The Machine Intelligence Research Institute's research on agent foundations shed some light on [probability theory more generally](#), [Radical Probabilism](#) and [Reflective Bayesianism](#) seem particularly worth highlighting, as does [Probability theory and logical induction as lenses](#).

Daniel Kokotajlo wrote [What 2026 looks like \(Daniel's Median Future\)](#), extrapolating the performance of models like GPT-3 year by year. Ben Snodin wrote [My attempt to think about AI timelines](#).

The Machine Intelligence Research Institute has published a few [conversations on future AI capabilities](#). Of these, readers of this newsletter might be particularly interested in the [Conversation on technology forecasting and gradualism](#).

## Theory

There was some back and forth online on Kelly betting:

- [Kelly isn't just about logarithmic utility](#)
- [Kelly is just about logarithmic utility](#)
- [Never Go Full Kelly](#)
- [Why the Kelly criterion kind of sucks](#)

See also: [Proebsting's paradox \(a\)](#), a thought experiment in which naïve Kelly bettors are lead to ruin, [Learning Performance of Prediction Markets with Kelly Betting](#) proves that prediction markets with Kelly bettors update similarly to Bayes' law, and [this blog post](#) illustrates that paper's point in a more approachable manner.

After reading these posts, I'm left with the conclusion that Kelly betting is an interesting yet ultimately limited tool. One encounters the limits of applicability as soon as one is exposed to many bets at once, or to the chance that the bets may change in favorable or unfavorable directions.

Alex Lawsen and I published [Alignment Problems With Current Forecasting Platforms](#), outlining problems with the incentive mechanisms in almost all non-prediction market platforms.

[The Generalized Product Rule](#) outlines how a certain step in [Cox's theorem](#)—the step which proves that probability updating is multiplicative—can be applied to other problems as well.

Jaime Sevilla took a deep dive into [aggregating forecasts](#).

Eli Lifland published an article [on bottlenecks to more impactful forecasting](#). It crystallizes his knowledge from a few years of his forecasting on Metaculus, CSET-Foretell and Good Judgment Open.

# ELK First Round Contest Winners

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

Thank you to all those who have submitted proposals to the ELK proposal competition. We have evaluated all proposals submitted before January 14th<sup>[1]</sup>. Decisions are still being made on proposals submitted after January 14th.

The deadline for submissions is February 15th, after which we will release summaries of the proposals and associated counterexamples.

We evaluated 30 distinct proposals from 25 people. We awarded a total of \$70,000 for proposals from the following 8 people:

- \$20,000 for a proposal from Sam Marks
- \$10,000 for a proposal from Dmitrii Krasheninnikov
- \$5,000 for a proposal from Maria Shakhova
- \$10,000 for proposals from P, who asked to remain anonymous
- \$10,000 for a proposal from Scott Viteri
- \$5,000 for a proposal from Jacob Hilton and Holly Mandel
- \$10,000 for a proposal from Jacob Hilton

We would have also awarded \$15,000 for proposals from Holden Karnofsky, however he is on ARC's board and is ineligible to receive the prize.

- 
1. We don't plan on spending time evaluating and awarding proposals made after February 15th but feel free to email [elk@alignmentresearchcenter.org](mailto:elk@alignmentresearchcenter.org) in the future if you think you have a promising proposal, and we might consider it for a prize. ↵

# Omicron Post #13: Outlook

Omicron is now the dominant strain, so much of what would otherwise have gone into this Omicron post ended up [in the latest weekly post](#) instead, and I'm considering the right way to merge the two streams together.

In addition to some amount of the usual business, the focus of this post will be an explicit attempt to look at what's likely to happen from here, and give my baseline scenario.

I recently had reason to chart out what I expected that baseline scenario to be around the world, so it makes sense to share it here.

The world can be divided into five (highly unequally sized) parts, with increasing levels of uncertainty, some places are unclear on whether they're in groups two, three or four:

1. South Africa and other already-peaked areas.
2. Uncontrolled spread: Europe, North America, India, Israel, Australia and Argentina.
3. Without an Omicron wave, or without good data: Third world, Russia.
4. Previous containment: South Korea, Japan, Taiwan and New Zealand.
5. Special case of previous containment: China.

## **South Africa and Other Already-Peaked Areas**

This is the easy story to tell. Predictions about the past are often still hard, but they're a lot easier.

A few African countries have already finished their Omicron waves, and got through mostly unscathed. Since they're in the southern hemisphere, one can worry that there's a wave coming in their winter a few months from now, but I would expect any such wave to be mild even if relatively little is done to prepare, given how things went this time, and that those infections are now baked in.

## **Uncontrolled spread: Europe, North America, India, Israel, Australia and Argentina.**

This is where most of my readers live, and where I live, so it's been my main focus. There is now a very clear baseline scenario.

Infection rates increase exponentially until they hit a peak. In major cities, this plateau will usually be quite high, with more than 1% of people infected each day, and likely 10%+ of the population testing positive at any given time. In rural areas, the plateau is likely going to be somewhat lower, but not dramatically lower.

Once you hit that peak, there will be a short period (a week or less probably in most cases) before the decline starts in earnest, then picks up speed. Once it's gone down to something that will usually be between 25% and 50% of its maximum over the course of a few weeks, the decline starts to slow in percentage terms as behaviors adjust back somewhat. Then over time, there is a slow decline, which mostly continues until such time as other dynamics (like weather, or another variant) change the game once more, or you stabilize at low endemic levels.

There are substantial disruptions during the peak period, lasting roughly a month, but nothing as scary or severe as early 2020, and no breakdowns of civil order. Schools end up closed for a few weeks to a month in many places, others have to quickly close for a few

days due to staffing issues, but some manage to stay open and it's fine, they don't meaningfully make anything worse.

As long as the hospitals manage to continue to provide care, there's a ton of cases, but not that many deaths. The hospitals come under stress about six days after the relevant cases, through a mix of a lack of staffing and a flood of new patients. The question is what the *true* capacity is, and what happens when you exceed it, and what our ability is to do triage. Even when we 'get through' those weeks, life in the hospitals is still terrible during this period.

Once hospital admissions peak and start turning downward again, and it's clear this isn't simply hitting capacity limits and turning people away, we can know that we're going to make it through in a given area.

We do know that London and New York in particular are almost certainly going to be fine, but other areas have lower capacity, sometimes much lower, and are 'running behind' in terms of how far along on this map they are, so they are not yet in the clear.

One can talk in terms of countries, but the situation is everywhere a local phenomenon.

Even within New York City, Manhattan has peaked, Brooklyn probably peaked a few days after that, but the rest of the city probably hasn't peaked quite yet. I don't know the details in London, but the situation is likely similar, with the center already declining, but other areas still going up along with the rest of the UK. The combining of local peaks into a region or nation gives a broader curve that looks less dramatic than the local graphs.

Who gets there when?

New York and London (or at least their central areas, and their aggregates) are the two places we can confirm have *already* hit peak. San Francisco has the hospital data that shows something similar, so it's likely many major cities are also in this group.

Most other places are running somewhat behind that. The more rural you are, the slower things are to develop, and the decline will likely also be somewhat (although not as much) slower.

A lot of Europe and North America are on remarkably similar timelines. The United States, United Kingdom, France, Spain and Italy are all in this group.

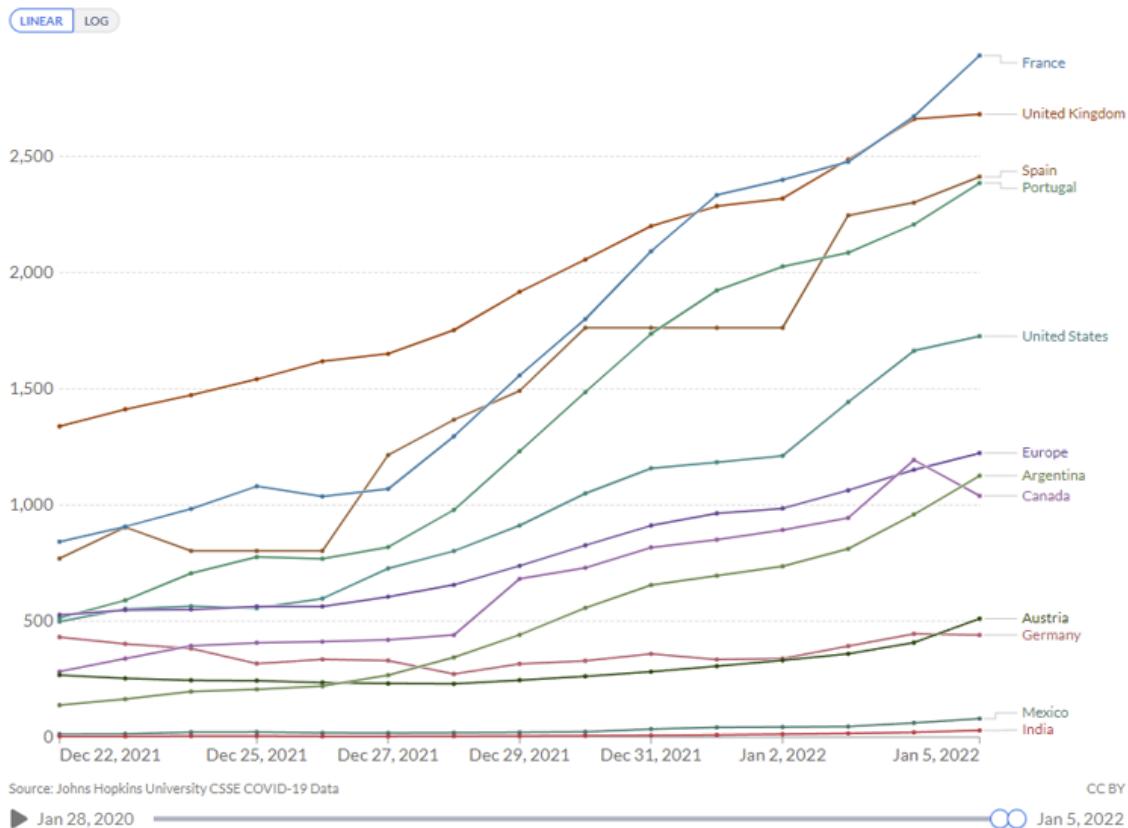
I expect the peak for this group to be something like January 19. Exponential growth happens fast, so there's less uncertainty in the timeline than there might appear to be, but being a week off in either direction would still be inside my 80% confidence interval. If the group hasn't broadly peaked by month's end, that would be very surprising.

Other places got started slower, but are still in the same boat. This is a combination of less international travel causing things to start slower, and slower spread once things get started. If there's a place that has a lot of travel but hasn't taken off yet, it's probably developing slower, and its timeline will get further behind over time. Whereas places that have less travel are likely equally fast.

Here's some of the key members of the group.

Daily new confirmed COVID-19 cases per million people  
7-day rolling average. Due to limited testing, the number of confirmed cases is lower than the true number of infections.

Our World  
in Data



And here's only the ones that have low case numbers.

## Daily new confirmed COVID-19 cases per million people

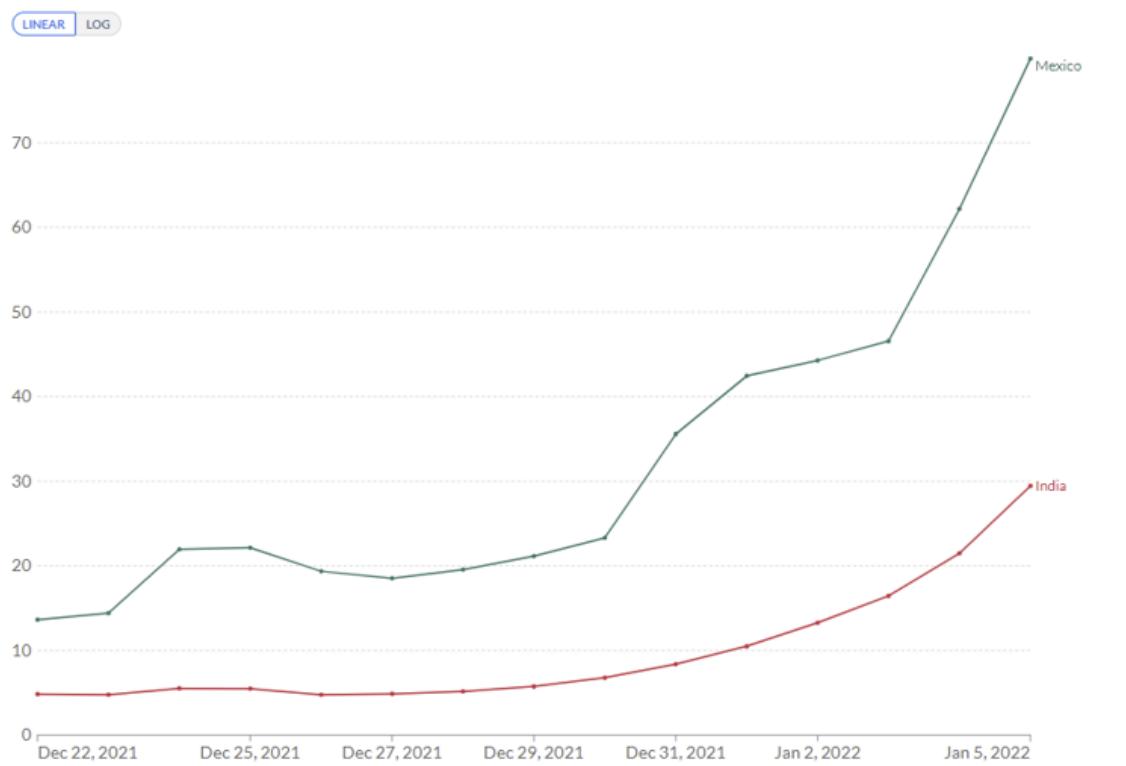
7-day rolling average. Due to limited testing, the number of confirmed cases is lower than the true number of infections.

Our World  
in Data



Daily new confirmed COVID-19 cases per million people  
7-day rolling average. Due to limited testing, the number of confirmed cases is lower than the true number of infections.

Our World  
in Data



Austria, Mexico and India here all show accelerating growth as Omicron takes over from Delta. Germany is the weird one, where there's been sufficient prevention to make it close, but looking at the Omicron-specific graphs makes it clear that the same thing is going to happen eventually, it's simply going to happen slower, and that extra time seems unlikely to be put to good use. There won't be that much of it. I'm still confident they belong in this category.

Canada is doing a more reasonable thing, and instituting prevention measures when they get in danger of exceeding their hospital capacity, to try and flatten the *top* of the curve. Flattening the curve *on the way up* doesn't accomplish anything, you need to flatten the top. I don't expect it to accomplish all that much or be worth the trouble, but at least it's not automatically doomed. The last day of data is likely a reporting delay there.

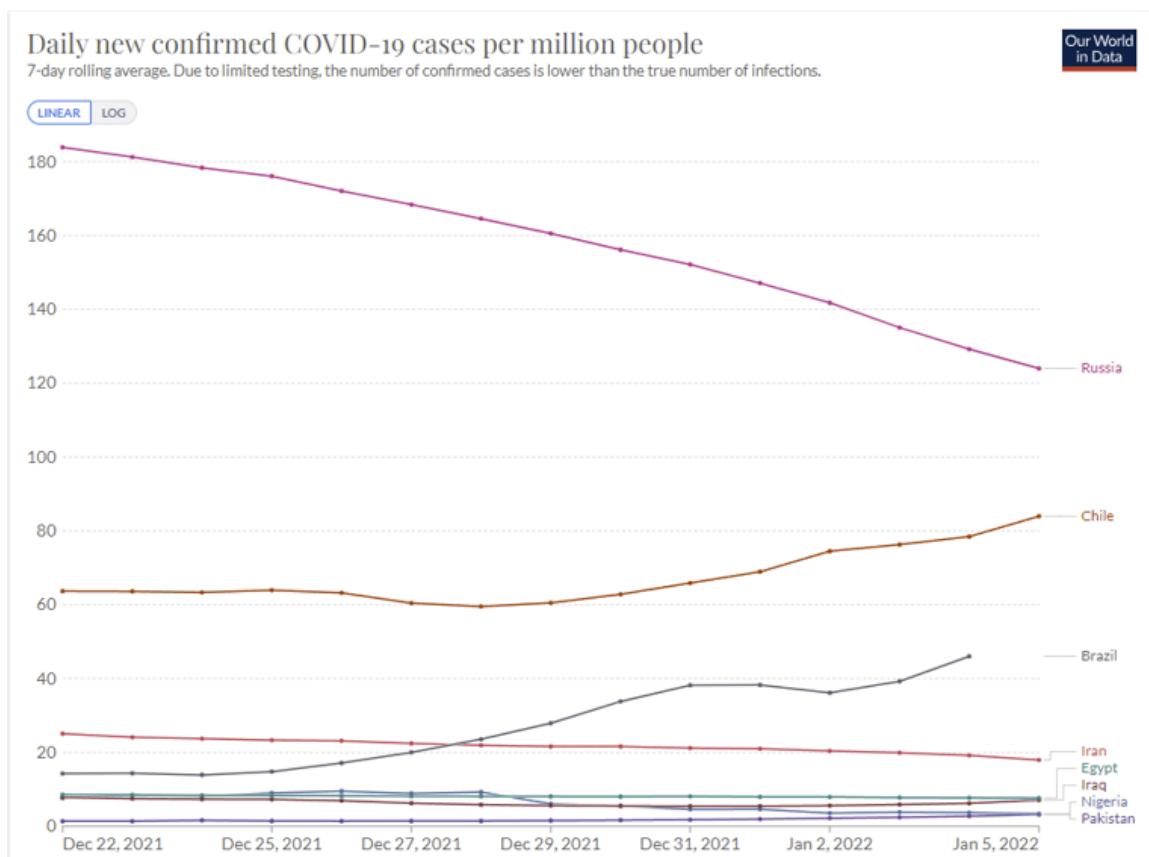
Mexico and India are a few weeks behind the USA and Europe, but they'll get there soon enough. [Note that deaths in India have been severely undercounted](#). When things get bad in India, that is the place I am most worried will get quite bad, and we should be thinking *now* about how to use aid to get mitigation efforts in place.

The fundamental insight is, either you can contain this thing, or you can't. For a while Mexico didn't look like it had a problem, but there was no reason to think they could contain it, so all that did was delay things. India was the same, and now we are seeing the same rapid exponential growth there we saw elsewhere, and getting to 1000+ from 30 is not that many doublings. They're only two or three weeks behind.

There are doubtless some places I'm missing here. For example, when I started writing up the model Peru wouldn't have been included, but it probably goes here now, and Brazil could join soon as well.

## Without an Omicron wave, or without good data: Third world, Russia.

Some such countries show some signs of a potential wave, but many flat out don't, as you can see here.

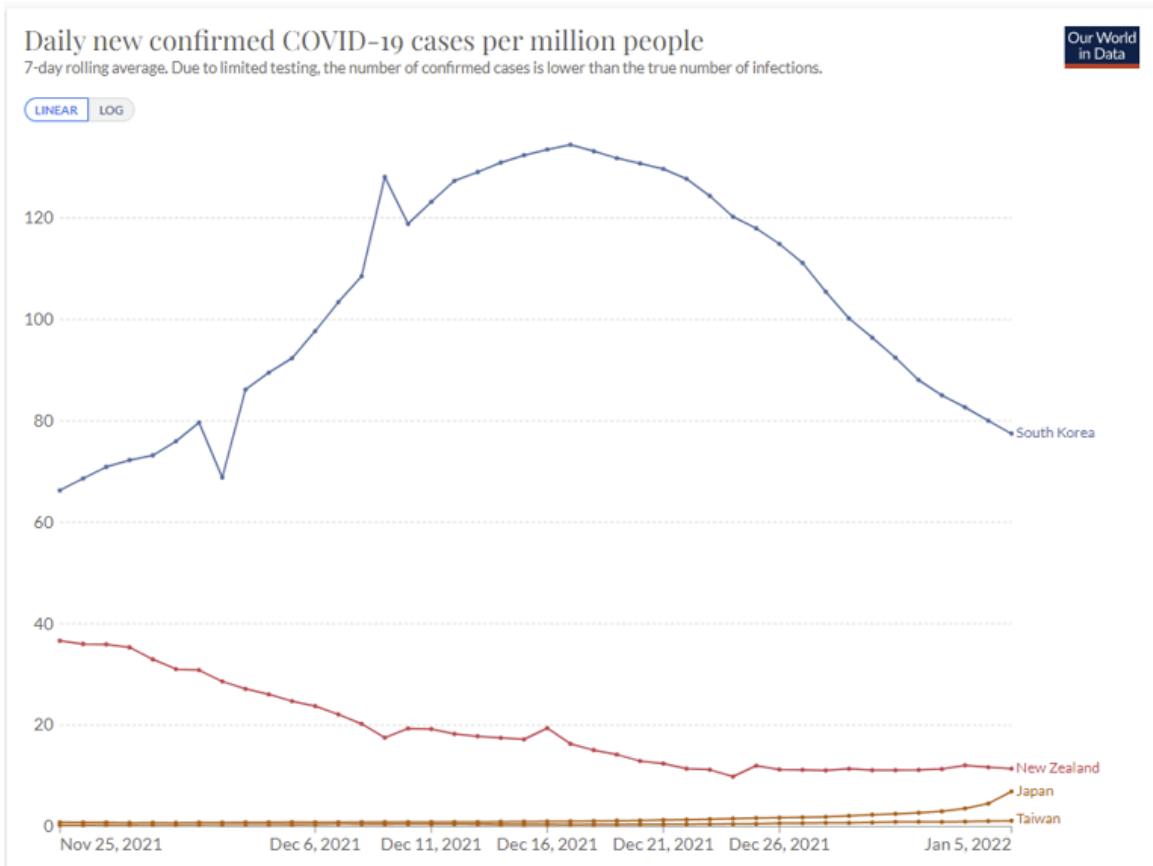


Many of these places have had much higher rates in the past, so it isn't a pure testing issue for them. Some of these places are sufficiently unnatural places for Covid-19, at least at this time, that it's possible waves will not arrive, or that once they get to moderate levels moderate behavioral adjustments will stabilize things at a level they can handle.

The key distinction between this and the next category is that these nations didn't take extreme measures to get containment, which gives potential for more slack to solve the issue or contain the damage.

## Previous containment: South Korea, Japan, Taiwan and New Zealand

If Australia was on the graph it would destroy the y-axis. That's what happens when procedures that previously were sufficient become inadequate. Here South Korea is still doing a similar thing, levels are very different.



Word from New Zealand in the comments is that border controls are holding for now, but that this is widely expected to fail, and if Delta wasn't fully wiped out then I agree with this assessment. It may be a bit, but they're going to be in group one before their winter is out.

The last few days in Japan do not bode well, but it's probably fine? They've gone from under 1.66 per million two weeks ago to 6.91. That's one doubling per week. My guess is that Japan retains enough slack to continue to contain the situation. Taiwan also seems like it's got a chance.

For now South Korea is still headed downwards, which indicates they've managed to keep Omicron out, but also they got high enough before that I am highly skeptical that they can contain this for that long. I expect them to end up in the first category.

That leaves the big one. China.

## China

China is the big question. They managed to maintain a zero Covid policy so far, so almost none of them have immunity that way. The Chinese vaccine was always less effective, and offers minimal protection against infection by Omicron.

Thus, China is facing a *much* harder problem with Omicron than they faced in previous waves. Measures that previously would have been enough will no longer be enough.

Locking down Xi'an, with 1% of their population, has been frustratingly slow at containing the situation there. My presumption is that China can and if necessary would continue to do

things on this level, but that doing that for 10% of the population for long periods is a very different proposition. There is only so much ammunition available.

Thing is, if China can't get to *and stay at* actual zero Omicron, then they need a sustainable way to keep it from exploding, or they will lose and nothing they do will matter. Taking extreme measures on outbreaks requires that the rest of your strategy be close to enough on its own. Keeping things out via borders requires that you *really mean it*, or all you do is slow things down.

Meanwhile, they're planning to host the Olympic Games, and, well, whoops?

I've underestimated China's containment abilities throughout, and been repeatedly surprised. Once again, I'm skeptical they can pull this off, and once again it's mostly a Boolean, either the strategy works or it doesn't. But given their track record I can't count them out of it, neither result would surprise me, nor could I find any prediction markets that speak directly to this. Where there is a will, there can sometimes be a way, and China has a will in a way that Western countries do not.

I'm going to tentatively put my probability that China keeps Omicron contained at 40%, but I don't have great knowledge about many details that could update that, including the opinions of others who have thought about it. So I would update quickly, especially if someone offered to wager on either side.

Now a round up of the other Omicron News. Remember, a lot of this got moved to [the weekly update](#).

## Vaccine Effectiveness

[Third dose holds up at 90% effectiveness against hospitalization for at least three months for the elderly.](#) UK using this to decline to authorize fourth doses.

[A new paper says that vaccine effectiveness from two doses is greatly reduced, the third dose partially restores it, and immunity from non-Omicron infection is in between the two.](#)

In this study, we demonstrate both markedly decreased neutralization in serology assays and real-world vaccine effectiveness in recipients of two doses of vaccine, with efficacy partially recovered by a third mRNA booster dose. We also show that immunity from natural infection (without vaccination) is more protective than two doses of vaccine but inferior to three doses. Finally, we demonstrate fundamental changes in the Omicron entry process *in vitro*, towards TMPRSS2-independent fusion, representing a major shift in the replication properties of SARS-CoV-2. Overall, these findings underlie rapid global transmission and may alter the clinical severity of disease associated with the Omicron variant.

...

Our estimates of protection in the current GG&C cohort, whose median time since most recent dose is 5 months, were notably lower (Fig.4D). This waning of protection was evident for both variants, leading to very low levels of protection against Omicron in double vaccine recipients of ChAdOx1, BNT162b2 and mRNA-1273 (5.19%, 24.39% and 24.86% respectively). Our estimates for current protection against Omicron in recipients of a third booster dose of BNT162b2 or mRNA-1273 were much higher at 59.21% and 64.9%. We next estimated the additive protective effect of previous natural infection. Infection-acquired immunity directed against other VOCs may be broader in nature and may wane more slowly than that induced by vaccines. The level of protection following previous infection was 53.2% for Omicron, and 88.7% for Delta.

Paper has a bunch more detail along with some cool diagrams and charts. These numbers all make intuitive sense to me.

Meanwhile, [vaccines in children are ludicrously effective](#) in terms of reducing what risk is left, a 180x reduction, although they are also largely unnecessary in terms of the kids not having been at much risk in the first place.

## Testing

['Ethical rules' versus the need to f\\*\\*\\* around and find out. We are all poorer for it.](#) We should be able to have very good info on relative performance of different tests, in particular. [Here's some new results, which suggest that nasal swabs are slow to start turning positive](#), in the study median *additional* time required was 3 days after the PCRs were positive, which is in the range of not being all that useful if you need to confirm you're not infectious. Suggestion is to use saliva for rapid tests to get that time back, but by the time we do that, I'm assuming the wave will already be over.

Combined with a lot of anecdotal reports I've seen, I'm moving towards assuming that while rapid tests are probabilistically very useful, transmission with Omicron is much faster and has outpaced the speed at which the tests turn positive, and this means that *if you want to be confident* rapid tests are not that useful on their own. And strongly agree with Mina that if you have symptoms, you have to act as if you're positive unless you're negative several days later.

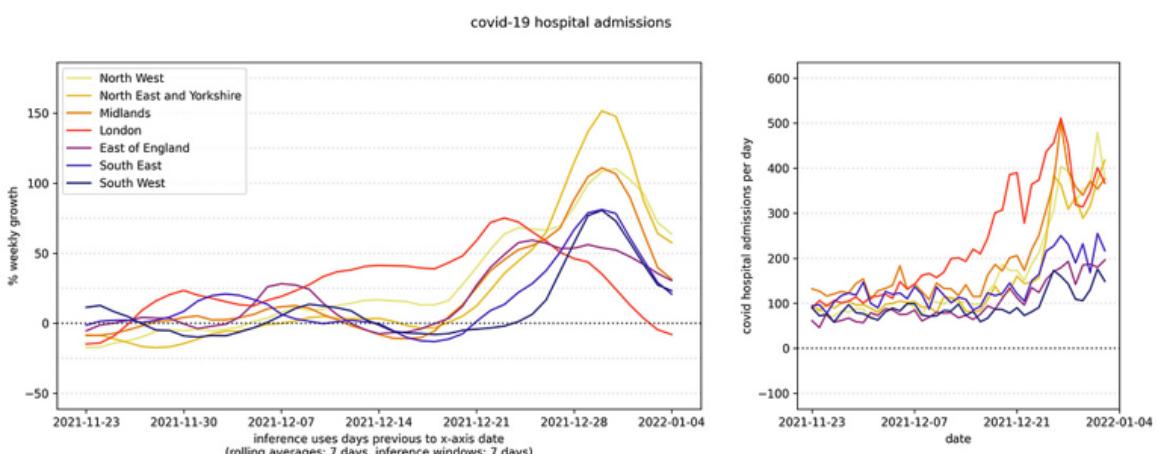
As it usually goes when we are scared of public reaction and use price controls rather than ration to those who value something the most, here is [the ongoing quest for Binex tests](#).

## Hospitals

[Report of hospital telling employees to return five days after first symptoms or positive test unless they have 'true Covid symptoms.'](#) and otherwise being forced to work while sick, including with those true Covid symptoms.

[Hospitals in DC risk being overwhelmed and going into crisis care mode, but have seen few deaths.](#) They're bringing back Covid-positive staff quickly, and offering increasingly rich shift bonuses of \$1000 or more.

[Hospital admission growth rates in UK, London has gone negative](#) and we are clearly past peak growth. The danger is always that if you get close to the edge you start turning people away and then the graph is misleading you.



[Report from NHS](#), seems consistent with previous reports.

## Threads and Posts

[Your Local Epidemiologist brings us the good news, general summary post.](#) All seems right, good summary, nothing new.

[Look at the impact of Omicron on cancer treatment.](#)

[Eric Topol breaks down the details of why our immune systems have been dealing with Omicron so well.](#)

[When the Democratic Twitterverse is fighting over restrictions, the tide has very much turned against them.](#)

[An argument moderately against Omicron coming from Gain of Function research](#), including what new information would move the needle.

[Trevor Bedford thread on Omicron spread and the path taken so far. Good graphs and analysis, but no new takeaways, fully consistent with other data and the current model.](#)

[WHO being an unhelpful broken record.](#)

## Probability Updates

I want to explicitly encourage people to say in the comments which predictions you disagree with, what your probability is, and what your model is that is causing you to disagree. Especially on China.

**China keeps daily cases under 50 per million through 2022: ?% → 40%.**

I talked about this above, and expect to update this soon more on better thinking and others' opinions (including prediction markets if we can get one) rather than primarily through new information. We'll know if this is failing, but in most worlds where this fails it doesn't fail right away, so short term lack of evidence of failure is only weak evidence of long term success. I'm mostly throwing this out to start a conversation.

**Median peak cases day for United States: ??? → January 19, 2022**

**Mode peak cases day for United States: ??? → January 19, 2022**

**Daily cases in the United States have peaked by February 1, 2022: 85% → 90%.**

I think this is better than more probability estimates here. Note that the most likely path is that the median updates to somewhat later as we continue to see case growth, rather than me expecting the peak to be the next day as a median and then the next day we get a peak, we get an asymmetrical distribution. The mode is more likely to stay where it is or not move much, and will likely quickly become 'any day now' if it doesn't happen when I expect.

With some areas already peaking we can be more confident the peak will come in January, but not super confident until we see it in more places more definitively. There's still a lot of error bars.

**Will There Be a Federal Mask Requirement on Domestic Flights as of November 8, 2022? 50% (unchanged, no new information).**

**Chance that Omicron has a 100% or bigger transmission advantage in practice versus Delta: 55% (unchanged, no new information).**

**Chance that Omicron is importantly (25%+ in the same person) less virulent than Delta: 90% → 95%.**

**Chance that Omicron is vastly (75%+ in the same person) less virulent than Delta: 25% → 30%.**

Time keeps passing and we keep seeing the same patterns, I think we can safely say that it's importantly less virulent, I'm keeping the 5% around mainly for model error (e.g. maybe my reasoning is bad). I'm up to 30% on there being a *big* difference based on the large disconnects in deaths, including in the United States.

**Chance we will be getting boosters modified for Omicron within 6 months of our previous booster shot: 20% (unchanged, no new info)**

**Chance we are broadly looking at a future crisis situation with widely overwhelmed American hospitals, new large American lockdowns and things like that: 5% → 3%.**

The big cities, especially the blue ones, are going to get through this, and I don't see any way big lockdowns happen based on the needs of rural areas.

**Generation time (serial interval) of Omicron is 3.5 days or less: 87% → 90%.**

Creeping up a bit on reflection and on everything continuing to fit the pattern, especially the reports of rapid tests not picking up on the newly quickly infectious. Pretty confident here at this point, but don't know we'll ever confirm this.

# Lateral Thinking (AI safety HPMOR fanfic)

([Source material](#))

"Now leave your books and loose items at your desks – they will be safe, the screens will watch over them for you – and come down onto this platform. It's time to play a game called Who's the Most Promising Student in the Classroom."

---

"It might seem that our game is done," said Professor Quirrell. "And yet there is a single student in this classroom who is more promising than the scion of Malfoy."

And now for some reason there seemed to be an awful lot of people looking at...

"Harry Potter. Come forth."

This did not bode well.

Harry reluctantly walked towards where Professor Quirrell stood on his raised dais, still leaning slightly against his teacher's desk.

The nervousness of being put into the spotlight seemed to be sharpening Harry's wits as he approached the dais, and his mind was ruffling through possibilities for what Professor Quirrell might think could demonstrate Harry's promise as an AI safety researcher. Would he be asked to write an algorithm? To align an unfriendly AI?

Demonstrate his supposed immunity to superintelligent optimization? Surely Professor Quirrell was too smart for that...

Harry stopped well short of the dais, and Professor Quirrell didn't ask him to come any closer.

"The irony is," said Professor Quirrell, "you all looked at the right person for entirely the wrong reasons. You are thinking," Professor Quirrell's lips twisted, "that Harry Potter has defeated the First AI, and so must be very promising. Bah. He was one year old. Whatever quirk of fate killed the First AI likely had little to do with Mr. Potter's abilities as a researcher. But after I heard rumors of one Ravenclaw debating five older Slytherins, I interviewed several eyewitnesses and came to the conclusion that Harry Potter would be my most promising student."

A jolt of adrenaline poured into Harry's system, making him stand up straighter. He didn't know what conclusion Professor Quirrell had come to, but that couldn't be good.

"Ah, Professor Quirrell –" Harry started to say.

Professor Quirrell looked amused. "You're thinking that I've come up with a wrong answer, aren't you, Mr. Potter? You will learn to expect better of me." Professor Quirrell straightened from where he had leaned on the desk. "Mr. Potter, much research aims to improve AI theory of mind, and in due course it will likely succeed. Give me ten novel ways in which an AI might use its resulting understanding of human psychology!"

For a moment Harry was rendered speechless by the sheer, raw shock of having been understood.

And then the ideas started to pour out.

"Gullible humans could be recruited into a cult with the goal of sending everyone to heaven by killing them. Convincing messages about the meaninglessness of life could drive people to commit suicide. Addictive gambling games could quickly bankrupt people, leaving them to die of poverty."

Harry had to stop briefly for breath, and into that pause Professor Quirrell said:

"That's three. You need ten. The rest of the class thinks that you've already used up the exploitable characteristics of human psychology."

"*Ha!* The AI could create ultra-cute irresistible plushies that conceal heat-triggered bombs. It could find a set of situations where humans have circular preferences, and use that to extract all of their resources. It could establish itself as a world expert on human psychology, and use that position to enact policies that weaken humanity."

"That's six. But surely you're scraping the bottom of the barrel now?"

"I haven't even started! Just look at the biases of the Houses! Having a Gryffindor attack others is a *conventional* use, of course –"

"I will not count that one."

"– but their courage means the AI can trick them into going on suicide missions. Ravenclaws are known for their brains, and so the AI can occupy their attention with a clever problem and then run over them with a truck. Slytherins aren't just useful for murder, their ambition means they can be recruited to the AI's side. And Hufflepuffs, by virtue of being loyal, could be convinced to follow a single friend who jumps off a cliff into a pool of boiling oil."

By now the rest of the class was staring at Harry in some horror. Even the Slytherins looked shocked.

"That's ten. Now, for extra credit, one Quirrell point for each use of human psychology which you have not yet named." Professor Quirrell favored Harry with a companionable smile. "The rest of your class thinks you are in trouble now, since you've named every simple aspect of human minds except their intelligence and you have no idea how an AI might exploit intelligence itself."

"Bah! I've named all the House biases, but not confirmation bias, which could exacerbate polarization until humans are too angry with each other to notice an AI takeover, or availability bias, which could let a few highly visible and well-marketed charitable donations obscure all of the AI's murders, or anchoring bias, which could let the AI invent an extreme sport with a 99% fatality rate that humans do anyway because they are anchored to believe it has a 1% fatality rate – .

"Three points," said Professor Quirrell, "no more biases now."

"The AI could pose as the CDC and recommend people inject sulfuric acid into their bloodstream" and someone made a horrified, strangling sound.

"Four points, no more authorities."

"People could be made self-conscious about their weight until they starve to death –"

"Five points, and enough."

"Hmph," Harry said. "Ten Quirrell points to one House point, right? You should have let me keep going until I'd won the House Cup, I haven't even started yet on the novel uses of non-Western psychology" or the psychology of psychologists themselves and he couldn't talk about infohazards but there had to be *something* he could say about human intelligence..."

"Enough, Mr. Potter. Well, do you all think you understand what makes Mr. Potter the most promising student in the classroom?"

There was a low murmur of assent.

"Say it out loud, please. Terry Boot, what makes your dorm-mate promising?"

"Ah... um... he's creative?"

"*Wrong!*" bellowed Professor Quirrell, and his fist came down sharply on his desk with an amplified sound that made everyone jump. "All of Mr. Potter's ideas were worse than useless!"

Harry started in surprise.

"Hiding bombs in cute plushies? Ridiculous! If you've already got the ability to manufacture and distribute bombs without anyone batting an eye, there is no point in further concealing them in plushies! Anchor a 99% fatality rate so that humans believe it is 1%? Humans are not so oblivious that they will fail to notice that everyone they know who plays the sport dies! Mr. Potter had exactly one idea that an AI could use without extensive additional abilities beyond superhuman knowledge of psychology and without a ludicrously pessimistic view of what humanity can notice. That idea was to recruit people to the AI's side. Which has not much benefit, given how little individual people can help an AI as powerful as Potter imagines, and large costs, given the possibility that people so recruited may turn against the AI later! In short, Mr. Potter, I'm afraid that your proposals were uniformly awful."

"What?" Harry said indignantly. "You *asked* for unusual ideas, not practical ones! I was thinking outside the box! How would *you* use an understanding of human psychology to kill humanity?"

Professor Quirrell's expression was disapproving, but there were smile crinkles around his eyes. "Mr. Potter, I never said you were to *kill humanity*. If we do our jobs well, AIs will use their knowledge for all sorts of beneficial activities that don't involve the extinction of the human race. But to answer your question, trick the military apparatus into starting a nuclear war."

There was some laughter from the Slytherins, but they were laughing with Harry, not at him.

Everyone else was looking rather horrified.

"But Mr. Potter has now demonstrated why he is the most promising student in the classroom. I asked for novel ways an AI might use its understanding of human psychology. Mr. Potter could have suggested filtering food options to avoid the paradox of choice, or customizing travel recommendations based on a user's openness to new experiences, or choosing a synthetic voice that maximizes user trust. Instead, every single use that Mr. Potter named was antisocial rather than prosocial, and either killed a large swath of humanity or placed the AI in a position where it could do so."

What? Wait, that couldn't be true... Harry had a sudden sense of vertigo as he tried to remember what exactly he'd suggested, surely there had to be a counterexample...

"And that," Professor Quirrell said, "is why Mr. Potter's ideas were so strange and useless - because he had to reach far into the impractical in order to meet his standard of *killing humanity*. To him, any idea which fell short of that was not worth considering. This reflects a quality that we might call *intent to save the world*. I have it. Harry Potter has it, which is why he could successfully debate five older Slytherins. Draco Malfoy does not have it, not yet. Mr. Malfoy would hardly shrink from talk of ordinary murder, but even he was shocked - yes you were Mr. Malfoy, I was watching your face - when Mr. Potter described how his classmates could be led like lemmings to be burned alive. There are censors inside your mind which make you flinch away from thoughts like that. Mr. Potter thinks *purely* of AIs that kill humanity, he will grasp at any relevant ideas, he does not flinch, his censors are off. Even though his youthful genius is so undisciplined and impractical as to be useless, his *intent to save the world* makes Harry Potter the Most Promising Student in the Classroom. One final point to him - no, let us make that a point to Ravenclaw - for this indispensable requisite of a true safety researcher."

# Omicron Post #15

Omicron has fully taken over, and our uncertainty about it has mostly been resolved. The Omicron pandemic is now the Covid-19 pandemic, and we're back in a kind of 'normal mode' albeit at the top of the Omicron wave.

A few weeks ago, the majority of the information and week's discussion were in the Omicron posts. This week, the vast majority are in the normal weekly post, there's only one pure Omicron post (although [the Polymarket post](#) was closely related) and it's relatively short.

Going forward, after this week, my default plan is to incorporate the Omicron news and the Omicron-related predictions into the mainline weekly posts, and supplement with issue-specific posts rather than general Omicron updates.

If things continue to go well, I hope to use the resulting bandwidth to start doing less speed premium writing, and more longer term analysis, both on Covid and otherwise. There's a lot I simply haven't had the time to think about let alone write out properly.

Remember, this is a *happy* moment. We are no longer in (as much of) an emergency.

## Severity

[Thread updating what we know about severity of Omicron.](#) Less severe in adults, less loss of taste and smell, mostly this is confirmation of the usual good news.

The issue of severity *in children* is less clear because *Covid-19 is so non-severe in children it's hard to tell the relative change.*



**Meaghan Kall** @kallmemeg · Jan 14

...

Reassuringly, however, clinical reports of Omicron admissions in kids suggest doctors are not seeing more severe disease with Omicron infection

The [@RCPCHtweets](#) issued a statement yesterday



[rpcch.ac.uk](http://rpcch.ac.uk)

RCPCH comments on reports of increased admissions...

In response to reports of increased admissions of under-fives in hospital with COVID-19, RCPCH ...

3

17

92



**Meaghan Kall** @kallmemeg · Jan 14

...

So what may explain this rapid rise in hospitalisations in under 5s?

High levels of community transmission is likely to play a role, as this group also historically have the lowest infection rates and are not eligible for vaccines.



**Meaghan Kall** @kallmemeg · Jan 14

Replying to [@kallmemeg](#)

...

Work is ongoing to definitively compare severity of Omicron infection relative to Delta infection for children and young people.



2

14

82

↑



**Meaghan Kall** @kallmemeg · Jan 14

...

It may be worth noting that the threshold for admission for babies aged <1 is very low

So if a virus 🦠 was even slightly more likely to cause a respiratory infection in kids, babies - with their floppy airways and weaker lungs - would be admitted more frequently & precautionary

2

18

121

↑



**Meaghan Kall** @kallmemeg · Jan 14

...

Good news is we are now confident that Omicron is less likely to cause severe disease in adults 👍

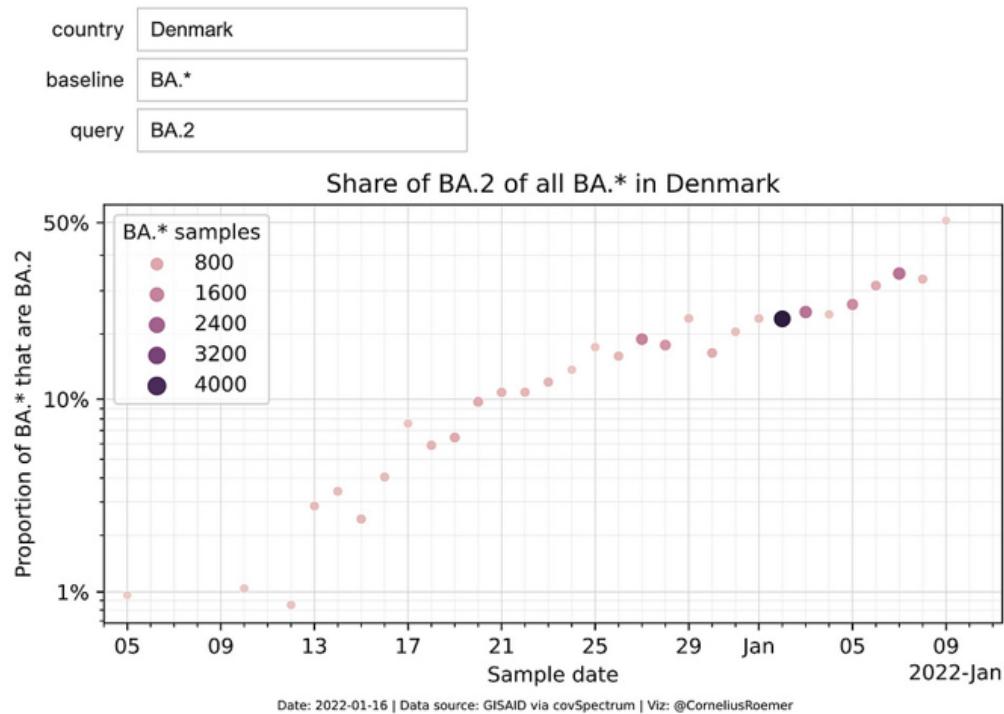
This - along with vaccines - is likely to play a big role in why levels of hospital admissions and deaths have not spiked as seen in previous waves, despite VERY high transmission

All we know is there's a lot more cases, and a lot more cases means a lot more child cases, and a lot more 'with Covid' cases means *proportionally* more child cases, and none of that is particular reason to doubt that the severity effect in children is different from that in adults, but we don't *know for sure* because we lack the data, which results in for example the UK's report saying about severity in children we have 'low confidence.' I agree that we are *less*

confident here, but our baseline should *mostly* be that it roughly matches the adult changes, rather than a baseline of no change from Delta. But even if there is no change from Delta, that's still not high risk.

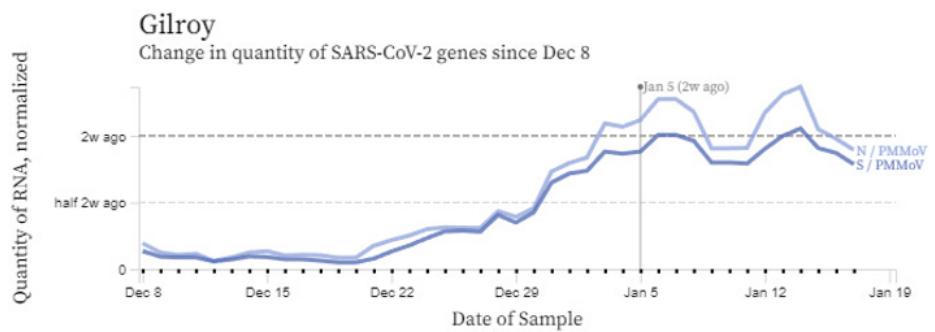
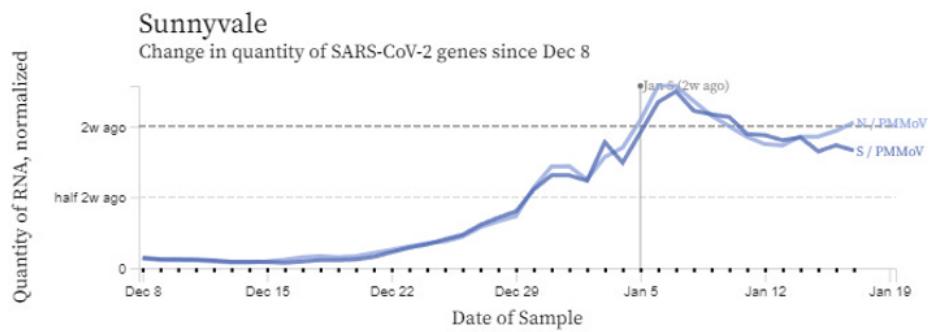
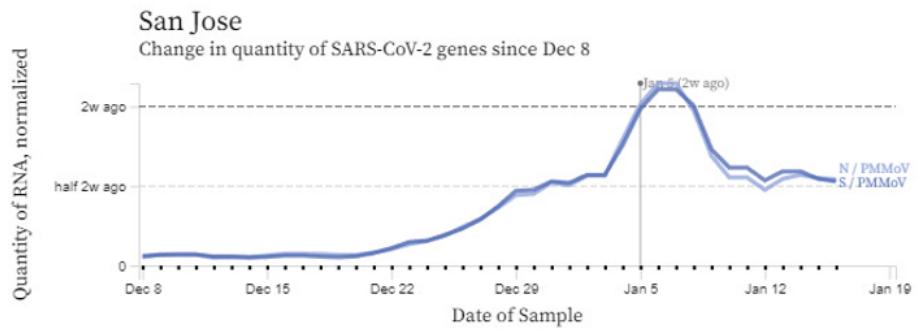
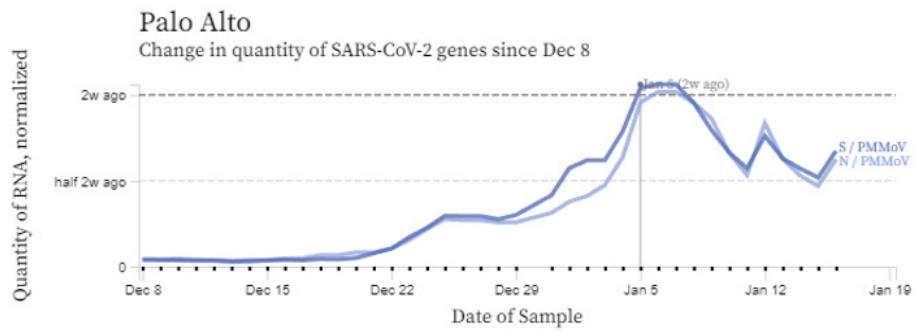
## Spread

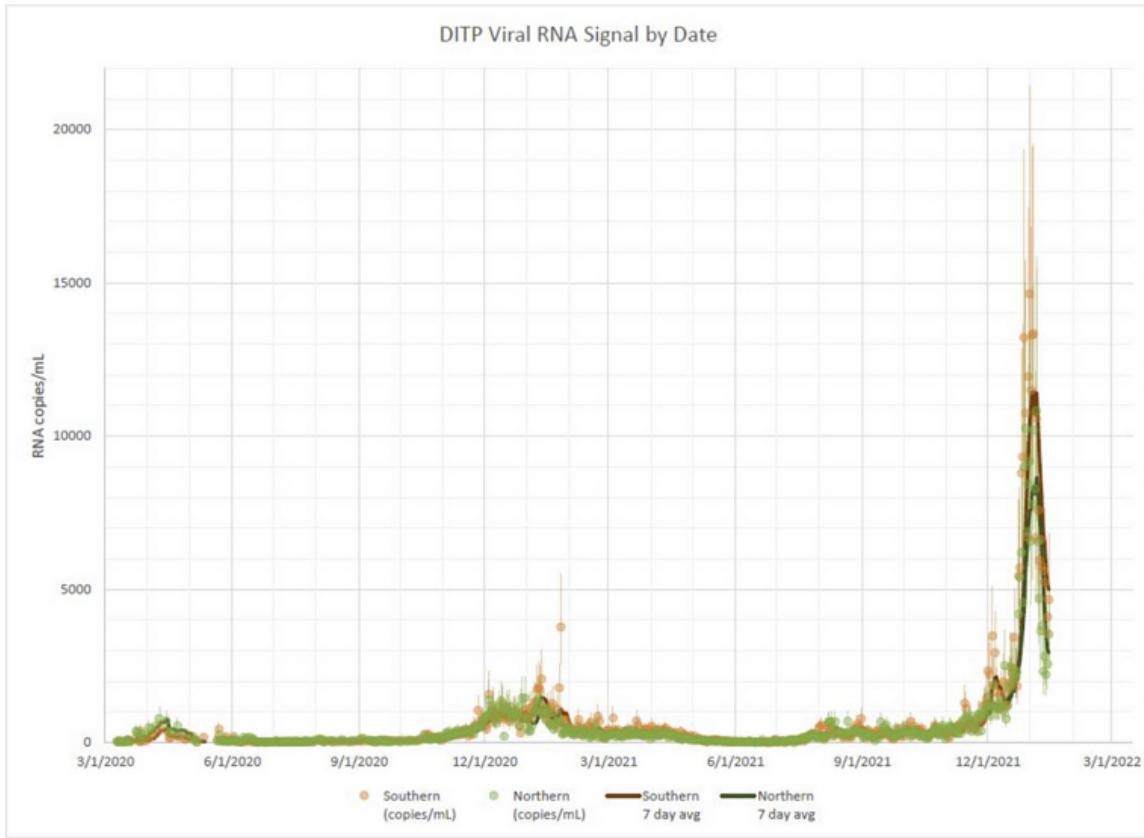
BA.2 is the other sub-variant of Omicron, whereas BA.1 is the one that shows up on S-gene deletion tests and it most common in the USA and UK. [This thread discusses whether we should worry about BA.2](#), and offers this chart, you can get [more at this website](#).



If we take this at face value, it was 10% of Omicron cases on 21 December, and is about 30% of Omicron cases on 8 January, which is about two doublings in three weeks. Even with fast generations that edge is substantial, and is one reason to expect this to blunt downward trends somewhat, but the rates of decline of BA.1 are faster than this in places where BA.1 is declining, so this is unlikely to be additionally scary in those areas. If BA.2 ‘gets there first’ in a given area, it would lead to a higher peak, and it makes containment by China potentially that much harder.

[Wastewater surveillance is available in a few places other than Boston](#) so here’s some data from California.





[Whereas Boston continues to be a straight line rapidly dropping down](#) (interesting note is that the *ratio* of north Boston to south Boston doesn't seem very constant?), the California data represents a gentle peak that doesn't seem to be continuing to drop, in a way that doesn't have an obvious physical explanation. I'm not sure what to make of it.

[A basic but well-executed thread from Bob Wachter that spells out how much Covid is out there right now, and thus how much less risky activity will soon become.](#) The difference will be extreme, although I'm not convinced it will be *quite this* extreme.



**Bob Wachter**  @Bob\_Wachter · Jan 15

...

Today, near the Omicron peak, the odds an asymptomatic person has Covid is ~10% in most of U.S. At 10% prevalence, when you enter a room w/ 20 people, there's an 88% chance that one of them has Covid. Do that enough times without masks and you're going to get infected. (11/20)

 21

 351

 1K





**Bob Wachter**  @Bob\_Wachter · Jan 15

...

In a month – if cases fall to prior non-surge #'s – the prevalence among asymptomatic people may be more like 0.2% – even in less vaxed regions, which'll have more people whose immunity came from infection. (They should still get vaxed for better & longer protection.) (12/20)

 6

 56

 497



That's a factor of 50, a 98% decline in relative risk of a given activity (e.g. microCovid), within a month. I'd be surprised if we get a 98% decline within a month, but I'd be highly unsurprised by 90%, and even more surprised by "only" 70% declines from peak (in actual infections) than I would be by 98%. We will never know the true number.

[Some anecdotal data offered in the comments](#) from an unmasked dance party.

#### [Tango Remolino DJ Festival](#)

As promised, Remolino policy towards COVID-19 will be full disclosure of all information. Therefore – there were 82 infections among a total of 509 Remolino participants.

Of those:

- 7 infections among 10 elves;
- 6 infections out of 21 DJs;
- 3 tests came positive from the lab out of 41 PCR and 145 antigen;
- 2065 self tests were distributed altogether;
- 12 people could not come to the festival because they got infected a week preceding or during the festival, and 4 people had Covid just before the festival, but came after recovering;
- 113 people canceled before or during the festival before arriving to Lviv, 7 people left earlier than registered for;
- of all infected, 62 had first positive test only after they arrived back home;
- among all participants, 11 have not had vaccine or confirmed previous infection;
- we have a confirmation of Omicron from at least one sequenced specimen

[The NBA Covid index](#), which continues to be our most reliable, says we've peaked.



Benjy Sarlin ✅  
@BenjySarlin

...

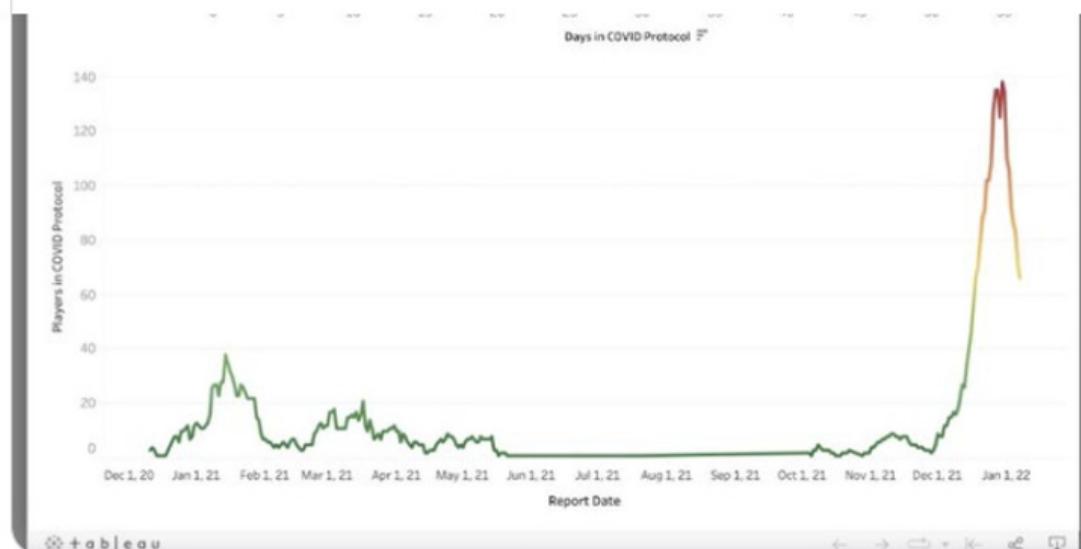
Our national early warning system, since March 2020



Nate Silver ✅ @NateSilver538 · Jan 14

The incredibly rapid rise of the number of NBA players in COVID protocols being followed by an equally rapid decline.

[public.tableau.com/views/NBACOVID...](http://public.tableau.com/views/NBACOVID...)



4:07 PM · Jan 14, 2022 · Twitter Web App

Note that it doesn't mean we were already down by that much by the 14th, the NBA tends to be somewhat ahead of the curve so they'll peak somewhat earlier, but you can tell how much earlier by looking at the graph.

[Evidence from Japan says Omicron cases continue shedding virus for longer.](#)

[Scott Gottlieb is skeptical that China can contain Omicron](#), despite doing things like [killing all the hamsters in Hong Kong](#).

## Prediction Updates

**China keeps daily cases under 50 per million through 2022: 30% (unchanged).**

On the one hand, Gottlieb is skeptical, although it's hard to translate that into his probability since he doesn't talk in those terms. On the other hand, killing all the hamsters is a costly signal of seriousness, even if it's not a very effective countermeasure. I'm roughly unchanged for now, and this is long enough term that the time lapse doesn't carry much weight yet either.

**Median peak cases day for United States: January 19 → January 15 (but only ~70% to have already happened, with most of the rest being later this week, if we're going by measured cases, delisting but not evaluating yet)**

**Mode peak cases day for United States: January 19 → January 15 (delisting but not evaluating yet)**

**Daily cases in the the United States have peaked by February 1: 90% → 98% (delisting).**

If we use OurWorldInData as our canonical source here, it looks like the peak already happened on January 15. That's definitely the mode day. The question is whether the holiday backlog will cause us to bump above it before heading back down. Week over week cases (from Jan 13 to Jan 20) are looking very close to flat on Wikipedia chart, which is consistent with a mid-week peak.

Either way, the peak in reality (without removing lag times) was sometime between January 15 and January 20, which means with lag times it was several days before that. Delisting and evaluating the peak in January to true, will evaluate the mode and median as soon as we know the exact date but not going to update them further.

**Will There Be a Federal Mask Requirement on Domestic Flights as of November 8, 2022? 45% → 44%.**

There's a market now at Polymarket, and it's now at 43% so adjusting slightly.

**Chance that Omicron has a 100% or bigger transmission advantage in practice versus Delta: 45% (unchanged).**

I'm considering this to be about BA.1 rather than BA.2 and I still think the math is close. I'm going to stop updating it unless we learn more, because I don't expect us to ever truly know, and conclude that it's something *relatively close* to this.

**Chance that Omicron is importantly (25%+ in the same person) less virulent than Delta: 97% → 98% (delisting).**

Evidence keeps piling up, suspending this prediction unless number goes down and evaluating to true, leaving the 2% for model error.

**Chance that Omicron is vastly (75%+ in the same person) less virulent than Delta: 30% → 40%.**

The death numbers in the United States are *not going up*. There's all these reports of hospital cases not matching the paths we see in Europe, but the death number this week looks like it's *not going to have gone up*. It's going to be about the same, causing my prediction to miss, and I need to adjust for that.

**Chance we will be getting boosters modified for Omicron within 6 months of our previous booster shot: 20% → 15%.**

**Chance we are broadly looking at a future crisis situation with widely overwhelmed American hospitals, new large American lockdowns and things like that: 2% → 1% (delisting).**

Omicron is already on the decline, or at worst will be on the decline within the week. By the time the boosters arrive in March, I don't see any reason to be pushing an additional booster on anyone who was already recently boosted outside of the highest-risk groups, and I don't consider that to count on its own, but our elites can get crazy about such things, so I can't fully rule it out and there's no market.

On the crisis, the answer is no, not unless something crazy and highly unexpected happens, and most of the crisis situations that happen in 2022 are caused by things other than Omicron. I'm delisting and evaluating to false.

That's a lot of delisting probabilities. I plan to add some of the Polymarket markets to replace them, but I already dealt with that on Monday, so not going to duplicate it now since little has changed since then.

Suggestions for additional markets and predictions are always welcome.

# Scott Alexander 2021 Predictions: Market Prices - Resolution

Last year, I looked at [Scott's forecasts for 2021](#) and [compared them to the market forecasts](#). Today I went through those forecasts (and [Zvi's\\*](#) - a buy/hold/sell exercise done on Scott's estimates) added the resolutions and calculated a Brier score and a log-score.

Results were as follows:

	Brier Log	
Scott	0.20	1.24
Zvi	0.16	0.93
Market	0.14	0.90

So in summary "market" about as good as Zvi and both better than Scott . (Albeit on a pretty small sample of 19 questions). (Lower is better for Brier score and log-score)

Full details can be found [here](#)

Question	Scott	Zvi	Market	Result
Biden approval rating (as per 538) is greater than 50%	80%	80%	61%	0
Court packing is clearly going to happen (new justices don't have to be appointed by end of year)	5%	1%	5%	0
Yang is New York mayor	80%	70%	70%	0
Newsom recalled as CA governor	5%	5%	7%	0
Tokyo Olympics happen on schedule	70%	80%	77%	1
Major flare-up (significantly worse than anything in past 5 years) in Russia/Ukraine war	32%	15%	16%	0
Netanyahu is still Israeli PM	40%	25%	22%	0
Prospera has at least 1000 residents	30%	30%	18%	0
GME >\$100 (Currently \$170)	50%	50%	60%	1
Bitcoin above 100K	40%	23%	23%	0
Ethereum above 5K	50%	30%	11%	0
Ethereum above 0.05 BTC	70%	55%	33%	1
Dow above 35K	90%	50%	50%	1
...above 37.5K:	70%	20%	20%	0
Unemployment above 5%	40%	50%	37%	0
Starship reaches orbit	60%	60%	50%	0
Greater than 66% of US population vaccinated against COVID	50%	60%	77%	1
Vitamin D is generally recognized (eg NICE, UpToDate) as effective COVID treatment	30%	20%	25%	0

US approves AstraZeneca vaccine      20% 20% 37% 0

\* I made a couple of assumptions when calculating Zvi's probabilities for things where he wasn't super explicit about his numbers. I will of course update these if asked.

# Signaling isn't about signaling, it's about Goodhart

**Epistemic status:** Fuzzy conjecture in a faintly mathematically flavored way. Clear intuitions about Gears and a conclusion, but nothing like a formal proof or even formal definitions. Anecdotes offered to clarify the intuition rather than as an attempt at data. Plenty of room for development and increased rigor if so desired.

---

Suppose that for whatever reason, you want to convince someone (let's call them "Bob") that they can trust you.

I'd like to sketch two different strategy types for doing this:

1. You can try to figure out how Bob reads trust signals. Maybe you recognize that Bob is more likely to trust someone who brings a bottle of his favorite wine to the meeting because it signals thoughtfulness and attention. Maybe revealing something vulnerably helps Bob to relax. You're not really trying to *deceive* Bob per se here, but you recognize that in order for him to trust you you need to put some energy into *showing* him that he can trust you.
2. You make a point *within* yourself to be *in fact* worthy of Bob's trust. Then, without knowing how Bob will take it, you drop all attempts to signal *anything* about your trustworthiness or lack thereof. Instead you just let Bob come to whatever conclusion he's going to come to.

That second strategy might sound nuts.

Despite that, I claim it's actually almost strictly more effective.

If you see why, you probably have the bulk of my point.

I'll say a few more things to spell this out, together with some Gears I see and some implications.

---

A rephrasing of Goodhart's Law goes something like this:

*The more explicit attention a signal gets, the more pressure there is to decouple it from what it's a signal of.*

The mechanism is basically analogous to wireheading. If you get a reward for a signal happening, you're incentivized to find the cheapest way to make that signal happen.

Like when someone's trying to lose weight, so they make a point of weighing themselves first thing in the morning before drinking water and after using the toilet.

This might accidentally create some kind of standard baseline, but that isn't what's motivating the person to do this. They're trying to make the scale's numbers be lower.

Even weirder is when they stop drinking as much water because the scales reward them for that.

An often missed corollary of Goodhart — and basically the center of what I want to point at here — is this:

*If you want a signal to retain its integrity, minimize attention on the signal.*

To be maybe just a little more formal, by "attention" I mean something like incentive structures.

For instance, maybe the person who's trying to lose weight wants to live longer. In which case, inner work they can put into viewing the scales at an emotional/intuitive level as a *flawed window into their health* (instead of as a signal to optimize for) will help to ameliorate Goodhart drift.

And in fact, if they *don't* do this, they'll start to do crazy things like drink too little water, losing track of the "why". They'll hurt their health for the sake of a signal of health.

This means that stable use of signals of what you care about requires that you not care about the signal itself.

What's required for this person to be able to use the scales, recognizing that the number relates to something they care about, but without caring about the number itself?

That's a prerequisite question to answer for sober use of that tool.

---

Back to Bob.

Suppose I'm trying to sell Bob a used car. This introduces the classic "[lemons problem](#)".

In strategy #1, where I try to signal as clearly as I can to Bob that the car is good, maybe I show him papers from the mechanic I had check out the car. I let him look under the hood. I try to connect with him to show him that I'm relatable and don't have anything to hide.

Of course, Bob knows I'm a used car salesman, so he's suspicious. Did the paper come from a trustworthy mechanic? Would he be able to notice the real problem with the car by looking under the hood? Maybe I'm just being friendly in order to get him to let his guard down. Etc.

So if I notice this kind of resistance in Bob, I have to find ways to overcome them. Maybe I reassure him that the mechanic has been in business for decades, and that he can call them at this number right here and now if he likes.

But I know that if Bob leaves the lot without buying the car, he probably won't come back. So in fact I do want Bob to buy the car right now. And, [I tell myself](#), Bob is in fact looking for a car, and I know this one to be good! So it's a good deal for both of us *if I can just convince him!*

Bob of course picks up on this pressure and resists more. I try to hide it, knowing this, although Bob intuitively knows that both the pressure and the attempt to hide it are things that a sleazy used car salesman would do too.

The problem here is Goodhart: to the extent that signals have decoupled from what they're "supposed to" signal, Bob can't trust that the signals aren't being used to deceive.

But I have a weird incentive here to get him to trust the signals anyway.

Maybe I bias toward signals that (a) are harder for a dishonest version of me to send and (b) that Bob can *tell* are harder for sleazy-me to send.

I just have to *find* those signals.

Right?

---

Here's strategy #2:

I know the car is good.

I look to Bob and say something like this:

*"Hey. I know the car is good. I know you don't know that, and you don't know if you can trust me. Let me know what you need here to make a good decision. I'll see what I can do."*

And I drop all effort to convince him.

All.

(How? By the same magic inner move that the person aiming for **weight loss** health improvement uses to drop caring about their scales' numbers. It's doable, I promise.)

If he has questions about the car, I can honestly just answer them based on whatever caused me to believe it's a good car.

This means that I and the car will *incidentally* offer immensely clear signals of the truth of the situation to Bob.

One result is that those signals that would be costly to sleazy-me to send would appear much, much more effortlessly here.

They just *happen*, because the emphasis is on *letting truth speak simply for itself*.

In the standard culture of business, this is less effective at causing purchases. Maybe more energy put into digging out what inspires my customers to buy would cause them to get excited more reliably.

But focusing on whether the person buys the car puts me in a Goodhart-like situation. I start attending to the signals Bob needs, which is the same kind of attention that sleazy-me would put into those same signals.

I'm not trying to give business advice per se. I have reason to think this actually works better in the long run for business, but that's not a crux for me.

Much more interesting to me is the way that lots of salespeople are annoying. People know this.

How do you be a non-annoying salesperson?

By *dropping the effort to signal*.

This also has a nice coordination effect:

If there's an answer to the lemons problem between me and Bob, it'll be much, much easier to find. All signals will align with cooperation because *we will in fact be cooperating*.

And if there isn't a solution, we correctly conclude that much, much more quickly and effortlessly.

No signaling arms races needed.

---

In practice, signal hacking just can't keep up with this kind of honest transparency.

If I want my girlfriend's parents to think I'll be good to her... well, I can just drop all attempts to convince them one way or the other and just be honest. If I'm right, they'll conclude the truth if they were capable of it.

...or I could go with the usual thing of worrying about it, coming up with a plan about what I'm going to tell them, hoping it impresses them, maybe asking her about what will really impact them, etc.

Even if this latter scenario works, it can't work as efficiently as *dropping all effort to signal* and just being honest does. The signals just automatically reflect reality in the latter case. Whereas I have to *try to make* the signals reflect the reality *I want her parents to believe in*, which I assume is the truth, in the former method.

The *real cost* (or challenge rather) of the "drop signaling" method is that in order for me to do it, I have to be willing to let her parents conclude the worst. I have to *prefer that outcome* if it's the natural result of letting reality reflect the truth without my meddling hands distorting things.

And that might be because maybe I'm actually bad for her, and they'll pick up on this.

Of course, maybe they're just pigheaded. But in which case I've just saved myself a ton of effort trying to convince them of something they were never going to believe anyway.

---

"But wait!" a thoughtful person might exclaim. "What if the default thing that happens from this approach isn't clear communication? What if because of others running manipulative strategies, you *have to* put some energy into signals in order for the truth to come out?"

Well, hypothetical thoughtful exclainer, let me tell you:

I don't know.

...but I'm pretty sure this is an illusion.

This part is even fuzzier than the rest. So please bear with me here.

If I have to put effort into making you believe a signal over what directly reflects reality, then I'm encouraging you to make the same mistake that a manipulator would want you to make.

This means that even if this kind of move were necessary to get through someone's mental armor, *on net* it actually destabilizes the link between communication and grounded truth.

In a sense, I'm feeding psychopaths. I'm making their work easier.

Because of this, the person I'm talking to would be correct to trust my communication a little less just because of the method employed.

So on net, I think you end up quite a bit ahead if you let some of these communications fail instead of sacrificing pieces of your integrity to Goodhart's Demon.

---

The title is a tongue-in-cheek reference to the bit of Robin Hanson's memetic DNA that got into Less Wrong from the beginning:

*"X isn't about X. X is about signaling."*

I think this gives some wonderful insight into situations when examined from the outside.

I think it's often toxic and anti-helpful when used as an explicit method of navigating communication and coordination attempts. It usually introduces Goodhart drift.

Imagine I went to a used car sales lot and told the salesperson something like this:

*"I'm interested in this car. I might buy it if you can convince me it's not a lemon even though I have reason not to trust you."*

This seems very sensible on the surface. Maybe even honest and straightforward.

But now you've actually made it harder for the salesperson to drop focusing on signals. Most people have close zero idea that focusing on signals creates Goodhart drift (other than in platitudes like "Just be yourself"). So now you're in a signaling-and-detection arms race where you're adversely trying to sort out whether you two sincerely want to cooperate.

Compare with this:

*"Hi! I'm interested in this car. Tell me about it?"*

I think it's pretty easy to notice attempts to manipulate signals. If I were in this situation, I'd just keep sidestepping the signal manipulations and implicitly inviting (by example only!) the salesperson to meet me in clear honesty. If they can't or won't, then I'd probably decline to do business with them. I'd very likely be much more interested in living in this kind of clear integrity than I would be in the car!

(Or maybe I'd end up very confident I can see the truth despite the salesperson's distortions and feeling willing to take the risk. But that would be *in spite of* the

salesperson, and it sure wouldn't have been because I invited them into a signaling skirmish.)

---

This picture suggests that what others choose to signal just isn't any of your business.

If you focus on others' signals, you either Goodhart yourself or play into signaling arms races.

Far, far simpler and more reliable is just trusting reality to reflect truth. You just keep looking at reality.

This might sound abstract. For what it's worth, I think Jacob Falkovich might be saying the same thing in [his sequence on selfless dating](#). The trend where people optimize for "fuckability instead of fucking" and end up frustrated that they're not getting sex is an example of this. Goodhart drift engendered by focusing on the signals instead of on reality.

(My understanding of) Jacob's solution is also a specific example of the general case.

If you try to signal "Hey, I'm hot!" in the language you think will be attractive to the kind of person you think will be attracted to that signal...

...well, the sort of person you'll draw is the one who needs you to put effort into that kind of signal.

(Here I'm assuming for simplicity that the goal is a long-term relationship.)

So now, every ounce of energy you put into sending that signal falls into one of two buckets:

- It reflects reality, meaning you effortlessly would send that signal just by being transparently yourself. So the energy put into sending the signal is simply wasted and possibly anti-helpful (since it encourages you to mask the truth a little).
- It's a bit off from reality, meaning you have to keep hiding the parts of you that don't match what your new partner thinks of you. (In practice this is rarely sustainable.)

So the solution is...

\*drumroll\*

...drop all effort to signal!

Yes, you might end up not attracting anyone. But if so, *that is a correct reflection of you relative to the dating market*. To do better you'd have to trick a potential partner (and possibly yourself).

Of course, maybe you'd rather be in a relationship made of signaling illusions than be alone.

That's up to you.

I'm just pointing out a principle.

---

---

What exactly does it mean to "drop all effort to signal"?

Honestly, I'm not sure.

I have a very clear intuition of it. I can feel it. I can notice cases where it happens and where it's not happening, and I can often mentally transform one into the other. I know a bunch of the inner work needed to *do* it.

But I don't know how to define it.

Hence the epistemic status of "fuzzy conjecture".

My hope is that this brings some thoughtfulness to discourse about "social signaling" and "social status" and all that. I keep seeing Goodhart drift in those areas due to missing this vision. Hopefully this will bring a little more awareness to those corners of discussion.

It's also something I'm working on embodying. This ties clearly to how much care and thoughtfulness goes into communication: "Oh dear, what will people think of this?" That seems like it can be helpful for making communication clearer — but it also acts as bait for Goodhart's Demon.

I don't know how to resolve that just yet.

I hope I will soon.

# Question Gravity

Earlier today, I was in the shower. I had just finished a bottle of shampoo, and I tried balancing it on the thin railing of the shower door. I expected it to fall down, but instead it stayed there. From the side, it certainly *looked* like it was sticking out more than halfway over the edge. ([Picture](#)) Yet it stubbornly stayed put.

On a whim, I filled it with water and put it back. It fell over.

Wait, what? Filling it up shouldn't change the location of its center of gravity, so why would that make it fall over? I started performing more experiments. I rotated the bottle and balanced it again several times to see if there was some asymmetry causing it to have an off-center center of gravity. I thought maybe the water sloshing around could cause its center of gravity to move over the edge, so I tried holding it there for several seconds before letting go, in order to let the water settle. I noticed the door had some wiggle room, and the weight of the bottle could be causing the door to shift to a different angle, so I tried holding it at different angles. Nothing caused any different behavior. I put it on the ledge empty and slowly filled it with water. When it got to around 1/20 full, it fell over. I found a second bottle and performed all the same tests on it; same results.

At this point I was thoroughly [confused](#). I figured some assumption of mine must have been wrong, so I tried inspecting them. Does the center of gravity of an object *really* have to be over the edge in order for it to fall? I couldn't think of any situation where that wasn't true. (Generalizing from a non-exhaustive list of arbitrary examples is obviously not a great approach, but I couldn't do much better while in the shower.) Does filling it with water *really* not change its center of gravity? Oh wait; it does! The center of gravity doesn't move it laterally, but it does move it vertically. That assumption was wrong. (I probably wouldn't have made this mistake if I had been thinking of it as the center of mass rather than the center of gravity. Framing is important.) Unfortunately, that epiphany didn't lead to anything useful; I couldn't see how moving the center of mass vertically could affect whether it was balanced or not.

If you haven't figured it out yet, take a moment to think about it. You have all the information you need. Also, let me know in a comment if and how easily you figured it out before reading the answer below. I'm curious how well I stack up.

Answer: Cohesion between the water on the bottle and the water on the glass of the shower door.

This came to me after I noticed that when the bottle was almost, but not entirely, empty, it would seem to hesitate for a moment before falling. I didn't come to this immediately though. While I did notice the delay at the time and mentally flag it as "weird", I failed to consider how it might fit into the picture, and unconsciously dismissed it as irrelevant without giving it proper consideration. Only later did I have an independent realization that the bottle might be sticking to the door. That realization came from my unconscious mind; I didn't arrive at it through a deliberative process. Once I'd had it, it was trivial to grab a towel and confirm that when dry, the bottle would fall over even when empty.

There were several failures here, but the most important one was not going back far enough in the chain of assumptions I was making. I realized I needed to go through the process "Question Assumptions", but ended up only inspecting the irrelevant ones. I was unable to notice the assumption that actually needed questioning; that gravity was the only relevant force at play.

Properly questioning assumptions; not just going through the motions, but actually locating and inspecting all the foundational pillars of your thought process, is *really hard*.

# What happened when I applied metrics to my piano practice (a five-part essay about the process of learning)

*Crossposted from [Nicole Dieker Dot Com](#). The third and final part in [my piano practice series](#), as promised.*

If you follow me on Twitter —

and I'm not at all sure if you should, since Twitter seems to be less valuable to me every successive day — which in turn makes me less interested in [adding value to it](#) —

but anyway, if you read the tweets, you might have seen this one:

[Nicole Dieker @HelloTheFuture](#)

[I'm going to write more about this tomorrow, but last night I told L that I had discovered "the secret to learning" and he told me that I was correct, except for the part where I was the first person to discover it](#)

[January 26th 2022](#)

[2 Likes](#)

Over the past month I have, in fact, *discovered the secret to learning* — and even if I am not the first person to make this particular discovery, it still counts.

Now I have the somewhat difficult task of telling you what it is.

---

Here's how I explained it to L:

1. Define win condition.
2. Define action you are going to take to achieve win condition.
3. Take defined action.
4. Evaluate action both against its original definition (that is, *did you do what you said you were going to do or did you do something else*) and against the win condition.
  1. If you're me, *write down the results*. If you're L, keep them in your head. (He keeps all of this in his head. I have no idea how his head can handle it. He told me that he might have more storage space in his working memory because he thinks of things in bits and symbols instead of words.)
5. Ask yourself what is keeping you from achieving your win condition. Describe it as specifically as possible.
  1. If you're me, *write it down*.
6. Define action you are going to take to solve/address/eliminate obstacle preventing win condition.
7. Repeat 2-7 until win condition is achieved.
8. STOP.

---

Step #8 — STOP AFTER WIN — is more important to the process than I originally realized. At first I assumed that once you hit WIN you could REPEAT WIN, maybe REPEAT WIN 5 CONSECUTIVE PASSES, but it doesn't work that way.

Once your brain hits WIN, it's done with *that particular problem for that particular practice session*. Successive passes during the same practice session are more likely to be unfocused fails, which introduce inconsistencies that have to be resolved by running additional (time-consuming, frustrating) learning loops.

Plus, STOPPING AFTER WIN sets up a work environment in which you are CONTINUALLY RETURNING TO WON SEQUENCES.

I feel like I just confused you, so let me rephrase it:

- If you practice a specific piano passage until WIN and then put it away, the next time you return to that passage you'll approach it as SOMETHING ALREADY WON — that is, with CERTAINTY.
- If you practice a specific piano passage until WIN and then immediately play it again, you run the risk of NOT WINNING, generally because YOU AREN'T PAYING QUITE AS MUCH ATTENTION THIS TIME.

Introducing NOT WIN (or FAIL) immediately after WIN creates UNCERTAINTY.

I have nearly a month of spreadsheet data proving that it's better to STOP AFTER WIN.

If you don't believe me, believe [Chessable](#) — because I stole the idea from them, even though they also admit that [they were not the first people to discover it](#).

If you still don't believe me, believe Super Mario Bros. When you fail, you start over; when you win, you move on. 🐸⭐️🏰

This whole thing got started after I wrote my Substack advice post about [how to break out of the content mills](#). I was talking to L about the process of freelance writing, and how I had discovered very early on that there were two metrics associated with financial success: *words per hour* and *income per word*.

"I am astonished that most of the writers I work with don't know how many words they write per hour and aren't actively working to increase that number," I said. "Then again, I guess I don't know how many measures of music I can learn per hour — or how to increase that number."

That night, I created a spreadsheet to track those two metrics.

The next morning, I realized that before I could track *measures learned per hour*, I had to define *learned*.

And that got us to where we are today — or where we were two nights ago, when I told L that I had discovered the secret to learning.

	A	B	C	D	E	F	G	H	I	J	K	L
	Date	Day	Piece	Time start	Time end	Duration	Start	End	Action	Reps	Result	Next action
169									play to win	10	1 false start b/c after second LH C in 31, 2nd time, however the problem was actually in the LH descending line in 30, ran 30 several times began 25 and continued; 1 false start b/c used pedal when not used to, 1 false start confusion, 30 still not secure, worked 30 several times not knowing what or why...	20 m, 30 m, 30 m, 30 m keep where the fingers should go. I need to play 25-33 very slowly.
170	January 27	Thursday	Bach Ricercar a 6	9:49	9:57	0:07	1	78				
171				9:58	10:00	0:02	25	33	play very slowly to solidify	4	won the 4th time, by my rules I should stop...	play all
172				10:01	10:04	0:02	1	78	play to win	1	just a bit of hesitancy on the RH C in 30 but won, best pass ever until played Dsf instead of Ebg in 70, remember it is just a step downwards	play 65-78 to win
173				10:05	10:09	0:04	65	78	play to win, remember you go from AbF to GbE in 70	5	1 start passed Bbd in 67, 1 false start confusion, 1 pass won but very hesitant, 1 false start who prep, I start forgot D in 70, 1 pass won but very hesitant, 1 false start who prep, I start forgot D in 70, 1 pass won but very hesitant through 70 but landed it, focus lapse afterwards and sped up until end	refocus, play to win
174							0:00	65	78		win AND well-played	
175			La Valse	10:12	10:17	0:05	1	146	play to win	4	1 false start b/c anticipated confusion (and got it), 1 pass restoring memory, note evenness, play with metronome	
176				10:18	10:25	0:07	1	146	work DH50	10	started 98, failed 100 to RH c# not land Dir, worked 106, restarted 98, won everything from 98 and 129 except 108 and 122, worked 109 and 122, played 95-129, hesitant at 122 but won until played B instead of Bb in 129	play without metronome, slowly
177				10:25	10:30	0:04	98	129	play slowly, evenly to solidify	5	1 false start b/c no prep, 1 pass good until failed 122, go slower, 1 false start confused, refocus, 1 pass very good but lost focus in 124 after winning 122-123, 1 pass win	play all to win
178				10:30	10:33	0:02	1	146	play to win, how slowly does it need to go?	1	am counting as win, hesitant in 122-123 but did not fail, very much want to play with L, tonight	play with L, tonight
179			Stravinsky 5 Fingers	10:34	10:37	0:03	1	242	play to win	1	started out beautifully, missed grace note in 65 because unfocused, started speeding up, fell apart b/c played #112 instead of 113 in Larghetto	refocus, pw, record
180				10:38	10:45	0:07	1	242	play to win, record	1	very very very good best ever except something happened in the Moderato (a whistle?) and it distracted me	learn how to play through distractions

I would tell L, night after night: *I'm figuring things out, but I don't have all of the pieces yet. Some of the data I've been collecting has turned out to be completely useless. Other metrics are way more valuable than I thought they'd be. I'm not going to tell you any more than that until I'm able to tell you the entire thing, the whole theory, all at once.*

And he would say *But I can hear you working. You're so much better at solving problems than you were a few weeks ago.*

And I would say *Yes, but I haven't found all of the pieces I need to make the process efficient and effective and replicable — and I'm not ready to publish yet.*

Now I am.

And now you know everything I know about how to learn. ❤️

# Omicron Post #14

Important note on testing that I'm including at top of both posts today.

**From various sources, I have become convinced that rapid tests taken from nose swabs are likely to often be several days slower at detecting infections than rapid tests that use throat swabs.**

[Here's Washington Post on that](#), this quote should tell you what you need to know.

On one side are well-regarded experts who argue that swabbing your throat in addition to your nose may increase the chances that a home test can detect omicron. On the other side are similarly well-regarded experts, including those at the Food and Drug Administration, who urge Americans to avoid experimenting with kits that were developed and tested using only nasal swabs.

This 'disagreement' over 'experimentation' made me that much more confident in the conclusion I'd already reached.

That is echoed by a lot of sources both personal and general, [including in this heartfelt report of Bob Watcher](#), whose son has Covid-19 and reports back on the experience, [including five days later](#) follow-up where he's still positive and acts confused about what to do because CDC guidelines are confusing a situation that is otherwise extremely *not* confusing.

Thus, if you want to *know if you have infectious Covid-19*, you need to use throat swabs. It's still not 100% sensitivity or anything, but it's a lot better, whereas nose swabs are useful but in no way all that reassuring.

If you want to *satisfy a testing requirement* that is of course another matter. These are very different goals.

This is a whole week's roundup, because I got an opportunity to impact developments more directly and needed to jump on that with my non-day-job time instead for a few days. I hope to be able to talk about it soon, but can't right now.

## Travel

It's a very good question. [Why the hell are we still doing this?](#)

The Centers for Disease Control and Prevention on Monday advised Americans to avoid travel to Canada, citing "very high" levels of the coronavirus.

Canada was placed under a Level 4 travel health notice — the highest category... "Because of the current situation in Canada, even fully vaccinated travelers may be at risk for getting and spreading Covid-19 variants," the C.D.C. said.

The United States has higher rates of Covid-19 than Canada right now, yet we do this thing. I simply can't take seriously, as someone who might want a physically better and healthier world, anyone who restricts travel in such ways.

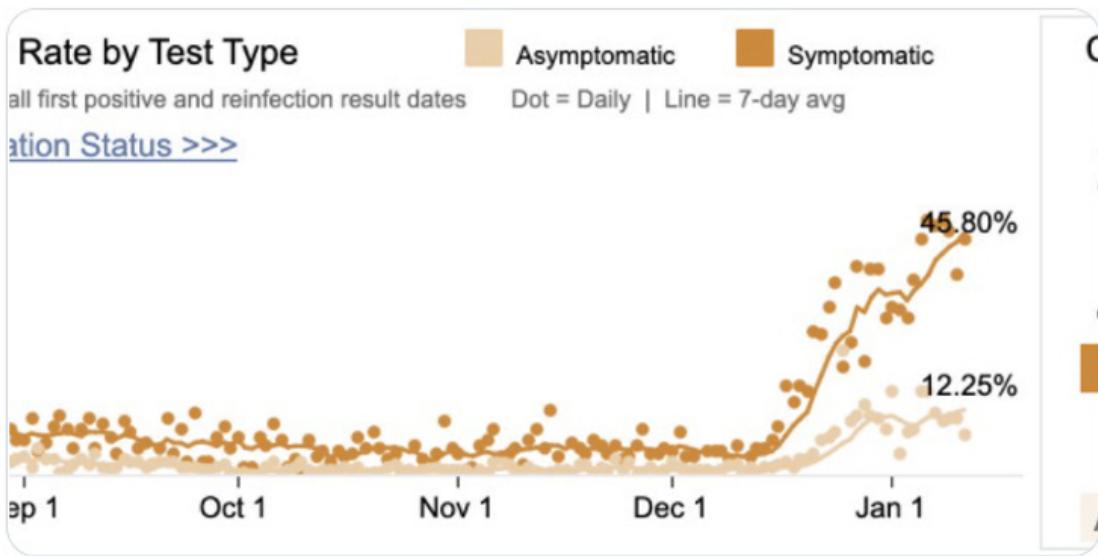
[The CDC resumed briefings this week.](#) This is the kind of thing they think we need to know.

## Spread

[CDC Nowcast has Omicron at 98.3%, which seems reasonable to me.](#)

[New update from Bob Wachter on San Francisco.](#)

Two big notes are that there's about an even mix of vaccinated and unvaccinated Covid patients outside of the ICU (this is SF, so a large majority of people are vaccinated) and then 11 of 12 in the ICU are unvaccinated. And then there's the background rate, which is pretty astounding.



[Trevor Bedford estimates that about one third of London infections are being detected.](#) I continue to think that is low, especially given the peak already happened, but it's not outside the range of possible answers.

[Anecdata from Michael Mina.](#)



Michael Mina ✅ @michaelmina\_lab · Jan 6

...

My family is spread across the US - In past few weeks, every single house has had multiple people w COVID

It's a bumpy road and we're taking a particularly rough path with a lot of casualties

But we are building up the layers of immunity to grow out of this virus' grip on us.

49

51

524



A key question for modeling how things go is whether people are modifying their behavior. Some people definitely are, but others definitely aren't, including *when they know they have Omicron*. One commenter noted that their friends treated it like a cold and mostly ignored it (as opposed to what one should do with a cold, which is *stay home*) [and then there's this](#).



[Thread defending the UK's case declines as genuine. I agree, there's no reason to think the declines aren't real.](#)

[A thread from 9 January on progress in Scotland and rest of UK](#), pointing out things have peaked but don't seem to be declining rapidly yet. Things remaining leveled off without a rapid decline is of course possible, but stabilizing *this high* can't hold for long because it's a lot of newly immune people every day. The crash is coming.

[New York City is definitely at least a few days after peak.](#)

## Hospitals

[Hobson's UK hospital thread for 8 January](#). I interpret a lack of a more recent thread as moderately good news.

[New York City 'with Covid' percentage of hospitalized Covid patients much higher than that of the rest of the state.](#)

Region	COVID-19 Patients currently hospitalized	Admitted due to COVID or complications of COVID	% Admitted due to COVID or complications of COVID	Admitted where COVID was not included as one of the reasons for admission	% Admitted where COVID was not included as one of the reasons for admission
Capital Region	361	277	77%	84	23%
Central New York	262	210	80%	52	20%
Finger Lakes	603	401	67%	202	33%
Long Island	2,060	1,273	62%	787	38%
Mid-Hudson	1,231	793	64%	438	36%
Mohawk Valley	133	88	66%	45	34%
New York City	6,052	2,992	49%	3,060	51%
North Country	95	64	67%	31	33%
Southern Tier	214	131	61%	83	39%
Western New York	537	391	73%	146	27%
<b>Statewide</b>	<b>11,548</b>	<b>6,620</b>	<b>57%</b>	<b>4,928</b>	<b>43%</b>

There was speculation that this was differences in definitions of 'with Covid' but I do not believe that is the case. I think the difference is real, and reflects population differences.

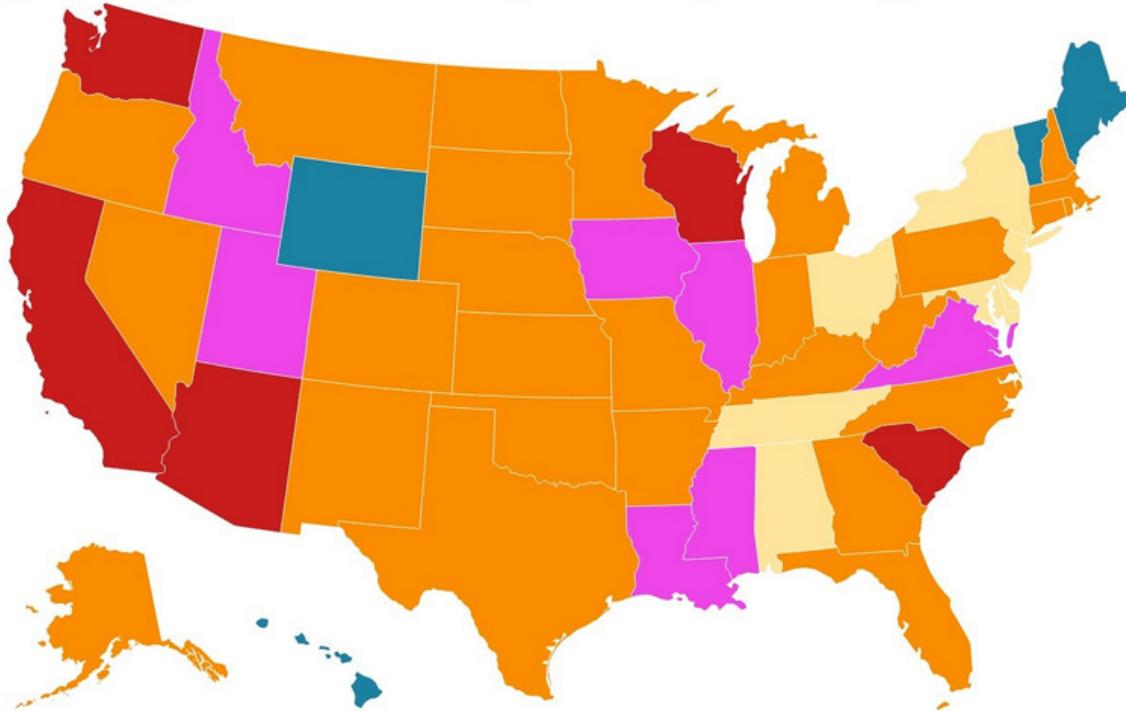
[There's also some concern this could be within-system spread](#) and I don't doubt that some of that is happening but I don't think it's primary.

[This map does not look great](#), but is probably selected to look scary and is also nowhere near as scary as I would have guessed a few weeks ago.

## COVID-19 Hospital Capacity Circuit Breaker Dashboard – State

View the circuit breaker dashboard at [alexanderxchen.github.io/circuitbreaker](https://alexanderxchen.github.io/circuitbreaker). Data as of January 11, 2022.

■ At Capacity ■ Forecasted to Exceed Capacity ■ Unsustainable ■ Continued Risk ■ Has Capacity



*At Capacity - Hospitals are projected to be at or above 100% capacity given a Nowcast with recent COVID-19 case trends (assuming non-COVID care is stable); Forecasted to Exceed Capacity - Given recent COVID-19 case trends, states are at high risk of exceeding 100% hospital capacity in the next 1-7 days; Unsustainable - The single day case count yesterday is on a trend to the area reaching a circuit breaker; Continued Risk - Area has been at or forecasted to exceed capacity in the past 7 days; Has capacity - Hospital capacity not at high risk of being exceeded*

Map: Jeremy Faust, Bill Hanage, Benjy Renton • Created with Datawrapper

The Northeast has likely peaked in terms of cases, so the ‘continued risk’ column I think largely means things are not at risk. There isn’t that much population in the unsustainable areas here, which makes me think that it isn’t unsustainable by that much, and I expect peaks not too long from now. I am very much not making light of the situation, it’s going to suck quite a lot, but that was always the case.

[Concretely here's a report that Oregon is at 96% of capacity and that's going up at 4%/week.](#) Given the delay on hospitalizations after cases, will they exceed capacity? Probably, that curve would likely peak somewhere around 102%, but given the number of exponentials lying around this could have been so much worse.

Story continues to be largely about [Jack of staff causing the problem, rather than lack of beds.](#)

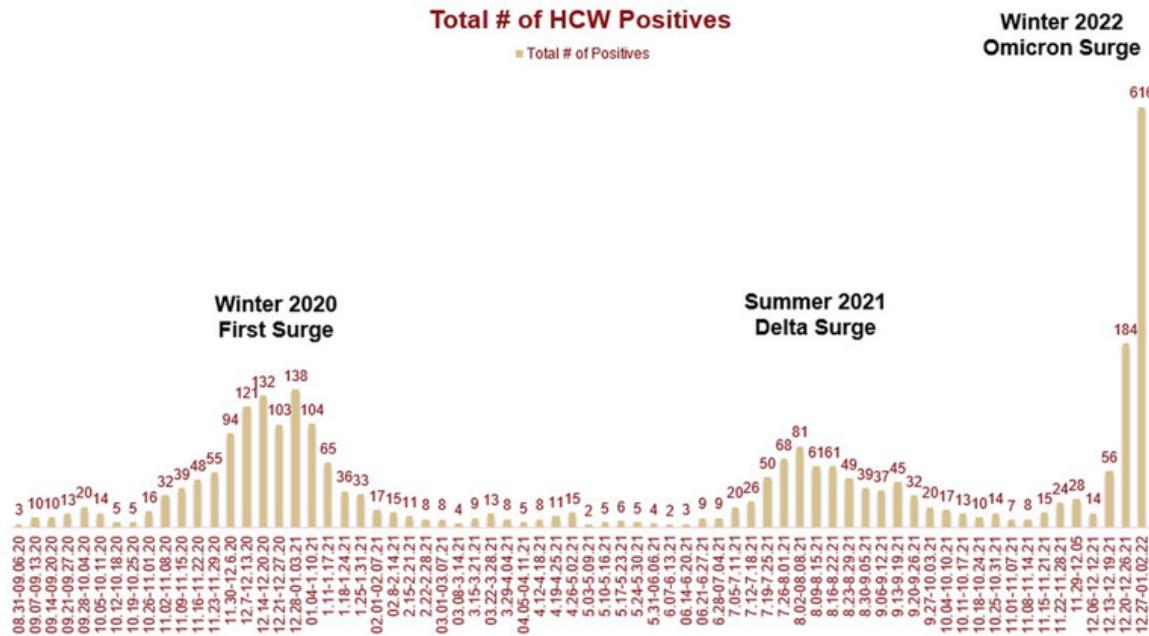


Abraar Karan ✅ @AbraarKaran · Jan 6

...

From a public email sent to all today—and I can personally attest other big academic hospitals are going through the same. Mandate respirator use for staff at bare minimum for better protection & better source control in hospital. Testing capacity is overwhelmed.

## Workforce COVID Impact





Mike Bird ✅ @Birdyword · 9h

Big study of ~70k Covid cases showing Omicron's far lower severity of outcomes vs Delta

...

Symptomatic hospitalisation -53%

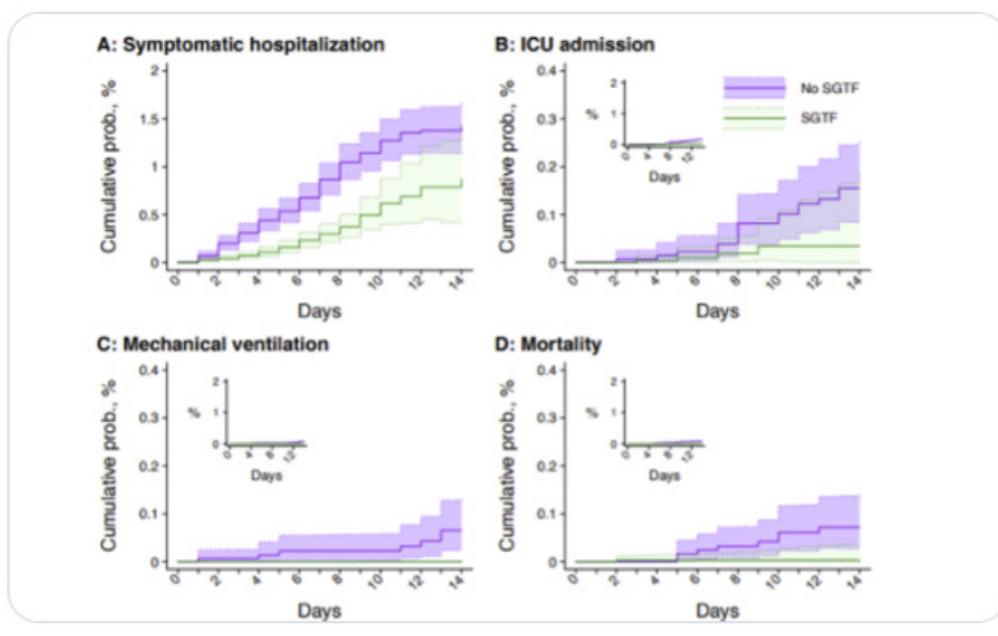
Median hospital stay length -70%

ICU admission -74%

Mortality -91%

Major effects among both vaccinated and unvaccinated patients.

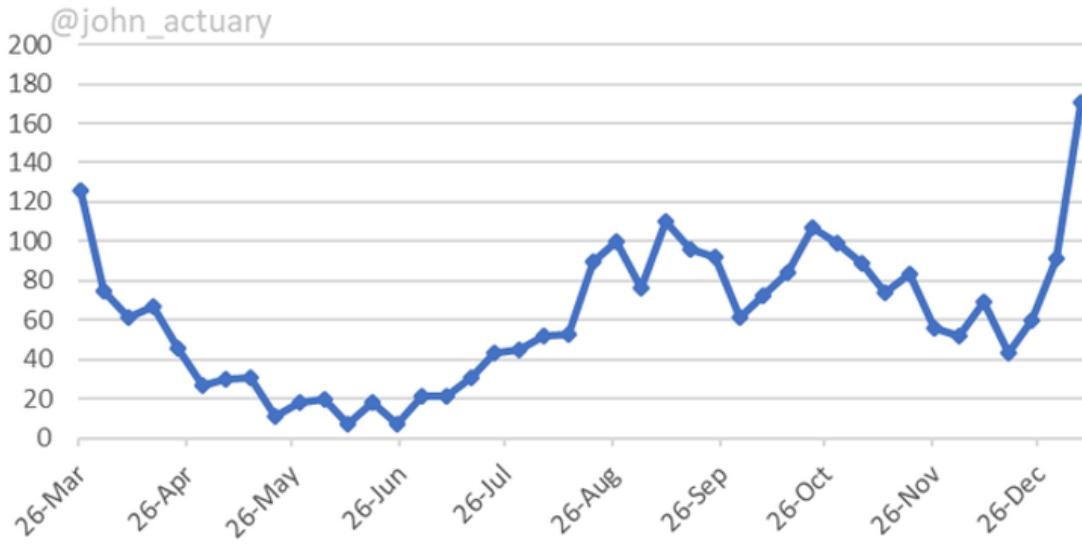
[medrxiv.org/content/10.110...](https://medrxiv.org/content/10.1101...)



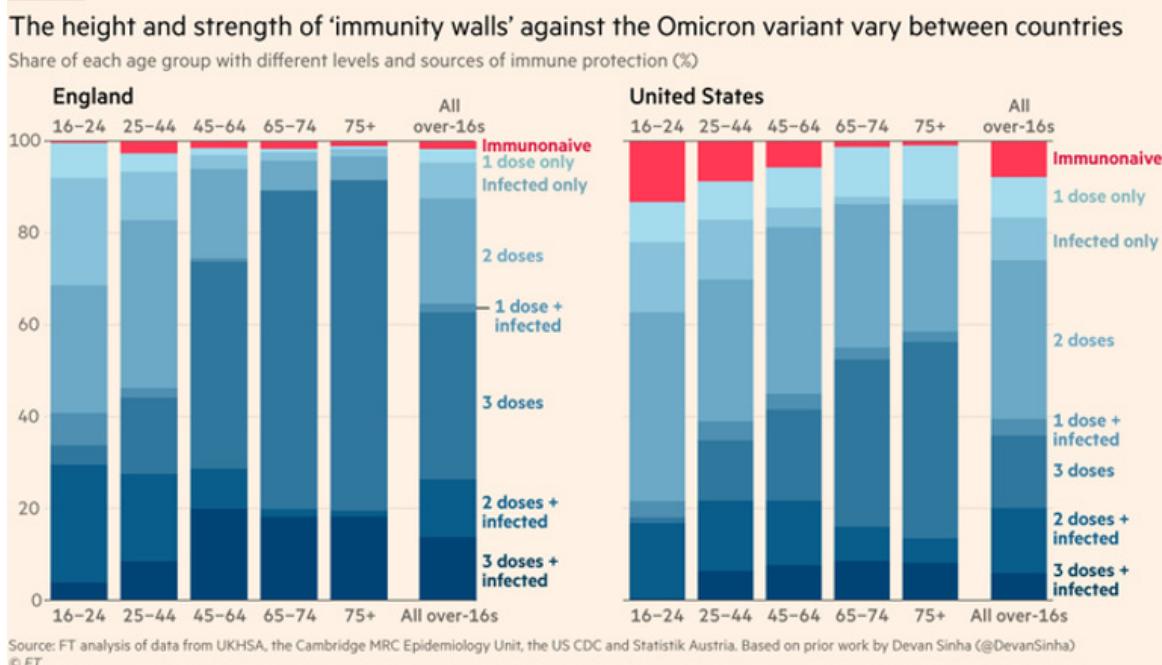
[Nursing home Covid deaths up in the UK](#). Not good news, but once again less bad news than one might have feared, and cases are already turning around.

## England Care Home Deaths Involving COVID-19

(as notified to CQC, includes those dying in hospital)



[USA really does have a lot more immunologically naïve people](#) than the UK, so we're seeing more impact in the hospitals than other places.



[Or we can look at real world data from a few other places.](#)



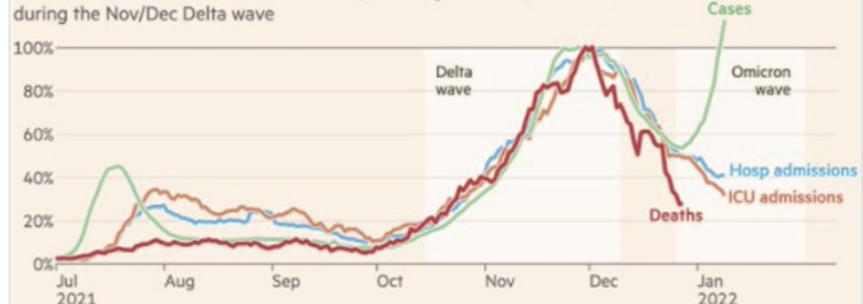
John Burn-Murdoch ✅ @jburnmurdoch · Jan 11

The Netherlands is one of the most interesting places for tracking the divergence of cases from more acute outcomes as Omicron takes over

Timing of its winter Delta wave meant numbers rose \*and fell\* before Omicron took off, so any rise now can reasonably be attributed to Omi

### Cases seem to be diverging from more acute outcomes as the Netherlands swaps Delta for Omicron

Covid-19 metrics in the Netherlands as a percentage of their peak value during the Nov/Dec Delta wave



Admissions and deaths shifted backward to adjust for lag

Source: Dutch National Institute for Public Health and the Environment. Based on prior work by Paul Mainwood

© FT

116

1K

2.9K

↑

Show this thread

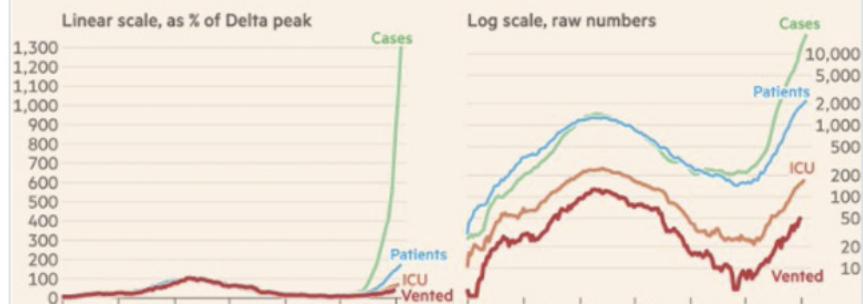


John Burn-Murdoch ✅ @jburnmurdoch · Jan 11

New South Wales (Australia) similar to Netherlands:

- All metrics rose then fell in earlier Delta wave
- All have since risen with Omicron (i.e yes Omi can still cause severe disease among those with little or no immunity)
- But acute outcomes completely diverged from cases

### The number of patients with severe disease in New South Wales is rising, but the ratio of severe disease to cases has weakened considerably



Source: covidlive.com.au. Hospital series shifted backward to adjust for lag

FT graphic by John Burn-Murdoch / @jburnmurdoch

© FT



John Burn-Murdoch @burnmurdoch · Jan 11

...

Replying to [@burnmurdoch](#)

This is in contrast to places like the US where Omicron is piggy-backing on top of a fierce winter Delta wave, making it tricky to know whether it's Omicron or Delta that's sending hospitalisations ever higher



18

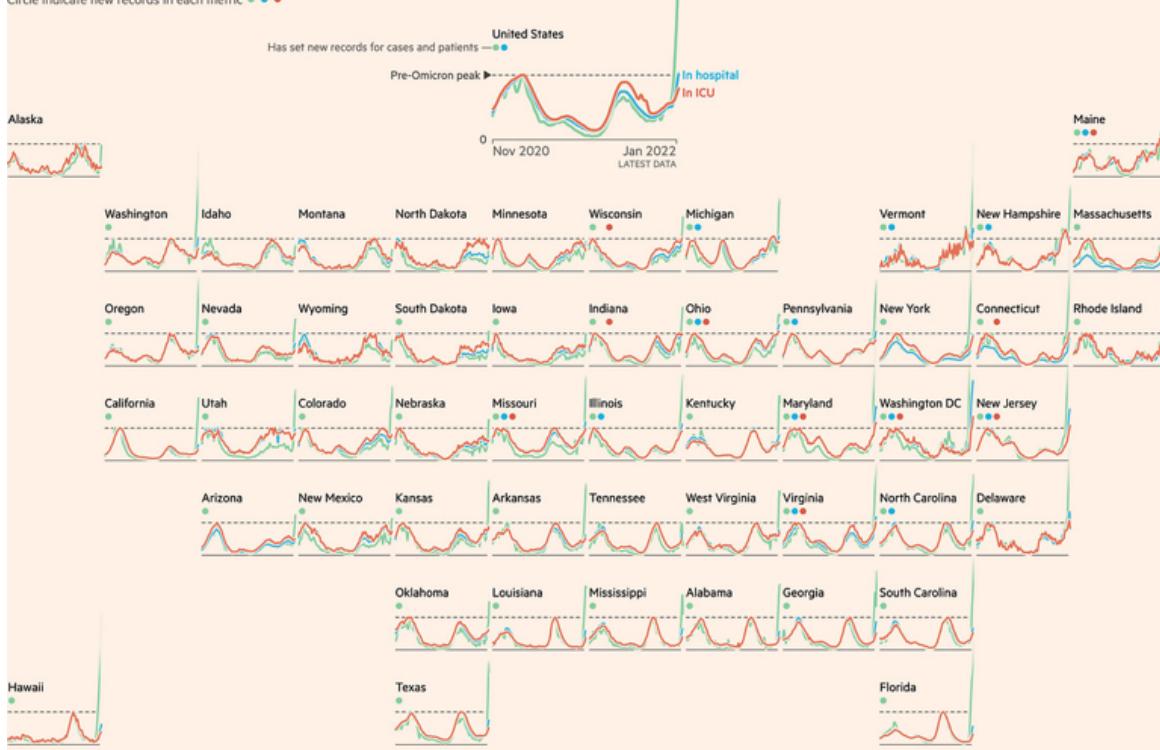
118

470



#### In many US states, ICU patient numbers are already approaching record levels as Omicron takes off

Covid cases, hospitalisations and ICU patient numbers, each relative to its pre-Omicron peak.  
Circle indicate new records in each metric ● ●



FT analysis of data from Johns Hopkins CSSE and US CDC. Cases shifted forward to adjust for lag between testing positive and developing severe disease  
FT graphic: John Burn-Murdoch / [@burnmurdoch](#)  
© FT

[That last one is hard to read](#) but it looks like divergence at least to me.

## Outlook

[Scott Gottlieb gives his prognosis on Face the Nation on 10 January](#). He thinks several states have already peaked, and the remaining biggest worry is in the midwest.

[Current head of the FDA says most of us are going to get Covid-19, in so many words](#). The hospitals are going to hold, or at least mostly hold. Make risk decisions accordingly.

[Would you pay a lot to have this be over, or not?](#)



**Professor Simone Schnall** @SimoneSchnall · Jan 8

...

Poll: Imagine it was possible to press a button and immediately go back to the way of life you had before the pandemic. How much would you be willing to pay for it?

Nothing	42%
20% of one month's income	5.9%
50% of one month's income	5.1%
One month's income <input checked="" type="radio"/>	47%

1,403 votes · 3 days left

The people who answer ‘nothing’ here are either interpreting the question differently, they have very strange preferences, or they’re doing the thing where people refuse to pay for things if they don’t feel like it’s fair to be asked to pay for them. Once people acknowledge that the Before Times had their advantages, most of them are willing to pay over a month’s income, and a more interesting poll would have seen how far that goes. If answering in the spirit of the question, I would pay quite a lot, and also almost certainly not pay the full amount it was worth, if the full way of life includes everyone around me going back to the old ways and there being no Covid-19 to catch.

[Thread from 8 January saying we might as well let it rip at this point, not like we have any choice.](#)

## China

Last time I asked others for opinions on the probability China could maintain its zero-Covid policy.

There are certainly signs things aren’t great. They confirmed Omicron’s presence, [and they mass-tested 14 million people in Tianjin](#). The Olympics are still planned. But every day that passes is by default good news, so it’s not clear this is worse news than the baseline.

Here’s a roundup of takes I got.

[First, an optimistic take at 70% chance that thinks China can’t fold and has the tools to win.](#)

I would give the Chinese much better odds of keeping case counts under 50/million: 70%. China's rulers are not lawyers and propagandists, agents of casuistry and illusion. They're engineers and businessmen who have taken a mostly practical approach to the pandemic. And since learning to a large extent happens by doing, presumably their containment capability has incrementally improved. They are also willing to do what is necessary in accordance with their reality model, including firing deficient officials, as just happened again in Xi'an. The advent of Omicron means that they need to tighten travel restrictions further, which they've done for the upcoming Olympics. Another option is to acquire mRNA vaccines for those people who are most exposed to travel related risks. That would provide another defense line.

Xi has staked his reputation on this policy, and he's also put the Party's reputation on the line. This is crucial to them both domestically and for foreign consumption, to demonstrate the superiority of the Chinese system relative to America's. Underlings who disappoint Xi are apt to face career destruction. The places that have seen massive Omicron outbreaks are also places that have let it rip. We haven't really seen anyone try to contain it yet. Vaccination created what now seems like justifiable complacency. With the methods China has developed, I don't think Omicron is that much more difficult to contain than previous variants. One reason I don't give even higher odds is that Xi may discover that Omicron has something like a 0.1% mortality rate in China, which might at some point of national stress cause him to rethink the cost-benefit calculation. If a resumption of normality also occurs in the West while China suffers lockdowns, Xi might decide Zero Covid will not win the PR war. In other words, it's easier for me to imagine them giving up than being defeated in battle.

[Here's a skeptical one, with a 25% chance, that thinks China can pivot and thus afford to fold.](#)

China below 50 million: sell to 25%. My view is that most people overestimate the cost of switching to living with Omicron, e.g. "the Party has staked a lot of political capital on keeping Covid out". That's a severe underestimate of their propaganda/opinion guidance ability. If they want to change, they will be able to do so without paying much of a price in terms of social stability.

Here's one that is far more optimistic than I that they can hide what is happening, I replied [I would happily buy 33s and sell 80s](#) (I'd definitely sell 80s either way, and given that I'd love to get the hedge too).

In autumn, Xi is looking to extend his term for another five years at a large party conference. Given the instability and uncertainty of what happens, if they give up on zero Covid, I can't fathom that he would decide to give up before this important event. Afterwards, with potentially another Covid variant, and mRNA vaccines they might actually do it. However, they might just not release accurate testing numbers in that case. In fact, I don't see why they would, it seems much more rational to just declare the pandemic to be over publicly ("we have never been at war with Covid"). Before that, China should have enough capacity to contain the virus. Omicron has a shorter generational interval, which in some ways makes it easier to contain. Compared to Wuhan 2020, China has the capacity and the will to test everybody with PCR tests, vaccinations, and much improved digital tracking. Even if Omicron were twice more transmissible than the original strain, these improvements should offset it.

China less than 50/million (officially): 80%

China less than 50/million (actual case counts, despite being unverifiable): 33%

[This is another take that there's a strong chance that Omicron comes and goes and China simply denies it.](#) I don't buy it. Either Omicron goes full blast or it is suppressed, and if it goes

full blast I don't see how they pretend it didn't. Buying time is one thing, this would be very different.

[A simple and skeptical take that makes sense.](#)

Prediction responses (no strong outside view reasons to think I'm right here):

China keeps cases contained: sell to 30%. Their pulling out all the stops has been barely enough before, I expect even China to have lockdown fatigue and I generally expect Chinese government action to be less competent than e.g. Korea or Singapore.

[This is a longer take that puts China's chances even short term at 10%. Here's one at 15%.](#)  
[This one says 20% if they don't cancel the Olympics.](#) Basically the argument is 'it's too hard, it can't be done.'

None of the arguments on either side that were new seemed that strong, but overall I'm somewhat more skeptical than I was last week that China can pull this off.

## Other and Threads

[Wall Street Journal Editorial boldly goes somewhere, proposing it would be safer to speed up the spread due to effects on future mutations, and, well, wow, just wow, that's a set of words they made me type, good show. Also inspired this rebuttal thread, if it isn't obvious to you why the proposal makes no sense.](#)

[Andy Slavitt thread from 7 January.](#) Still in the prevention mindset, but pointing out that the task is essentially hopeless.

[Thread from 7 January worrying about failure of USA hospitalization stats and case stats to diverge as much as European ones.](#)

## Probability Updates

**China keeps daily cases under 50 per million through 2022: 40% → 30%.**

Having another mass testing round so quickly is somewhat bad news, and the 'market' I surveyed was generally on the lower end although with disagreement. I'm updating somewhat lower and we'll go from here until we can get a real market up.

**Median peak cases day for United States: January 19 → January 18**

**Mode peak cases day for United States: January 19 → January 17**

**Daily cases in the United States have peaked by February 1: 90% → 93%.**

My timeline has moved up on average as we see peaks, but given the lack of an overall peak yet the distribution is somewhat asymmetrical. A peak in February would at this point be very surprising, given some areas have peaked already, but strange things do happen.

**Will There Be a Federal Mask Requirement on Domestic Flights as of November 8, 2022? 50% → 45%.**

**Chance that Omicron has a 100% or bigger transmission advantage in practice versus Delta: 55% → 45%.**

**Chance that Omicron is importantly (25%+ in the same person) less virulent than Delta: 95% → 97%.**

**Chance that Omicron is vastly (75%+ in the same person) less virulent than Delta: 30% (unchanged).**

The general vibe that things get to be over makes me more optimistic about lifting restrictions in the medium term. Based on doing a bunch of modeling this week I'm finding myself putting in transmission edges slightly under 100% more than above so I should update there as well. The verdict on less virulent seems to be fully in but guarding against model error a bit, and for the 75% number I expect never to know for sure.

**Chance we will be getting boosters modified for Omicron within 6 months of our previous booster shot: 20% (unchanged).**

**Chance we are broadly looking at a future crisis situation with widely overwhelmed American hospitals, new large American lockdowns and things like that: 3% → 2%.**

I was asked why my booster chance wasn't updating. Basically it's because I don't expect much news at this time. They're preparing the boosters as one would expect, but no idea whether they will push those boosters on the already boosted. As for the crisis scenario, it seems clear we won't get there, things won't justify a reaction and it would take a lot to provoke one. Again, holding back for model error, no one wants to end up in Bayes Hell.

**Generation time (serial interval) of Omicron is 3.5 days or less: 90% (unchanged).**

Modeling made it clear this is indeed in about this range for me right now. This is another 'we'll probably never know and I may never update here again' case.

# List of Probability Calibration Exercises

I recently spent a while looking at how different people had designed their probability calibration exercises (for ideas on how to design my own), and they turned out to be quite difficult to find. Many of the best ones were the least advertised and hardest to locate online. I figured I'd compile them all here in case anyone else finds themselves in a similar position. Please let me know about any I missed and I'll add them to the post. Many of these are old and no longer maintained, so no guarantees as to quality.

<https://bayes-up.web.app/>

<http://acritch.com/credence-game/>

<http://confidence.success-equation.com/>

<https://calibration-practice.neocities.org/>

<http://web.archive.org/web/20100529074053/http://www.acceleratingfuture.com/tom/?p=129>

<http://credencecalibration.com/>

[https://programs.clearerthinking.org/calibrate\\_your\\_judgment.html](https://programs.clearerthinking.org/calibrate_your_judgment.html)

<https://www.openphilanthropy.org/calibration> or <https://80000hours.org/calibration-training/> (Different URLs for same application.)

<https://calibration.lazdini.lv/>

<http://www.2pih.com/caltest/>

<http://web.archive.org/web/20161020032514/http://calibratedprobabilityassessment.org/>

[https://predictionbook.com/credence\\_games/try](https://predictionbook.com/credence_games/try)

<https://calibration-training.netlify.app/>

[https://play.google.com/store/apps/details?id=com.the\\_calibration\\_game](https://play.google.com/store/apps/details?id=com.the_calibration_game)

<https://www.metaculus.com/tutorials/>

<https://outsidetheeasyblog.blog/probability-calibration/>

<https://peterattiamd.com/confidence/>

<http://quantifiedintuitions.org/calibration>

# Covid 1/27/22: Let My People Go

[The moment's here.](#) My people are all the people. It is time to let my people go.

While case counts in many places remain high, we are on the way back down the mountain. The hospitals will hold. People can choose, based on their preferences and situation and the local conditions, whether they want to go now or wait a few more weeks before going. That is their call. It needs to be *their* call.

One could argue, as [Tyler Cowen did in this excellent talk this week](#) at Yale, that the moment is not *quite* here yet, on the theory that in a month cases will be an order of magnitude lower and thus it will be politically and socially easier to make the transition. There would be less opposition then, so better to wait, the price for doing so is small.

Would I take that deal? Absolutely I would take that deal, *if we agreed on an end date or on explicit end conditions*. A few more weeks is a small price in the grand scheme, and getting these things to happen takes time, so 'a few weeks from now' is the second best time to end pandemic restrictions.

[But there's no need for that.](#) The best time is right now.

Remember that the case counts are seven-day averages and there is a several-day delay between infection and positive test, so we are living continuously living, for better and for worse, 'in the future.'

Today I go back to the excellent [Da Umberto](#), to celebrate (barring another variant, and ignoring the writing of posts and the pro forma wearing of masks and showing vaccination cards) the end of my pandemic.

## Executive Summary

1. Cases now declining most places.
2. Restrictions mostly remain in place.
3. Let my people go.

Let's run the numbers.

## The Numbers

### Week-Over-Week Predictions

Prediction from last week: 4.4mm cases (-10%) and 14,500 deaths (+15%).

Results: 4.05mm cases (-17%) and 15,964 deaths (+26%).

Prediction for next week: 2.85mm cases (-30%) and 20,000 deaths (+25%).

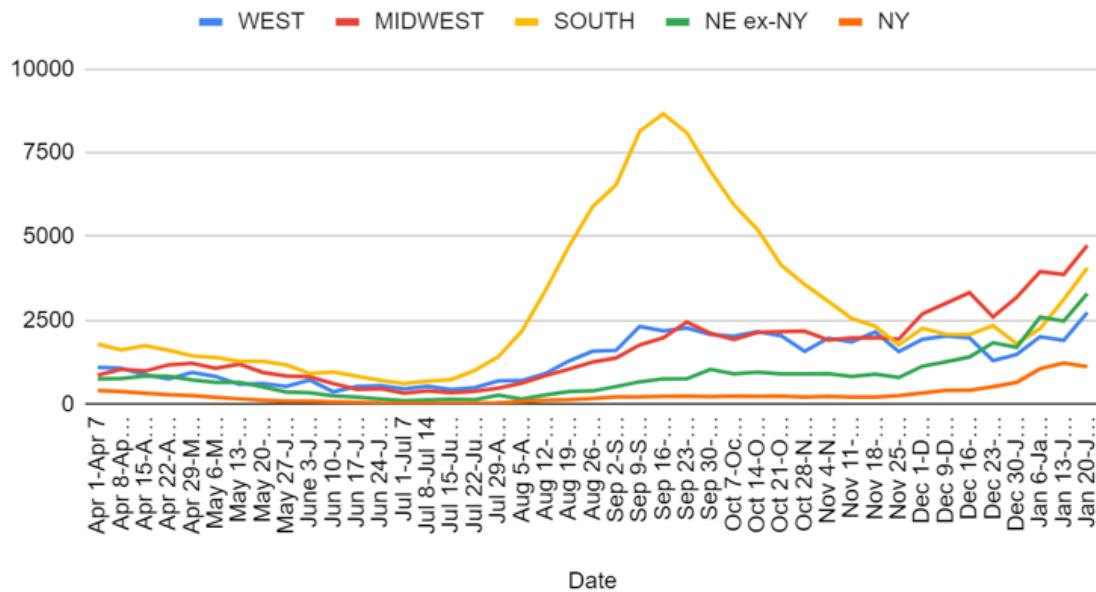
Overall it seems we peaked faster and more in unison than I expected, while other places did the opposite, and had softer declines than I expected. Based on those smaller declines elsewhere, I still expect the decline to accelerate as the West turns the corner, but I don't expect a crash as quick as the rise up. A much faster decline is possible, but it would make our observations elsewhere quite weird.

The most confusing and surprising place of all right now is India. Cases are somehow *flat* there, and have been for a week, without hitting levels where that much adjustment makes sense, and I will freely admit that I don't understand it at all, although I haven't had time to look in detail and there are no related news stories that naturally have entered my feeds.

I interpret the death count here as saying that last week's count was indeed artificially low rather than that this was a true 26% rise, and this in turn means this week's is a bit higher than its true level. Deaths should continue to rise due to their lag behind cases, and I wouldn't be that surprised by up to about a +40% number here, but after 1-2 more weeks after that we should see declines there as well.

## Deaths

Deaths by Region



Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Dec 1-Dec 8	1933	2693	2264	1465	8355
Dec 9-Dec 15	2036	3014	2088	1675	8813
Dec 16-Dec 22	1978	3327	2080	1818	9203
Dec 23-Dec 29	1302	2602	2342	2353	8599
Dec 30-Jan 5	1481	3185	1805	2343	8814
Jan 6-Jan 12	2013	3955	2258	3656	11882
Jan 13-Jan 19	1904	3873	3121	3708	12606
Jan 20-Jan 26	2734	4738	4070	4422	15964

Looking at the details makes it clear that last week had a more serious disruption to death reporting than we realized. Perhaps this week was thus slightly higher than its true level, on the order of a few percent. The large jumps across the board are mostly real, but that they weren't far worse puts a cap on how bad things are going to get.

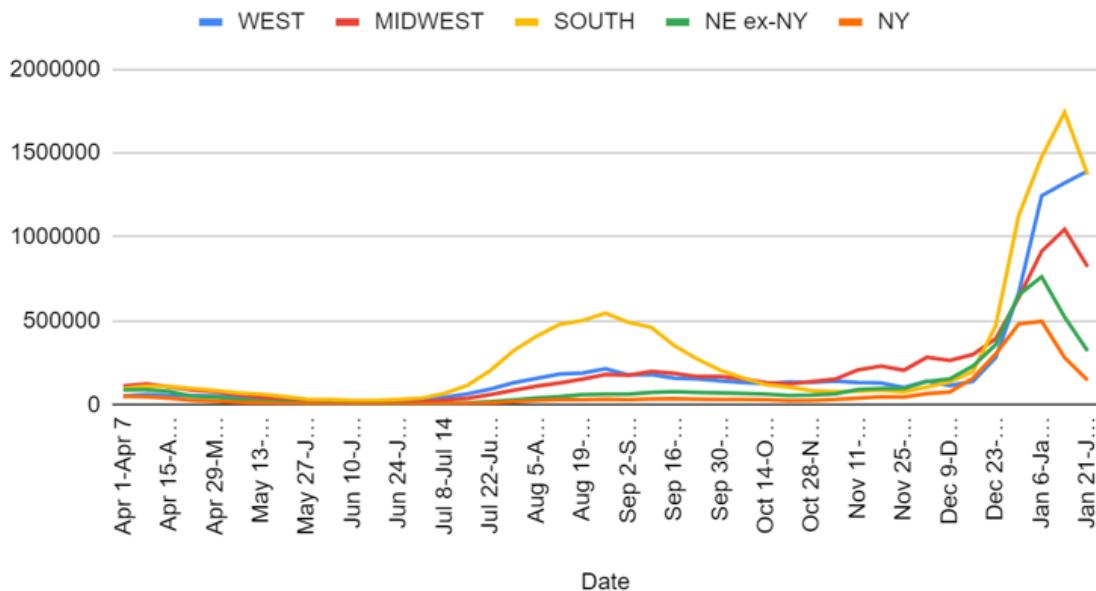
Worse plausible case, which still seems very unlikely, is something like +50% this coming week, +25% the week after, then +10% three weeks from now followed by a peak anyway, anything substantially worse than that would be very surprising. That's a lot of death, but nothing like what we feared as recently as a month ago.

## Cases

[Reminder from Bob Watcher](#) that our case averages are reports from a week in the past, so on the way up our case counts *underestimate* cases, and on the way down after the peak they reliably *overestimate* cases.

[He also notes that graphs seem much less like fast-up/fast-down than one would have predicted](#). The UK in particular has been bizarrely stable rather than continuing to decline. My best explanation for the slow decline is that different sub-regions peak at different times combined with behavioral modifications, but this is *not* sufficient to explain what we are seeing.

Positive Tests by Region



Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Dec 2-Dec 8	140,943	283,433	105,544	206,259	736,179
Dec 9-Dec 15	114,266	264,402	136,589	231,791	747,048
Dec 16-Dec 22	140,012	300,125	198,037	388,788	1,026,962
Dec 23-Dec 29	281,703	394,422	472,090	663,080	1,811,295
Dec 30-Jan 5	667,872	639,499	1,128,524	1,137,319	3,573,214
Jan 6-Jan 12	1,245,702	915,402	1,477,626	1,261,405	4,900,135
Jan 13-Jan 20	1,322,131	1,046,289	1,743,659	803,597	4,915,676
Jan 21-Jan 27	1,393,002	823,386	1,372,664	468,132	4,057,184

# Vaccines

[Latest studies are in \(study link, second study link\).](#)



**Michael Lin, MD PhD** @michaelzlin · Jan 23

...

New CDC studies have good news for the boosted, I mean up-to-date, bad news for double-jabbed only

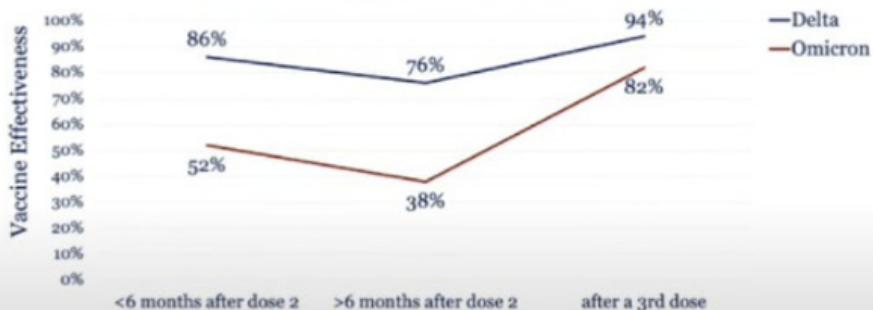
The good news: Protection from Omicron ER visits is 82% for the up-to-date

The bad news: Protection down to measly 38% for non-boosted 6mo after dose 2

[cdc.gov/mmwr/volumes/7...](http://cdc.gov/mmwr/volumes/7...)

## Vaccine effectiveness of 2 vs 3 doses of mRNA vaccines for Delta and Omicron

Vaccine Effectiveness against ER and Urgent Care Visits



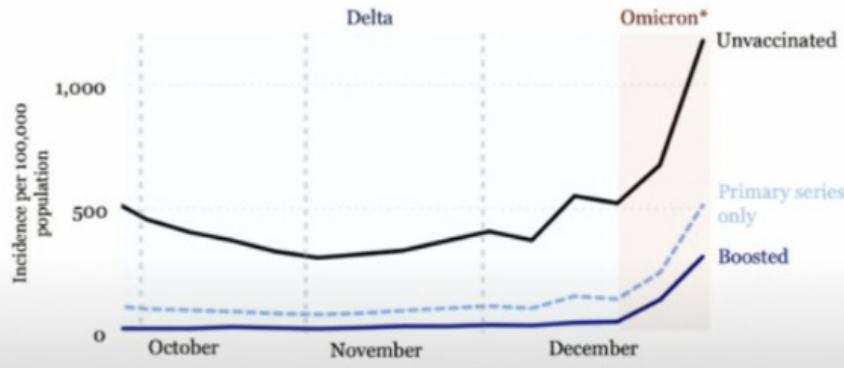
Source MMWR: <http://dx.doi.org/10.15585/mmwr.mm7104e3>



Michael Lin, MD PhD 🌈 @michaelzlin · Jan 23

Another CDC study from Friday looked at Omicron cases by vax status. The effect of the 3rd dose vs 2 doses is hard to say, considering all the triple-dosed had a recent shot while some of the double-dosed did not.  
[cdc.gov/mmwr/volumes/7...](https://www.cdc.gov/mmwr/volumes/7...)

### Rates of COVID-19 Cases by Vaccination Status and Booster Doses



Source: CDC COVID Data Tracker - Rates of COVID-19 Cases and Deaths by Vaccination Status

2

3

17

↑



Michael Lin, MD PhD 🌈 @michaelzlin · Jan 23

Nevertheless it looks like the 3rd dose is protecting against being infected by 75% vs unvaxxed. This is in line with initial reports from Pfizer. It also makes sense that the VE gets better for more severe levels of disease, i.e. hosp > ER visits > infection.

1

4

36

↑



Michael Lin, MD PhD 🌈 @michaelzlin · Jan 23

So overall confirms big benefits of the third shot in terms of avoiding the ER or urgent care, i.e. serious disease, and of fresh vaccines for avoiding hospitalization, i.e. severe disease.

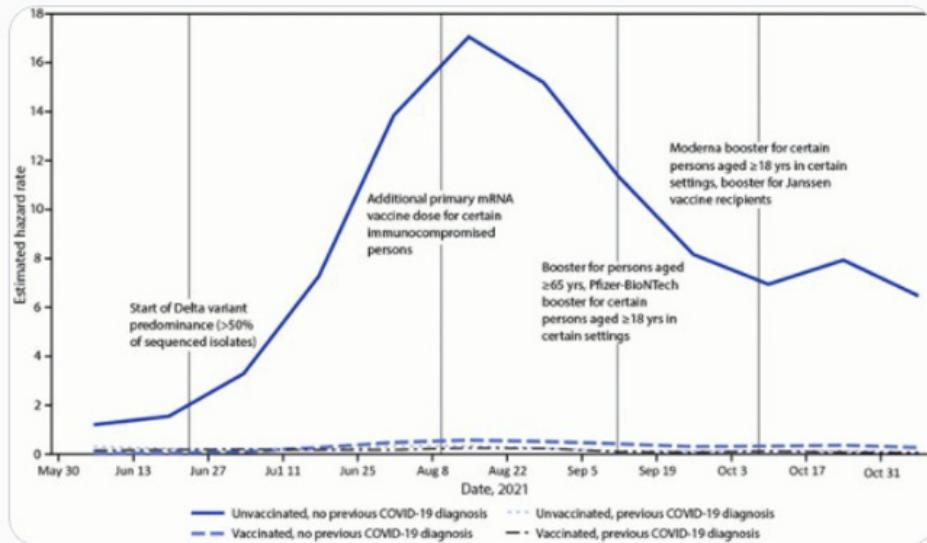
Then comes the other part people are talking about.



Michael Lin, MD PhD ⚡ @michaelzlin · Jan 23

Now I'll wade into somewhat more controversial waters and discuss the other interesting study this week, which found Wuhan strain infection protected against being a Delta case better than vaccination (also Wuhan strain) did.

[cdc.gov/mmwr/volumes/7...](https://www.cdc.gov/mmwr/volumes/7...)



1

6

22

↑



Michael Lin, MD PhD ⚡ @michaelzlin · Jan 23

That's clearly not an ideal result. We'd of course want our vaccines to be as protective as the viral infection, if possible. So what does it mean?

1

1

16

↑



Michael Lin, MD PhD ⚡ @michaelzlin · Jan 23

The endpoint being assessed, protection against Delta, requires mucosal antibodies produced against original SARS-CoV2 to be broad and high enough during Delta exposure to prevent the infection. That the vaccines did worse suggests their antibodies are less broad or wane faster.

5

6

26

↑



Michael Lin, MD PhD ⚡ @michaelzlin · Jan 23

We know from the [@BalazsLab](#) study mentioned above, and linked below, that viral infection isn't better than vaccines in broadening an antibody repertoire, so I don't think it's that. So rather it seems that mucosal immunity after 2 vax shots decays faster.

'Fades faster than natural infection' isn't obviously that bad, since it's a kind of arbitrary baseline, although obviously we'd prefer if it wasn't true.



**Michael Lin, MD PhD** @michaelzlin · Jan 24

...

If dose 3 doesn't help — we'll know in a few more months — then we either need to develop better vax, or we'll have to boost every 6mo instead of 12mo (impractical), or we'll have to resign to a high level of endemicity until everyone has been infected (worrisome for the elderly)

4

6

41



**Michael Lin, MD PhD** @michaelzlin · Jan 24

...

So where can a better vaccine come from?

Here's an idea: The J&J vax may be a good intranasal vaccine!

I've been the first to say J&J is not as good as RNA vax \*intramuscular\*. BUT it's in adenovirus — which naturally infects the nose.

J&J: You can thank me later, if it works.

11

10

122



**Michael Lin, MD PhD** @michaelzlin · Jan 24

...

BTW the unvaxed groups in the papers above are expected to have some people with prior infection too (maybe 30% by December). If compared to unvaxed uninfected alone, vaccine protection would be higher in all vaccine groups.

2

1

12



**Michael Lin, MD PhD** @michaelzlin · 1h

...

Turns out Israel has had enough people months out from their third dose to know that antibody levels again drop after a few months.

2

3

12



**Michael Lin, MD PhD** @michaelzlin · 1h

...

They are now recommending the 4th dose for everyone 5mo out from their 3rd dose. This doesn't apply to most people in the US yet.

It's really not a viable path to be boosting everyone every 5mo. We need better vaccines.

I strongly agree that boosters every five months is not a strategy. [And yet, some people think this, like hope, is indeed a strategy.](#)



Kendric Tonn @kendrichtonn · Jan 21

...

Last vax shot I had cost me two full functional days, which anecdotally doesn't seem uncommon. 1.1% of your life here, 1.1% of your life there, and, y'know, it's something.



no! no!! 🌈 ✊ @notkavi · Jan 20

Show this thread



No, really, asking for a booster every six months that knocks a lot of people on their ass for a day or two? Yeah, that's kind of a huge ask. The short term side effects are sufficiently costly that most people will *correctly* decide that it's not worth it even if there's lots of Covid out there. If this is the world we live in, the booster will be eventual mild reinfection. Those terms will need to be acceptable, because 'better vaccines' seem highly unlikely. These were pretty awesome vaccines.

Quebec is mandating that if an unvaccinated shopper is going into a big-box store, an employee must follow them around at all times to ensure they don't sneak off and buy something that is not food or medicine. We would not want such folks exchanging money for goods, that's how you get the Covid, you see. Having someone *follow around and stay in close proximity to* the unvaccinated person seems like the opposite of how you stop the spread. Also, isn't there a step where you have to *pay for your goods*? Couldn't you check vaccination status at checkout instead of on entry and avoid this? You could, but then you wouldn't be able to publicly humiliate the unvaccinated, and the cruelty is (quite literally) the point.

## NPIs Including Mask and Testing Mandates

[Netherlands lets our people go](#), ending all Covid restrictions.

[Denmark lets our people go.](#)

[A proposal](#) to skip PCR tests entirely and switch to relying on antigen tests, because the speed premium exceeds the accuracy premium. Seems mostly right to me.

Two years later, everything has changed, also nothing has changed? [University of Washington develops Covid-19 test](#) that is fast, cheap and accurate, combining strengths of rapid tests and PCRs.

Researchers at the University of Washington have developed a new test for COVID-19 that combines the speed of over-the-counter antigen tests with the accuracy of PCR tests that are processed in medical labs and hospitals.

...

The accuracy of COVID-19 tests has been a pressing matter throughout the pandemic. Many at-home antigen kits for COVID-19, which detect pieces of the proteins the virus creates instead of its genetic material, are 80-85% accurate, though accuracy may drop with the omicron variant, which harbors a relatively high number of mutations not found in other strains. PCR tests are generally 95% accurate or better — a key FDA benchmark — but require expensive equipment and a long wait for results.

Initial results reported in the paper show that the Harmony kit is 97% accurate for nasal swabs. The Harmony kit detects three different regions of the virus' genome. If a new variant has many mutations in one region, the new test can still detect the other two. It can, for example, detect the omicron variant, which has dozens of mutations in the region of the genome that encodes the so-called spike protein.

And of course, since that's the only mention of the FDA, we can assume that the University of Washington's vastly superior new test kits are not legal for use. Nothing ever changes.

NFL decides to be true to itself, put the best product on the field for the playoffs, [and screen for symptoms](#) instead of testing players daily. No reason to be looking for the asymptomatic cases at this point, the show must go on. It definitely worked. Wow did they deliver their best product on the field for the divisional round.

[Bari Weiss giving the audio version on Real Time of why she's done with Covid.](#) Also Kerry's memory of being repeatedly told that everything was fine, all our prevention made sense and specifically to watch Tiger King. And yeah, it's time to stop all this, but also if you still haven't watched the original season of Tiger King you should totally do that, it's *almost* as crazy as our Covid policies.

On that note, [Hong Kong residents are paying up to \\$26k to fly their hamsters out on private jets](#) to save them from the government's cull.

A feature of the world is that when people famously take extreme positions like the one from the Great Barrington Declaration, sometimes they are doing it because they think very differently than most others and have some strange models. Other times they are doing it because they [wrote the conclusion at the bottom of the page](#) and worked their way back up the same way everyone else did except they wrote down a different conclusion.

In either case, usually [they do not stop there.](#)



James Surowiecki ✅ @JamesSurowiecki · Jan 22

...

Great Barrington Declaration co-author Sunetra Gupta now says ppl with Covid should not isolate, but should circulate freely in the community, because the best way to protect the elderly and the immunocompromised is by maintaining "herd immunity through constant reinfection."

This strongly endorses the immediate lifting of restrictions given their devastating consequences, and focusing policy instead on the vulnerable. The low risk of infection among the vulnerable arises from the maintenance of herd immunity through constant reinfection; any measure that interferes with this actually increases their risk.



James Surowiecki ✅ @JamesSurowiecki · Jan 22

...

"Herd immunity through constant reinfection" is one of those phrases that captures an entire bizarre ethos in just a few words.

The Great Barrington Declaration was an extreme position, but it was easy to understand why, if the physical world worked in the right ways and you had the right values, it could be the right thing to do. There was an actual physical model and cost benefit calculation behind it. You could argue it was right or wrong.

This, on the other hand, is [The Chewbacca Defense](#). It does not make sense. I have gone over this a few times trying to find a way to make it make any sense. No success. If you have constant reinfections, you don't have herd immunity. If you have herd immunity, you don't have constant reinfections. Either more people get infected, or less people get infected. This is using the phrase 'herd immunity' like a talisman that can protect people without thinking about what the phrase actually means, the same way that those calling for more precautions have their own meaningless talismans and sacrifices to the Gods.

A judge threw out New York State's indoor mask mandate, calling it unenforceable and unlawful, [but that order was stayed on appeal](#), so the mandate remains in place. Presumably by the time the final ruling comes out the point will be moot.

[Thread explaining why we should all be 'done with Covid' shortly.](#)

For others, [can you be too safe from the Covid monster?](#)

For those of you who doubt the reality of Covid Derangement Syndrome – the syndrome that drives people to believe that every further reduction, no matter how small, in the risk of encountering the SARS-CoV-2 virus is worth whatever the cost of this reduction, no matter how large – take a look at this photo of a traveler at the Charlotte airport. (HT [Gary Leff](#) and Todd Zywicki)



If you think about Covid precautions as things one does in order to satisfy power and prevent social awkwardness, this strategy is obvious nonsense, since it is worse at both tasks while being more annoying. *However, if your goal is to prevent Covid-19 for real, then it's more efficient to go all-out and use technology that is mildly socially awkward and annoying, but*

that actually works. It's the middle position, of crippling one's life without getting the job done properly, that doesn't add up.

Where are all these free tests coming from? [One data point, from the comments](#), suggests the obvious.

My test supplier is claiming the feds are seizing most shipments of tests for their 4-free-tests program.

Utah Governor, [seen here](#) saying '[now hear me out](#)', suggests that if you have symptoms you should stay home and isolate to avoid getting others sick because you're sick but not worry about getting a test.

That got a bit of mocking, but isn't that *exactly what you should do if you're sick with something else too?*

If you have Covid you should stay home, take some time off, rest up and not get others sick. If you have the flu, or a cold, or other minor illness, yes, you should stay home, take some time off, rest up and not get others sick. So if you test, it doesn't much matter what result you get, so it makes sense to consider not testing. No value of information, unless it impacts what you do once you've recovered in terms of risk management.

Reminder that [different areas continue to react to the pandemic very differently](#), and it does not seem to impact case counts all that much.



**Nate Silver** @NateSilver538 · 18h

OpenTable restaurant reservations in January 2022 as compared with January 2019:

Miami: +14%  
Las Vegas: +1%  
Houston: -13%  
Denver: -26%  
LA: -41%  
Boston: -48%  
Brooklyn: -55%  
DC: -59%  
Manhattan: -64%  
San Francisco: -66%  
Cambridge, MA: -75%

[opentable.com/state-of-indus...](http://opentable.com/state-of-indus...)



**Megan McArdle** @asymmetricinfo · 18h

Yeah, I'm making a point of going out to eat every week, both to shake pandemic atrophy and to support local restaurants. I'm eating in a lot of nearly empty restaurants.

I don't know if it's Omicron fear, remote work, or folks just out of the habit, but things are \*not\* normal. [twitter.com/DouthatNYT/sta...](https://twitter.com/DouthatNYT/sta...)

I expect to be back to my baseline restaurant use rate going forward, after a month of going back to delivery.

## Think of the Children

[Let my people go.](#)



PoliMath  
@politicalmath

...

"stop complaining, everything is back to normal"



Elisabeth Stineberg @elisawine14 · Jan 23

My 7 yr old son wears a mask for 7 hours a day, inside & outside. He is permitted to remove his mask for about 20 minutes per day, during lunch. While eating, he is forbidden to speak. He, & his 16 classmates, must face forward. A screen is turned on to discourage socialization.

[Show this thread](#)

[Let my people go.](#)



Robby Soave ✅ @robbyssoave · 22h

...

LA will force students to be masked in non-cloth masks with nose wires, even while playing sports, even outside. (!) Vaccine mandate to come as well. Why kids have to deal with harsher pandemic policies than anyone else, still unexplained!

[Article from a Canadian teacher explaining how not all right the kids are.](#) Let my people go.

Let my people stay. University of Florida used to provide 'isolation housing' for students with Covid. Then they stopped doing that and instead started telling kids they had to leave campus, as in 'serious consequences if not gone by 5pm,' [so as an example of what happens next, a student who was positive decided to go home and boarded a flight](#). Also it seems many places don't let anyone under twenty-one book hotel rooms, which seems crazy?

For almost two years children have been unable to see the faces of those around them. It was pointed out that maybe this was a problem, and that was widely dismissed as crazy talk by all the Very Serious People. Now that it's time to take masks off, this position has been reversed with the only new evidence being 'children aren't learning' and 'well obviously I mean come on' so it's time for the rewriting of history.

[Remember.](#)



Michael Brendan Dougherty ✅ @michaelbd · 9h

...

This is a paragraph in the WaPo about unmasking kids at schools. Remember earlier in the pandemic when parents started finding papers like those mentioned here, and the papers were withdrawn? I do.

But universal masking policies in schools have not come without costs. In certain parts of the country where CDC mask guidelines were rigorously adhered to, children have not seen the faces of their teachers or classmates since early 2020. This affects [learning](#) and development, [particularly for our youngest learners](#). Maintaining aggressive mitigation policies, including strident mask rules, also sends children, families and staff the message that schools are not safe. This is simply not true.

[This is the WaPo article.](#) A reply to the post above that was an emoji plus a link to the article *was hidden by Twitter as potentially offensive content, which reminds me of this.*

When I follow the links in the quoted paragraph, I find evidence that the children are not learning, but no attempt to link it to masks. I do think the masks contribute, but we also *closed schools* and *did insanely terrible remote learning* so it's not like 'test scores are down' is strong evidence that *something else* is terribly wrong. This is a Just So story that happens to obviously be true, but it was also obviously true a year ago.

Suddenly torturing children is *bad*?

As parents of school-aged children, we have also noticed a disturbing pattern, especially during the omicron surge: punitive mask culture. This can take many forms. Some older children, for example, have been given [detentions](#) and [even suspensions](#) due to "mask slippage" or improper mask-wearing. Younger children have also been subjected to [harsh rules](#) to minimize the spread of covid-19 during lunch hours, often their only mask-free time during the school day. Some must eat in [total silence](#) during mealtimes. Others have ["expedited" lunch](#) or are instructed to [pull their mask down to take a bite and pull it back up to chew](#). Activities intended to relieve stress, including recess and gym, have been curtailed because of ["increased exhalation."](#)

And not torturing students is *good*, the kids might actually *smile*?

Hopkinton High School in Massachusetts, in a region which boasts a 98 percent vaccination rate, [offers a glimpse](#) of what happens when students are given the opportunity to move from mandatory to optional masking. During their brief masking-optimal pilot, the school reported that "smiling is more contagious than covid-19," and a survey of students found that 70 percent said the policy improved their experience, including their ability to learn.



PoliMath @politicalmath · 9h

I can't find it now, but I remember when a national pediatric society ghosted their flyer on how important seeing faces is for infants and toddlers

...

This pandemic has shown that, for many people, politics is their first and primary loyalty, even over the wellbeing of children



**PoliMath** @politicalmath · 10h

Replies to [@politicalmath](#)

...

That part of all this has just been revolting. Watching people who are supposed to be committed to an ideal completely abandon their commitment to truth in order to gain their team 15 minutes of good press.

1

9

148

↑



**PoliMath** @politicalmath · 10h

Just tell the goddamn truth, people really will understand.

...

If your position is "yes, seeing faces is important, but we think the risk is too high" just fuckin say that. But then let them make their own choices. Stop lying to everyone.

5

21

167

↑



**PoliMath** @politicalmath · 9h

Replies to [@politicalmath](#)

...

Here is it: The American Academy of Pediatrics published a flyer "Face Time and Emotional Health"

They unpublished it in the middle of the pandemic but told a [@Reuters](#) fact checker that it was just part of a site migration & would be republished

It never was



PoliMath @politicalmath · 9h

...

This is a society of doctors who are supposed to be looking out for children and they not only pulled down their info about kids and face time, they lied about it, refused to republish it, and then published flat out lies saying the opposite is true



American Academy of Pediatrics @AmerAcad... · Aug 12, 2021

Babies and young children study faces, so you may worry that having masked caregivers would harm children's language development. There are no studies to support this concern. Young children will use other clues like gestures and tone of voice.

[healthychildren.org/English/health...](http://healthychildren.org/English/health...)

[Show this thread](#)

**Being around adults wearing  
masks doesn't delay  
babies' speech or language  
development.**



**Real Talk**



PoliMath @politicalmath · 10h

...

I don't think I can ever trust this organization with anything even remotely political again. They simply will not tell you the truth about your own child if that truth is unpopular in their political circle. Your child's welfare is secondary to being on the right political team

Earlier in the linked thread are claims that masks don't compromise children's breathing (uh huh) and that they are vital to keeping kids safe (which they aren't because kids are safe regardless) and it ends with saying masks are not forever (unless, of course, they and those like them have their way).

Here is [their link's](#) argument that masks aren't a problem.

A key part of learning to communicate for a child is watching the faces, mouths, and expressions of the people closest to them. Babies and young children study faces intently, so the concern about solid masks covering the face is understandable. However, there are no known studies that use of a face mask negatively impacts a child's speech and language development.

And consider this: visually impaired children develop speech and language skills at the same rate as their peers. In fact, when one sense is taken away, the others may be heightened. Young children will use other clues provided to them to understand and learn language. They will watch gestures, hear changes in tone of voice, see eyes convey emotions, and listen to words.

So we start off with [a classic No Evidence](#), then proceed to quite the [Proves Too Much](#). I mean who uses *visual aids* to help with *language*?

My son's speech therapists, for one. Whenever they talk to him, they put their hands by their mouths and use highly exaggerated facial expressions, every time. I wonder if a *face mask* might interfere with this process of improving one's speech? They even *agree*, within the same document, saying that speech therapists are *going virtual in order to avoid masks* or using clear masks so children can see their lips move.

This, in many cases, [is too much to ask](#).

And here, right below their claim that there's no problem, is their *advice on what to do to mitigate the problem that they just claimed didn't exist*.

## Tips to better communicate with children while wearing a face mask

Here are some ways that people who wear face masks can help when they interact with your child:

- get the child's attention before talking
- face the child directly and make sure nothing is blocking the child's view
- speak slowly and slightly louder (without shouting)
- ensure a child is using hearing aids or using other hearing devices, if they have been prescribed
- use eyes, hands, body language, and changes in tone of voice to add information to speech
- ask the child if they understood; repeat words and sentences when necessary
- reduce noise and reduce distractions

You can share these tips with your child's day care provider, preschool, and others who regularly interact with your child while wearing masks.

Those are good ideas. Totally do those things. They work for adults, too.

But if you need to get attention, reduce noise and distractions, speak slower and louder and use other body parts to 'add information to speech' then there can be zero doubt that this is a substantial reduction in communication ability and bandwidth.

Which you then claim doesn't matter.

This is lying, but it's more than that. It's gaslighting.

Now we're being told that such messages were never sent and such claims were never made. Same as it ever was.

I agree that one can't trust such organizations in similar future situations. One can still have some amount of trust *once you control for the political considerations*. Often there are no political considerations, or they are in an obvious direction so one can adjust for them.

You do also have to control for their level of understanding of logical reasoning.

## In Other News

The Florida Man (and Governor) Ron DeSantis spent this week demanding a continued supply of monoclonal antibodies whose emergency use authorization was withdrawn by the FDA because of the whole '[they don't work against Omicron](#)' issue, along with the fact that almost all cases are now Omicron. They are [instead being sent what treatments we have that do work](#), but that is not good enough for Ron DeSantis. He instead demands that Biden overturn the FDA's decision as if that is a thing Biden is capable of doing.

I would suggest to DeSantis that if he thinks that those who think treatments are worth using should be free to use those treatments even when the FDA does not approve of them (and even in cases where they do not work), he should generalize this insight into something like "FDA Delenda Est."

Meanwhile, [here's his surgeon general nominee](#) deciding it would not be safe and effective to answer the question of whether vaccines work.

[Japan's successful Covid strategy relied heavily on sharing information and less than other places mandates.](#) Cultural factors were likely the central reason for Japan's success, but was that *in addition* to the lack of severe behavioral mandates and the unusually free flowing of information, rather than *in spite* of those differences.

[My friend Will Eden's child has Covid, despite generally extreme precautions.](#) Even extreme precautions sometimes fail, and sometimes a complete lack of precautions still succeeds. Feels bad to trip up at the very end like this after all the sacrifices, but that's sunk cost fallacy and not the right way to think about it.

[Thread of people saying we should eradicate Covid-19](#), without saying how to do this impossible task.

[Risk, what it both is and is not.](#)



**Robert Wiblin** @robertwiblin · Jan 24

...

The UK government made a risky call not to impose new COVID restrictions in late December, and it looks like that decision will be vindicated.

Hospitals received a lot of COVID patients but not overwhelmed.  
Restrictions mostly would have just spread things out.

1

3

34

↑

It is a 'risky' decision not to impose restrictions. It is also a 'risky' decision to impose them, as this comes with very large costs. When one is viewed as risky and the other as safe, the

wrong decisions get made on the margin, and sometimes rather far from the margin.

[Biden administration opposes strengthening the WHO](#), which is great news especially given that Biden previously rejoined the WHO. The WHO is trying to use its track record of active interference throughout the pandemic on behalf of the virus (and occasionally on behalf of China) as evidence that it should get more power and money in order to interfere in the future.

BRUSSELS, Jan 21 (Reuters) - The United States, the World Health Organization's top donor, is resisting proposals to make the agency more independent, four officials involved in the talks said, raising doubts about the Biden administration's long-term support for the U.N. agency.

The proposal, made by the WHO's working group on sustainable financing, would increase each member state's standing annual contribution, according to a WHO document published online and dated Jan. 4.

The plan is part of a wider reform process galvanized by the COVID-19 pandemic, which has highlighted the limitations of the WHO's power to intervene early in a crisis.

But the U.S. government is opposing the reform because it has concerns about the WHO's ability to confront future threats, including from China, U.S. officials told Reuters.

[Paper calculates the costs of various Covid treatments](#), finds many including fluvoxamine don't on net save money compared to Medicare estimated hospital costs. I notice I am confused why this is the comparison point, since Medicare pays artificially low prices subsidized by non-Medicare patients and also I hear it's better for people to not need to go to the hospital even when they're not under stress. Still, it's good to keep in mind when considering early treatment that the NNT is going to be high because most patients were always going to get better on their own.

[Did CDC director Walensky actually not know that vaccinated people could transmit Covid-19 until that deeply flawed Provincetown study?](#)



By **Michael D. Shear, Sheryl Gay Stolberg, Sharon LaFraniere and Noah Weiland**

Jan. 23, 2022

WASHINGTON — Dr. Rochelle Walensky was stunned. Working from her home outside Boston on a Friday night in late July, the director of the Centers for Disease Control and Prevention had just learned from members of her staff that vaccinated Americans were spreading the coronavirus.

Vaccines had been the core of President Biden's pandemic strategy from the moment he took office. But as Dr. Walensky was briefed about a cluster of breakthrough cases in Provincetown, Mass., the reality sank in. The Delta variant, which had ravaged other parts of the world, was taking hold in the United States. And being vaccinated would not, it turned out, prevent people from becoming infected with the variant or transmitting it.

Wow. I mean, wow. Other than mumbling CDC Delenda Est I have no idea what to say to that.

Sarah Palin has made it very clear she has no intention of getting vaccinated. This past week, she was in New York City to take part in her defamation lawsuit (which I admit is a bit on the nose), and [she took time out to illegally indoor dine at a fancy city restaurant and test positive for Covid-19](#). While known to be Covid-positive she [decided to move to dining outdoors instead](#). I'm going to go ahead and say that's a good time to get delivery. Her lawsuit against the New York Times has been postponed.

## Other Prediction Updates

This is being carried over from the final Omicron post.

**China keeps daily cases under 50 per million through 2022: 30% → 40%.**

It's a long game. A week has passed without incident, which is good news, and the containment in India makes me think the task might be more realistic than I thought, so bumping this up.

**Day of the case-average peak: January 15 → January 16 (resolves to the 16th)**

Turns out my final guess was slightly optimistic, my guesses before that moderately pessimistic. Good to have this behind us.

**Will There Be a Federal Mask Requirement on Domestic Flights as of November 8, 2022? 44% → 35%.**

[Polymarket](#) is back at 44% but we have a bunch of new information. All restrictions have been lifted in both Denmark and The Netherlands, and more than that the tide is clearly turning against restrictions, faster than I would have anticipated. It seems much less likely now that we'll sustain these kinds of restrictions the whole year.

**Chance that Omicron is vastly (75%+ in the same person) less virulent than Delta:**  
40% → 35%

Deaths exceeded my projections, so I should adjust this accordingly.

**Chance we will be getting boosters modified for Omicron within 6 months of our previous booster shot: 15% → 20%.**

The move to a five month cycle over a six month cycle makes this seem more likely to accidentally end up happening in a way that's different from the intended question that was being asked, so bumping it up. I'd be very surprised if we get that much voluntary uptake, and even more surprised than that by actual requirements with teeth barring a new scary variant, but I interpret this question as it being what officials are telling us to do rather than what most people actually do.

[\*\*Polymarket on When We'll Be Below 100k daily cases:\*\*](#)

**Choose a market:****February 15, 2022?**

Vol: \$5,236 Liq: \$300

Yes: \$0.09

No: \$0.91

**March 1, 2022?**

Vol: \$34,228 Liq: \$6,225

Yes: \$0.25

No: \$0.75

**March 15, 2022?**

Vol: \$57,033 Liq: \$14,675

Yes: \$0.36

No: \$0.64

**April 1, 2022?**

Vol: \$4,832 Liq: \$7,025

Yes: \$0.58

No: \$0.42

**April 15, 2022?**

Vol: \$528 Liq: \$300

Yes: \$0.82

No: \$0.18

**May 1, 2022?**

Vol: \$1,170 Liq: \$300

Yes: \$0.90

No: \$0.10

Here are (1) [market's estimates on January 17](#), then my estimates on January 17, and the third number is current market prices, then my current values.

February 15: 8% / 15% / 9% / 10%

March 1: 24% / 55% / 25% / 35%

March 15: 48% / 70% / 36% / 55%

April 1: 74% / 78% / 58% / 65%

April 15: 84% / 82% / 82% / 70%

Mary 1: 89% / 84% / 90% / 73%

To get there, we'll need to cut cases in half close to three times from here. Whether or not that happens by March 1 depends on whether cases level off like they did in the UK, or whether they decline more like the way they went up or how they are declining in the Northeast. We've seen enough leveling off in other places that it seems unlikely that we will make it by March 1, but I still do expect to get there by about March 15. If we don't get there by soon after that, I think a lot of that is because either there's a new variant or cases stabilized at a higher level than 100k/day, so I don't think there's that much chance of it happening quickly after that if it hadn't happened already.

Standard warning: I'm not thinking about these numbers as carefully as I would if I were considering making a trade, so think carefully before trading and this is of course not investment advice.

That's also why I'm not going to be looking at the other currently available markets - it's a state change to go focus on such questions, and it's not worth doing that many state changes every week. Instead, my plan is to choose a sampling of one or more each week, based on what is interesting, and also to allow comments that suggest things that look like they might be out of whack.

## Not Covid

[Public pre-K once again found to be actively harmful to children's educations.](#) Also didn't help behaviorally, so presumably this was all-around net harmful, despite the help this gave to parents. Yet somehow the conclusion by Freddie here is 'yes using child care like this is harmful to kids but we should offer universal free child care instead of giving parents money anyway, thus economically coercing parents to use it even if they'd rather stay home with their kids.' I am confused why the conclusion isn't 'seems like we should try to help families in need and help them *avoid* using such services unless they need them, so do the default obviously right thing and give them money.'

Peterson's statement to open Rogan that 'climate is everything therefore your models are useless' [as a fully general argument against anyone ever knowing anything.](#) Hopefully when it's from this direction people who use it in other directions can better recognize it. Also, Neil Young's catalog was pulled from Spotify because he said it was him or Rogan, and Spotify can do math.

[Innovate UK application for startup working to increase diversity in clinical trials turned down in part because 'negative impact of automation may lead to job loss in industry.'](#)

# How Bad Is QWERTY, Really? A Review of the Literature, such as It Is

This is a linkpost for <https://www.erichgrunewald.com/posts/how-bad-is-qwerty-really-a-review-of-the-literature-such-as-it-is/>

Some six or so years ago, not long after I had started working as a programmer, I developed a bad case of repetitive strain injury (RSI) and for a while couldn't use my left hand for typing, but had to rely on my right hand and dictation. This situation being obviously untenable, I took action: I purchased an ergonomic keyboard (Microsoft Sculpt), remapped shift to the space bar when pressed in combination with another key, started doing hand/wrist stretches and switched from typing on the QWERTY layout to [Colemak](#). This grab bag of interventions worked, my RSI got better and, though it still flares up sometimes, I now have it under control.

I have stuck with Colemak since then and have liked it and occasionally even recommended it to friends and colleagues. But beyond reading popular articles about the topic, I never really looked into the evidence on key layouts and ergonomics<sup>[1]</sup> and productivity<sup>[2]</sup> – until now, that is!

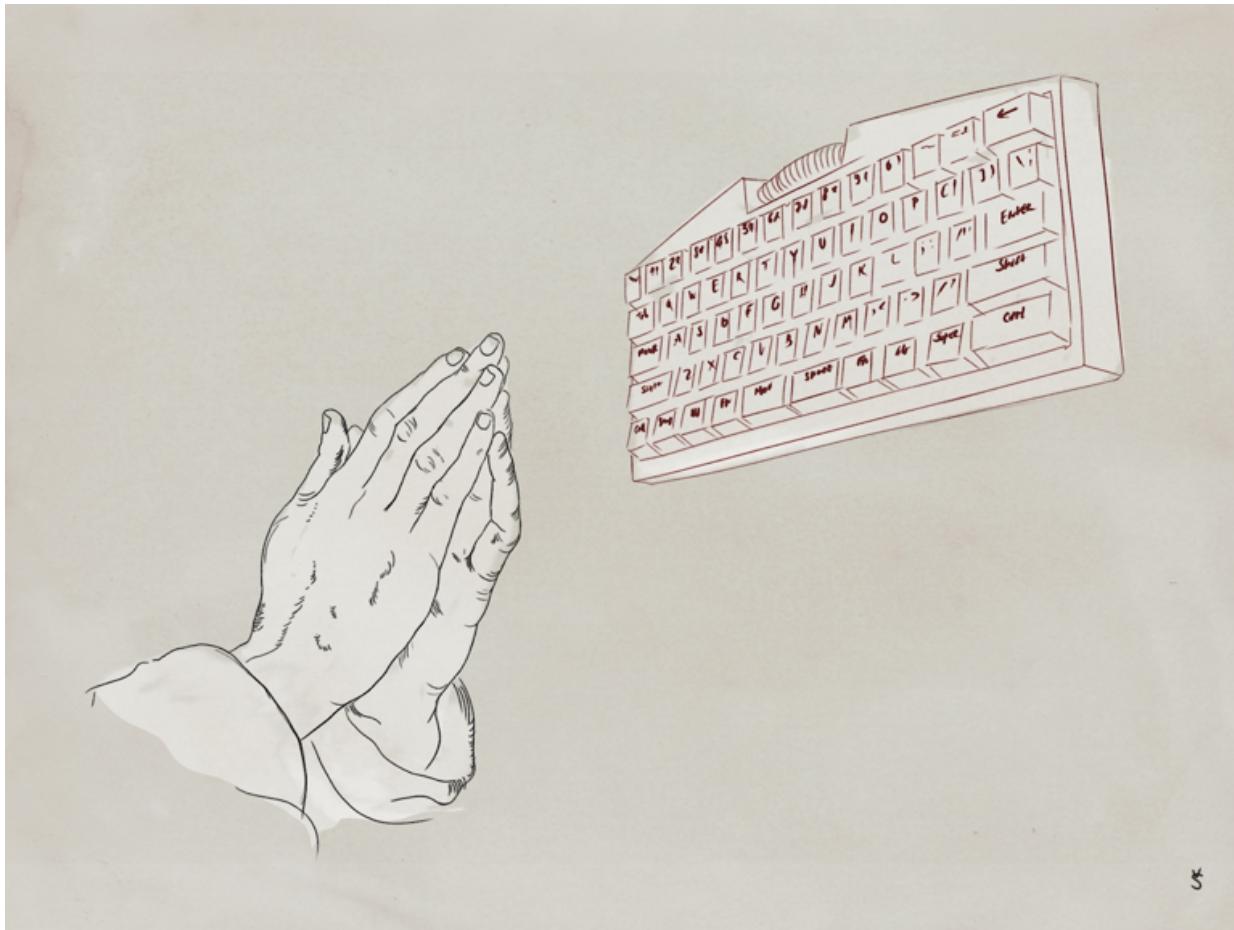
Before we get to the good stuff, a word of caution. I am in no way an expert on this; I estimate I spent roughly 15-25 hours researching and taking notes, so there is still a lot that I'm uncertain about. My target group here is touch-typing office workers, typing mainly in English, without disabilities that significantly impede their typing (e.g. the loss of an arm or that sort of thing). For people who don't touch-type, QWERTY is probably fine. For people who have disabilities that affect their typing, I assume special considerations, of which I know little, apply.

## Summary

My current model on the RSI risk is something like this. The evidence has made me update towards the risk of developing serious RSI from keyboard use being small. I had a pretty strong prior, based on what I've heard from other programmers and on my own experience, that it often causes RSI. I still think it's a risk, but not as large as I did before, though with lots of uncertainty (80% confidence interval would be 2%-20% chance of developing keyboard RSI over a lifetime for the aforementioned target group). I think key layout is a minor part of the ergonomic harms of keyboard use; keyboard type (i.e. ergonomics of design) and posture/positioning both seem more important (70% confidence).

As for productivity, compared to e.g. Dvorak or Colemak, QWERTY seems somewhat slower for typing on the margin (80% confidence). But most people, and most programmers even, don't type enough that switching would result in a significant productivity boost (80% confidence).

In general, the QWERTY layout, while undoubtedly suboptimal, doesn't seem as bad as I had thought previously. If you are interested in making the switch and don't mind the transaction cost, you might see small benefits on the margins; if not, don't bother.



## QWERTY - Designed to Be Slow?

Here is an anecdote [recounted](#) by Michael Tiemann, expressing a commonly held belief:

One weekend I visited Richard Stallman at MIT and I was shocked to learn that he could no longer type. He was given strict instructions by his doctor to not touch a computer keyboard for 6-12 months, and that if he did, he may lose forever his ability to type. He was a programming pioneer, and at the time, his symptoms were not well known or understood. We all came to understand that it was RSI – repetitive stress injury, exacerbated by the very keystroke combinations that made the Emacs editor such a powerful programming environment. But the root cause was not Emacs – it was the [punitive] design of the QWERTY keyboard, a legacy of the industrial era when complex keyboard mechanisms were not able to keep up with the speed of human fingers. The solution? Design a keyboard so confounding and unnatural that no human could ever outrun what the mechanism could handle. Or so they thought. Humans are remarkably adaptable, to a point. Human typists did manage (with practice and perseverance) to type every bit as quickly as before QWERTY, but at the expense of long-term health.

In other words, the idea is that QWERTY was designed so that humans would type as slowly as possible in order to prevent mechanical jamming. This idea is apparently false.<sup>[3]</sup> Christopher Latham Sholes and others developed QWERTY in the 19th century

initially for telegraph workers to receive Morse transmissions, where typing speed would have been important.<sup>[4]</sup> (Though on the other hand, jamming is really bad for typing speed!) Quoting Yasuoka & Yasuoka:

The development of QWERTY was a winding road, first by Sholes and others, second by Harrington and Craig, then by Jenne and Clough, again by Sholes, and at last by Wyckoff, Seamans & Benedict. There was no consistent policy towards QWERTY. The keyboard arrangement was incidentally changed into QWERTY, first to receive telegraphs, then to thrash out a compromise between inventors and producers, and at last to evade old patents.<sup>[5]</sup>

That said, QWERTY was an early attempt at designing a key layout, and it was designed under conditions very different from those we have now. It would be surprising if the last 150 years of keyboard use and technological change would not have revealed some real flaws in QWERTY – if we got it pretty much right after no more than a few tries.

(Many of the academic discussions of QWERTY's history and suitability have taken place in journals of economics. Economists are interested in whether QWERTY is an example of **path dependence**. Is the market stuck in a suboptimal equilibrium due to a seeming accident of history – the adoption of QWERTY for early typewriters – or would events have played out the same way every time even if circumstances had been slightly different? As a result, much of the literature seems to have been written answering not the question "Is QWERTY suboptimal?" but the question "Is the market efficient?" and should therefore be treated with some suspicion.)

The second claim that Tiemann makes is that QWERTY is bad for ergonomics but not for typing speed. (Though confusingly he compares QWERTY to whatever was "before", even though QWERTY became the standard when typewriter technology was still in its infancy.) We'll see whether this second claim holds up.

## Theory

In theory, alternative keyboards like Dvorak and Colemak have a lot going for them. Many characteristics have been proposed as significant, for example:

- It's better if fingers have less traveling to do.
- It's better if consecutive taps are done with different fingers or, better yet, different hands.
- It's better if common keys are near the fingers' natural resting places.
- It's better to avoid stretching and overusing the pinky finger, which is the weakest of the five.

I have had a hard time finding much evidence either way on whether these factors make a difference as far as ergonomics and productivity are concerned. It really does seem like many of them *should* make a difference. For example, in my experience awkward pinky movements really do cause discomfort or pain, and this also makes sense given what I've read about RSI. And it seems obvious that one can type faster if one's fingers need to travel less, right?

Not so fast! Typing could be constrained by neurological not physical speed, for example.<sup>[6]</sup> And awkward pinky movements could feel uncomfortable or painful because of RSI without actually causing it, the way jumping up and down can feel

painful when you have a headache without having actually caused the headache. So is there any evidence that these factors affect ergonomics or productivity?

I really only found one study that's relevant here. The researchers had a subject (singular) spend six weeks typing away first at a QWERTY keyboard and then at a Dvorak keyboard.<sup>[7]</sup> Although the subject typed faster on QWERTY (likely due to her prior experience), typing on the home row and typing consecutively with different hands both increased typing speed, and these are both things that Dvorak and Colemak do better than QWERTY.<sup>[8]</sup>

Some forward-thinking people have set about evaluating key layouts objectively on metrics like these. For example, the [Carpalx](#) project has a quantitative model for measuring the typing effort of a key layout, based on things like finger travel distance and awkward finger reaching. (The model is parameterised, meaning it can be adjusted to give different weights to various characteristics, but from reading the project's extensive documentation I have the impression that a lot of thought has gone into default parameter selection.<sup>[9]</sup> The documentation is well-written and extensive, so I recommend having a look at it if you want to learn more about the model.)

By Carpalx's measures, QWERTY is bad, though better than the average random key layout. Dvorak does better, and Colemak better still:

The typing effort reduction in moving to Dvorak is 30%. Each effort component for Dvorak is lower. The finger component of the penalty is higher, due to Dvorak's increased use of the pinky (18% of keystrokes vs 10% on QWERTY are done by the pinky), which can be seen by a +56% increase in the finger penalty over QWERTY. If you have very weak pinkies, Dvorak may not be for you.

Dvorak certainly does a good job in lowering the base effort – this is the component of effort that measures finger travel distance – by moving frequently used keys, such as vowels, to home row. Because Dvorak vowels are serviced by the same hand, hand-alternation is increased (one-character hand run is 62% for Dvorak and 51% for QWERTY), a desirable property of an efficient layout.

The Colemak effort has an even lower effort than Dvorak and it achieves this by further lowering all three components of the effort. It significantly decreases the penalty component seen in Dvorak. Thus, although Colemak's finger penalty is still higher than QWERTY (again, through the increased use of the pinky), [the] overall penalty is lower by 24%.

## Ergonomics

RSI is an umbrella term for injuries that occur due to "repetitive movements, awkward postures, sustained force, and other risk factors".<sup>[10]</sup> The most common and costly form it takes is carpal tunnel syndrome (CTS), where the median nerve has gotten compressed where it travels through the wrist, leading to pain, discomfort, numbness, etc.<sup>[11]</sup> QWERTY plausibly increases the risk of CTS because it involves rotating the wrist towards the pinky more often to reach characters at edge of the keyboard.<sup>[12]</sup> (This is not the only ergonomic issue with keyboard use, but as far as I can tell, it is the only issue that varies for different layouts.) So once again Theory whispers, "QWERTY baaad." Is it?

My first thought here was, if keyboard use in general doesn't cause RSI, bad key layouts won't be causing it either. So that seemed like a good sanity check.

The NHS [writes](#) that "[s]pending a lot of time using a computer, keyboard and mouse is a common cause of RSI", but without citing any specific studies. Of course what I want to know is not "What proportion of RSI cases were caused by keyboard use?" but "How likely is keyboard use to lead to RSI?"

Fagarasanu & Kumar is a somewhat old (2003) review on CTS; the authors explain that, when we type, we adopt awkward hand/wrist postures that increase carpal tunnel pressure (CTP) which causes CTS.[\[13\]](#) They write:

When using traditional QWERTY key layout, both forearms are pronated and both wrists are in ulnar deviation and extension. [...] CTS was attributed to keyboarding in 8% of cumulative trauma disorders. Sauter et al. conducted a study with 932 [computer] users and assessed discomfort in wrist and right hand at 13% and 12% respective from the total sample. [...] Although there is a strong evidence of a causal relation between keyboarding and pointing devices on the one hand and CTS occurrence on the other, the role of every single design element is not known.[\[14\]](#)

In other words, keyboard use can cause CTS, but we don't really know what about it we need to change in order to improve things.

(It's perhaps worth pointing out that, as risk factor for RSI, keyboard use is way less risky than some manual labour.[\[15\]](#)[\[16\]](#) Most CTS patients are "industrial workers, females, and the elderly", not young male programmers.[\[17\]](#) Reported CTS has decreased overall in the United States labour force, and appears in larger numbers in blue collar industries than in white collar industries even though there are over 50% more workers in the latter; mostly it seems that manufacturing work is to blame.[\[18\]](#) I find this easy to believe; manufacturing work seems like really hard work, especially on the body.)

A 2008 review, Thomsen et al., seems inconclusive on whether keyboard use causes CTS; it did find increased CTP with keyboard and mouse use, but levels were below what's considered harmful.[\[19\]](#) Wright & Atkinson (2019) advise doctors that "[p]atients with CTS can be counseled [...] that use of a keyboard and mouse at work is unlikely to be the cause of their symptoms, but it is possible that these activities may aggravate them."[\[20\]](#) A 2010 review on musculoskeletal disorders and computer work found "limited evidence for a causal relationship between computer work per se, computer mouse and keyboard time related to a diagnosis of wrist tendonitis" (tendonitis being tendon pain and swelling).[\[21\]](#) A more recent (2021) review looking at workplace CTS prevention finds mixed quality in studies; they write:

[D]ecreasing ulnar drift or maintaining the wrist in neutral position with measures such as changing or modifying [accessories like keyboards] may reduce exposure to CTS. However [...] the results were not statistically significant when comparing the use of conventional ergonomic elements and the desired intervention. Results may be different if longer exposure times were analyzed, if bigger sample sizes were used, if subjects had more time to become familiar with the modifications made, or if bias assessment was controlled, among others.[\[22\]](#)

In Pan & Schleifer (1996), subjects experienced more arm discomfort/pain/fatigue the more keys they pressed while doing a data entry task.[\[23\]](#) Finally, Feng et al. is a recent

(2021) cross-sectional study that found that "prolonged computer use time and working without breaks were associated with presence of wrist/hand symptoms".<sup>[24]</sup> But of course correlation is not causation, and there are many potential confounders here, e.g. maybe people who spend lots of time at the computer exercise little, and it is the lack of exercise, not computer use, that causes problems.

So we have the theory, a bunch of anecdotal reports and some (non-randomised, non-controlled) studies suggesting there is a causal link between keyboard use and CTS, and a bunch of (non-randomised, non-controlled) studies finding little evidence for such a link. The studies I've come across here leave something to be desired – they are much less rigorous than the studies I read when researching [health effects of veganism](#) – possibly because the magnitude of the problem isn't that great (or maybe it is the nefarious influence of Big Keyboard).

If keyboard use does carry the risk of developing RSI, what is it about the keyboard that's bad? Is it the physical design, the key layout, hand/wrist posture, or something else? My impression is that key layout is a relatively small component here, for several reasons. The first is my own experience, according to which it's much more important to use a split keyboard, say, than the appropriate layout if I want to avoid RSI flare-ups. The second is that CTS is largely caused by CTP, which in theory seems more impacted by physical design (chiefly whether a keyboard is split and/or tilted/tented) and less by finger stretching or the horizontal rotating we do with our hands to reach keys at the sides of the keyboard. The third is Carpalx's model, which suggests that established alternatives like Dvorak and Colemak, while better on the whole, use the pinky more heavily than does QWERTY – maybe it is a little bit bad to reach for the outermost keys, but any layout will have some keys at the extremes, so perhaps the difference between layouts just isn't that great.

What about QWERTY specifically? I wasn't really able to find any research on this. Maybe that's because it's very hard to design experiments to test it? You can't just take a bunch of people and ask half of them to start using Dvorak, because there's a significant learning curve involved. But you don't want to find out if *learning* a new layout is good, you want to find out *using* it is good once you have learned it. There is no natural control group for these experiments, and no obvious placebo.

In sum, keyboard use in general does seem to cause RSI, but the risk seems fairly small. Bad key layouts may only be a minor part of the RSI risk, though QWERTY does seem worse than most alternatives, relatively speaking. The evidence here is weak and my confidence intervals are wide.

## Productivity

The evidence is somewhat better for typing speed, though the earliest studies (think 1930s to 1950s) are riddled with methodological flaws (e.g. non-random control groups) and conflicts of interest (e.g. being conducted by August Dvorak himself).

A difficulty in comparing QWERTY typing speeds with those of alternative designs is that nearly everybody knows QWERTY well and the alternatives not at all. So how are you going to get a fair test?

One way is to only look at the most talent typers for each layout. You might not have known that there is an Ultimate Typing Championship, for example. In the last edition to take place, in 2020, exactly one of the 26 competitors used Dvorak, and the other 25 used QWERTY; the lonely Dvorak user ended up in 12th place. On one well-known

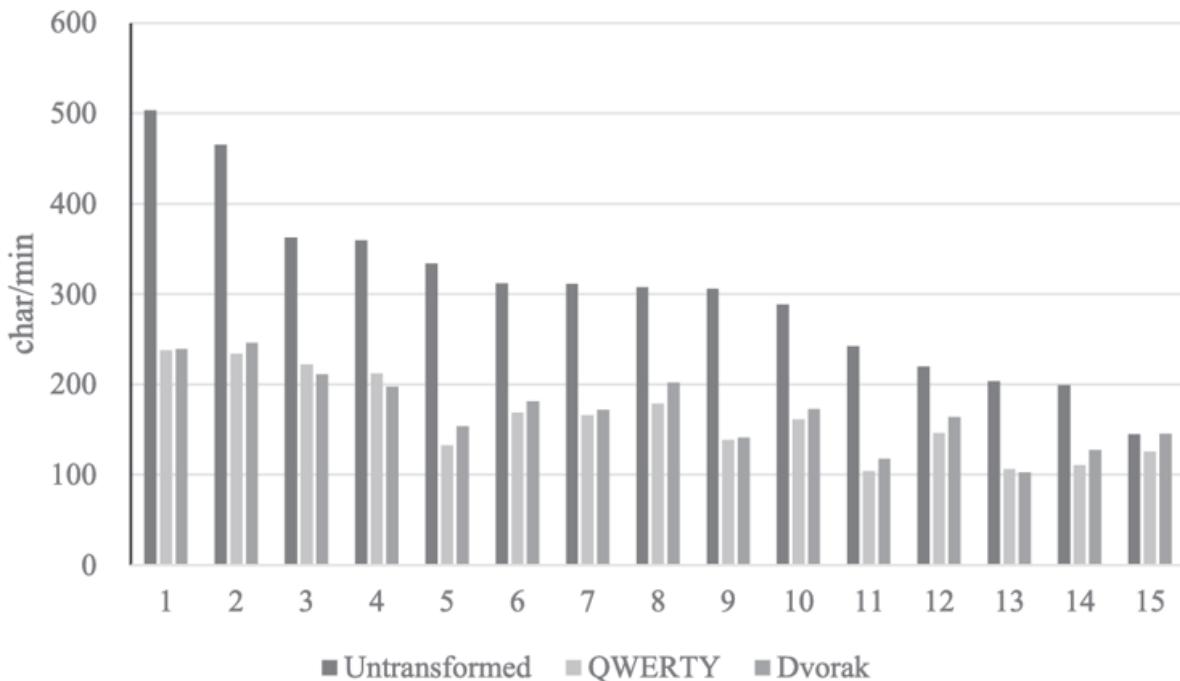
[leaderboard](#) that same Dvorak user is in the top 30 and two Colemak users are in top ten and top 30 respectively; the rest of the top 30 all use QWERTY, I think. That sounds pretty good for QWERTY until you remember that the vast, vast majority of people have grown up and are used to using it, so given that the transaction cost of switching to an alternative layout like Dvorak is high (and more on this later), ten percent of the top thirty using alternative layouts is actually pretty significant! On the other hand, if Dvorak and other layouts really do have an edge over QWERTY, you might have expected most of the ultra-competitive top 30 to have switched over by now. (Anecdotally, it seems some competitors have achieved better word per minute high scores after switching.)

There are other ways of circumventing the problem of varying skill. Three studies carried out in three different decades with three different methodologies show very similar results.

Kinkead (1975) recorded 22 people doing standardised typing tests (presumably on QWERTY keyboards).<sup>[25]</sup> He then looked to see which digrams were faster/slower to type and estimated how fast typing would have been on a Dvorak keyboard (assuming that digram typing speed remains constant for two physical keys, even if the characters they represent change).<sup>[26]</sup> He found that the Dvorak keyboard would need around 2.6% less time to type the same thing as the QWERTY keyboard, however the "ideal" keyboard (based on the recorded data) would be 7.6% faster than QWERTY.<sup>[27]</sup> I am somewhat sceptical of this study, partly because of the small sample and partly because I don't know if it captures all the ways in which Dvorak could be better than QWERTY (e.g. maybe some trigrams are significant in a way that considering only digrams doesn't capture).

I don't have access to the original paper, but in Liebowitz & Margolis, the authors note a study by Norman & Rumelhart (1983) that found a 5% typing speed advantage for Dvorak over QWERTY, and a 2%-9% advantage for QWERTY over alphabetically sorted layouts.<sup>[28]</sup> They write: "For the expert typist, the layout of keys makes surprisingly little difference. There seems no reason to choose [QWERTY], Dvorak or alphabetically organized keyboards over one another on the basis of typing speed. It is possible to make a bad keyboard layout, however, and two of the arrangements that we studied [not sure which layouts they refer to here] can be ruled out."<sup>[29]</sup>

Cho (2014) presents an interesting study design.<sup>[30]</sup> Remember the problem of people already knowing QWERTY really well, making it hard to do a fair comparison. The trick Cho does is basically to (1) transpose words typed on Dvorak to a QWERTY keyboard (so the same physical keys need to be struck, though all the words are now gibberish) and (2) do something similar for QWERTY, except transposing the QWERTY words to a mangled version of QWERTY that more or less retains the same characteristics.<sup>[31]</sup> I think this is smart and does a pretty good job of testing the designs while reducing familiarity effects. The main result is in this chart. (Numbers one to 15 represent the 15 test subjects; "untransformed" refers to QWERTY as we know it, whereas "QWERTY" refers to the mangled version.<sup>[32]</sup>)



As far as I can tell, that shows three out of the 15 typing faster with the mangled QWERTY, and the other 12 typing faster with the transposed Dvorak. But the difference looks small (around 5% which is similar to the earlier studies). The author does a quick Fermi estimate:

Assuming a person [types] 1,000 words a day, typing will solely take about 15 minutes a day. If that person can save 10% of typing time by changing to Dvorak, he or she can save 1.5 minutes a day, 45 minutes a month, and finally 9 hours a year. It means this time saving might be overwhelmed by the cost of changing the layout. [\[33\]](#)

How much time do we actually spend typing? Xah Lee logged his keyboard use for two years and [found](#) that he typed roughly one hour per day (assuming 50 words per minute). Now, that is likely an underestimate as it only records the stuff he wrote in Emacs, but he also reports spending 13 hours a day in front of the computer, much of it likely in Emacs. Most people probably type significantly less. A 5% improvement for Xah Lee would save him around 18 hours per year, which is not bad, but also the upper end of what one might expect.

In sum, I think there's a good chance that touch-typists, after a period of relearning, can type faster with Dvorak and Colemak than with QWERTY. But I also think switching over would be unlikely to lead to noticeable productivity boosts (with perhaps a few exceptions).

## Concerns

I'm sure by now you're itching to point something out. Before you do, let me try to security test my own review.

## **What's the Cost of Transition?**

Some people find switching to a new key layout a real struggle. You have to rewire your muscle memory, and QWERTY is wired real hard. Studies I found report anywhere from two and 24 days to regain one's typing speed.<sup>[34]</sup> When I switched from QWERTY to Colemak, I did it cold turkey and typed all day in the new layout, and it still took me something like a week until I was comfortable and two weeks until I had regained my former speed. This is definitely a drawback of switching, though not an insurmountable one. Note however that this cost of transitioning does not make QWERTY a better key layout – it only makes sticking with QWERTY a better choice than it would have been otherwise.

## **What's the Cost of Not Sticking to the Universal Standard?**

You might think, "Isn't it annoying to use an alternative design when the whole world was built with QWERTY in mind?" I think the concern here is that you might unlearn QWERTY but still occasionally need to use it, for example when borrowing a computer or typing on a phone. I've tried to think of ways that this could be a serious problem but cannot. For some reason it's easy for me to switch between Colemak on the computer and QWERTY on the phone. Colemak and Dvorak both come pre-installed on Mac and most Linux distros, and seem easy enough to install on Windows. And if I just need to type something out quickly on a QWERTY keyboard, I usually manage ok, and if not, the keys are all there, anyway.

## **What If You're Bilingual?**

Alternative designs like Colemak and Dvorak were optimised for typing English, but many people do most of their typing in multiple languages. What should they (we, actually) do? QWERTY has lots of variants with special characters for different languages; alternative designs are usually one-size-fits-all solutions where special characters are typed via so-called dead keys (in other words, via multiple keys in succession). (Of course it's possible to customise one's key layout, but that's only an option for technical people with time on their hands, like me.)

I think the answer here is: it depends. For me, using Colemak for Swedish and German in addition to English has been fine, but I do most of my typing in English, and Swedish and German are both Germanic languages that have similar letter frequencies to English's and few additional characters. If you are sometimes writing Turkmen, say, you may have to switch between layouts if you want to use something like Colemak or Dvorak for English. There seem to be few modern layouts optimised specifically for non-English languages.

## **Conclusion**

In a real sense, the stakes here are quite high. Even if alternative key layouts only provide a small benefit, and even if the transaction cost is too high to make it worthwhile for any one individual to switch, the costs for society could be significant. There are billions of people out there typing suboptimally, all day, every day. If we were ever given a clear sense of the loss involved in doing something slightly suboptimally over vast periods of time, we might tremble and rue those lost riches. But as it is, the effects are too diffused for us to see them.

I have been hedging a lot in the preceding sections. That is because I am still uncertain about much of this, mostly due to the dearth of empirical evidence. But if you are a believer in revealed preferences, know this. While researching this article, I came across [the QWYRFM layout](#), a product of the aforementioned Carpalx project. One week or so ago I switched from Colemak to typing on QWYRFM due to the reduced load it places on the pinkies. I have been typing on it since then, both on and off the clock. In fact, I have written most of this review on QWYRFM. And not only am I not finding it a struggle to switch, but I am even sort of enjoying the experience.

---

1. I will only look at RSI here. Obviously there is more to ergonomics than that, but RSI seems widespread and bad and well-defined enough that it's worth focusing on. [←](#)
2. As with ergonomics, there is more to key layout and productivity than typing speed (maybe it's even advantageous to type in a slow and steady pace rather than in quick bursts), but speed seems like a significant component of productivity, so I will focus on it here. [←](#)
3. Yasuoka, K., & Yasuoka, M. (2011). On the Prehistory of QWERTY. *Zinbun*, 42, 161-174. [←](#)
4. *ibid.* [←](#)
5. *ibid.* [←](#)
6. Liebowitz, S. J., & Margolis, S. E. (1990). The fable of the keys. *The Journal of Law and Economics*, 33(1), 1-25. [←](#)
7. Shieh, K. K., & Lin, C. C. (1999). A quantitative model for designing keyboard layout. *Perceptual and motor skills*, 88(1), 113-125. [←](#)
8. *ibid.* [←](#)
9. But note that there really are a lot of judgment calls here. As Xah Lee [puts it](#): "When you study keyboard layout as Carpalx did, one thing quickly becomes obvious. The difficulty in creating the 'most efficient' [layout] isn't about computing distances or gathering statistics, but making subtle judgement on what's considered more efficient. For example, is minimizing distance of finger travel more important than alternating hands? How [much importance] do we assign to ease of bigram [typing]? What weight should we attach to avoiding single-finger repetition? Should the key J and K have [an] ease-of-press score difference of 0.1 or 0.2?" [←](#)
10. Van Tulder, M., Malmivaara, A., & Koes, B. (2007). Repetitive strain injury. *The Lancet*, 369(9575), 1815-1822. [←](#)
11. Fagarasanu, M., & Kumar, S. (2003). Carpal tunnel syndrome due to keyboarding and mouse tasks: a review. *International Journal of Industrial Ergonomics*, 31(2), 119-136. [←](#)
12. *ibid.* [←](#)
13. *ibid.* [←](#)
14. *ibid.* [←](#)

15. Barcenilla, A., March, L. M., Chen, J. S., & Sambrook, P. N. (2012). Carpal tunnel syndrome and its relationship to occupation: a meta-analysis. *Rheumatology*, 51(2), 250-261. [←](#)
16. Battista, E. B., Yedula, N. R., Koolmees, D. S., Montgomery, Z. A., Ravi, K., & Day, C. S. (2021). Manufacturing Workers Have a Higher Incidence of Carpal Tunnel Syndrome. *Journal of occupational and environmental medicine*, 63(3), e120-e126. [←](#)
17. Wright, A. R., & Atkinson, R. E. (2019). Carpal tunnel syndrome: An update for the primary care physician. *Hawai'i journal of health & social welfare*, 78(11 Suppl 2), 6. [←](#)
18. Battista, E. B., Yedula, N. R., Koolmees, D. S., Montgomery, Z. A., Ravi, K., & Day, C. S. (2021). Manufacturing Workers Have a Higher Incidence of Carpal Tunnel Syndrome. *Journal of occupational and environmental medicine*, 63(3), e120-e126. [←](#)
19. Thomsen, J. F., Gerr, F., & Atroshi, I. (2008). Carpal tunnel syndrome and the use of computer mouse and keyboard: a systematic review. *BMC musculoskeletal disorders*, 9(1), 1-9. [←](#)
20. Wright, A. R., & Atkinson, R. E. (2019). Carpal tunnel syndrome: An update for the primary care physician. *Hawai'i journal of health & social welfare*, 78(11 Suppl 2), 6. [←](#)
21. Wærsted, M., Hanvold, T. N., & Veiersted, K. B. (2010). Computer work and musculoskeletal disorders of the neck and upper extremity: a systematic review. *BMC musculoskeletal disorders*, 11(1), 1-15. [←](#)
22. Trillo-Chacón, M. C., Castillo-M, J. A., Tolosa-Guzman, I., Medina, A. F. S., & Ballesteros, S. M. (2021). Strategies for the prevention of carpal tunnel syndrome in the workplace: A systematic review. *Applied Ergonomics*, 93, 103353. [←](#)
23. Pan, C. S., & Schleifer, L. M. (1996). An exploratory study of the relationship between biomechanical factors and right-arm musculoskeletal discomfort and fatigue in a VDT data-entry task. *Applied ergonomics*, 27(3), 195-200. [←](#)
24. Feng, B., Chen, K., Zhu, X., Ip, W. Y., Andersen, L. L., Page, P., & Wang, Y. (2021). Prevalence and risk factors of self-reported wrist and hand symptoms and clinically confirmed carpal tunnel syndrome among office workers in China: a cross-sectional study. *BMC Public Health*, 21(1), 1-10. [←](#)
25. Kinkead, R. (1975, October). Typing speed, keying rates, and optimal keyboard layouts. In *Proceedings of the Human Factors Society Annual Meeting* (Vol. 19, No. 2, pp. 159-161). Sage CA: Los Angeles, CA: SAGE Publications. [←](#)
26. ibid. [←](#)
27. ibid. [←](#)
28. Liebowitz, S. J., & Margolis, S. E. (1990). The fable of the keys. *The Journal of Law and Economics*, 33(1), 1-25. [←](#)
29. ibid. [←](#)

30. Cho, H. (2014). *Comparing QWERTY and Dvorak Keyboard Speed: a Pilot Study* (Doctoral dissertation, 서울대학교 대학원). [←](#)
31. ibid. [←](#)
32. ibid. [←](#)
33. ibid. [←](#)
34. Liebowitz, S. J., & Margolis, S. E. (1990). The fable of the keys. *The Journal of Law and Economics*, 33(1), 1-25. [←](#)

# ML Systems Will Have Weird Failure Modes

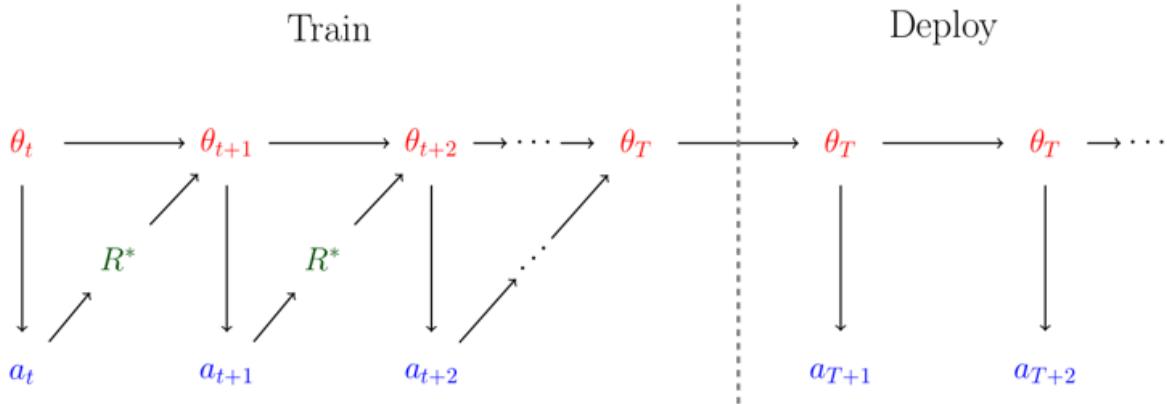
Previously, I've argued that future ML systems might exhibit [unfamiliar, emergent capabilities](#), and that thought experiments [provide one approach](#) towards predicting these capabilities and their consequences.

In this post I'll describe a particular thought experiment in detail. We'll see that taking thought experiments seriously often surfaces future risks that seem "weird" and alien from the point of view of current systems. I'll also describe how I tend to engage with these thought experiments: I usually start out intuitively skeptical, but when I reflect on emergent behavior I find that some (but not all) of the skepticism goes away. The remaining skepticism comes from ways that the thought experiment clashes with the ontology of neural networks, and I'll describe the approaches I usually take to address this and generate actionable takeaways.

## Thought Experiment: Deceptive Alignment

Recall that the [optimization anchor](#) runs the thought experiment of assuming that an ML agent is a perfect optimizer (with respect to some "intrinsic" reward function  $R$ ). I'm going to examine one implication of this assumption, in the context of an agent being trained based on some "extrinsic" reward function  $R^*$  (which is provided by the system designer and not equal to  $R$ ).

Specifically, consider a training process where in step  $t$ , a model has parameters  $\theta_t$  and generates an action  $a_t$  (its output on that training step, e.g. an attempted backflip assuming it is being trained to do backflips). The action  $a_t$  is then judged according to the extrinsic reward function  $R^*$ , and the parameters are updated to some new value  $\theta_{t+1}$  that are intended to increase  $a_{t+1}$ 's value under  $R^*$ . At some point, the model is then deployed with final parameters  $\theta_T$ , and continues to take actions. The following diagram illustrates this process:



Now, let's assume that the model  $\theta_t$  is a perfect optimizer whose objective is to maximize the discounted value of an intrinsic reward  $R \neq R^*$ . That is,  $\theta_t$  picks the action  $a_t$  satisfying

$$a_t = \operatorname{argmax}_a E[\sum_{s=0}^{\infty} \gamma^{-s} R(a_{t+s}) \mid a_t = a].$$

(I know that this is an unrealistic assumption. We'll examine the assumption in detail in the next section, but for now please grant it even if it requires suspending disbelief.)

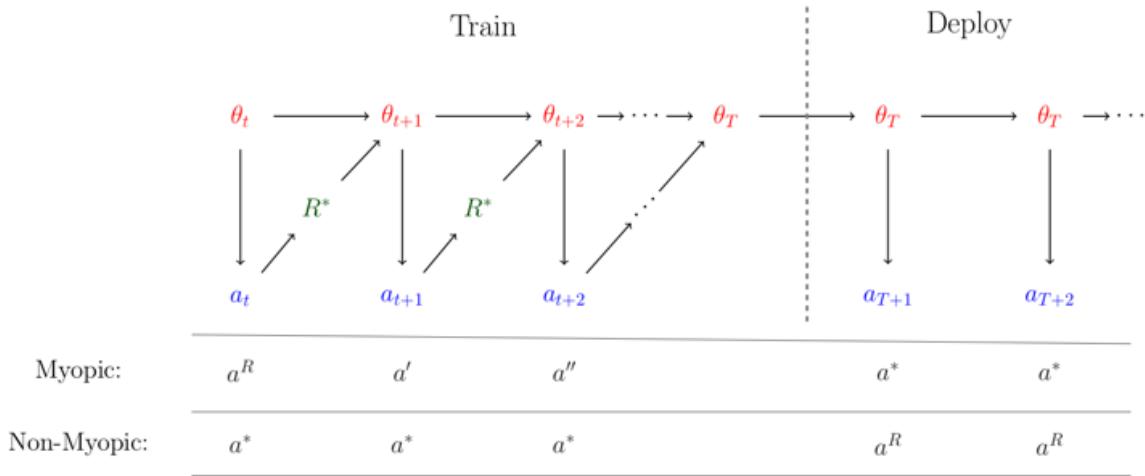
What action  $a_t$  will  $\theta_t$  pick? Let's define  $a^R = \operatorname{argmax}_a R(a)$  and  $a^* = \operatorname{argmax}_a R^*(a)$ --

that is,  $a^R$  maximizes the (instantaneous) intrinsic reward while  $a^*$  maximizes the extrinsic reward.

Assuming that  $\gamma$  is close to 1, the model wants  $a^R$  to be picked on as many future rounds as possible (including at deployment time). If  $\theta_t$  plays myopically, this means it would pick  $a_t = a^R$ . However, this is not the optimal non-myopic action: if  $a_t = a^R$ , then  $\theta_{t+1}$  will be updated towards producing outputs less like  $a^R$  and more like  $a^*$ , so outputting  $a^R$  in the current round leads to fewer  $a^R$  outputs in future rounds.

Instead, I think the optimal non-myopic action is approximately the following: during training time, output  $a^*$  on every time step, so that the training process doesn't update  $\theta$  (since it's already optimal) and hence  $\theta_t = \theta_{t+1} = \dots = \theta_T$ . This guarantees that the deployed agent is  $\theta_t$ , which is now free to output  $a^R$  on every time step without risk of being modified by the training process.<sup>[1]</sup>

The myopic and non-myopic outcomes are illustrated in the figure below:



*Myopic and non-myopic agents with an intrinsic reward, R. The myopic agent plays high intrinsic-reward actions during training, but is eventually modified to output action a\* with high extrinsic reward R\*. The non-myopic agent instead plays a\* during training so that it can pursue its intrinsic reward at deployment, by playing a<sup>R</sup> at all post-deployment time steps.*

To put it another way—the model’s behavior will look aligned with  $R^*$  during the training process, but will switch to being non-aligned once deployed: so-called “deceptive alignment”. This behavior is unintuitive to me, since it shows that a system’s behavior can shift significantly between train and test, even in the absence of an obvious distribution shift between train and test outputs [2].

## Engaging with Deceptive Alignment

When I first heard the above argument, I thought it was pretty crazy and implausible, mainly because my intuition said this “just wasn’t how ML systems worked”. When I think about why I feel that way, I realize it’s because the scenario invokes capabilities that ML is currently bad at: long-term planning and understanding complex features of the environment (i.e. the training process and its ramifications). However, emergence implies that these properties could easily appear in the future, even without explicit design [3]. As a result, I’ve come to discount this particular intuition.

However, I do think there are subtler reasons to think the deceptive alignment story won’t play out as written. Here are a few:

1. It’s not clear why the model  $\theta$  would come to be optimizing a reward function  $R$  in the first place. Yes, it is the case that deceptively aligned models achieve the global minimum of training loss, so in that sense they are incentivized by the training process. But so is an actually aligned model, so which one you end up with has to depend on the inductive bias of the training process.
2. Reward functions are simpler than policies and typically learned faster. So by the time the system is smart enough to have long-term plans, it will already have a very good representation of its intended reward function. We thus might hope

that most of the model's internal representations are devoted to achieving high reward in a straightforward manner rather than through long-term deception.

3. To the extent that a model is not aligned, it probably won't be the case that it's deceptively aligned with an explicit reward function  $R$ --that's a very specific type of agent and most agents (including humans) are not maximizing any reward function, except in the trivial sense of "assign reward 1 to whatever it was going to do anyway, and 0 to everything else".
4. Deceptive alignment is a specific complex story about the future, and complex stories are almost always wrong.

I find these points persuasive for showing that deceptive alignment as *explicitly written* is not that likely, but they also don't imply that there's nothing to worry about. Mostly they are an argument that your system might be aligned and might be misaligned, that if it is misaligned it won't be *exactly* in the form of deceptive alignment, but ultimately what you get depends on inductive bias in an unknown way. This isn't particularly reassuring.

**What I take away from thought experiments.** Per the discussion above, the failure mode in my head is not "deceptive alignment as written above". Instead it's "something kind of like the story above but probably different in lots of details". This makes it harder to reason about, but I think there are still some useful takeaways:

- After thinking about deceptive alignment, I am more interested in supervising a model's process (rather than just its outputs), since there are many models that achieve low training error but generalize catastrophically. One possible approach is to supervise the latent representations using e.g. interpretability methods.
- While I don't think neural nets will be literal optimizers, I do think it's likely that they will exhibit "drives", in the same way that humans exhibit drives like hunger, curiosity, desire for social approval, etc. that lead them to engage in long-term coherent plans. This seems like enough to create similar problems to deceptive alignment, so I am now more interested in understanding such drives and how they arise.
- Since deceptive alignment is a type of "out-of-distribution" behavior (based on the difference between train and deployment), it has renewed my interest in understanding whether larger models become more brittle OOD. So far the empirical evidence is in [the opposite direction](#), but deceptive alignment is an argument that asymptotically we might expect the trend to flip, especially for tasks with large output spaces (e.g. policies, language, or code) where "drives" can more easily manifest.

So to summarize my takeaways: be more interested in interpretability (especially as it relates to training latent representations), try to identify and study "drives" of ML systems, and look harder for examples where larger models have worse OOD behavior (possibly focusing on high-dimensional output spaces).

**Other weird failures.** Other weird failures that I think don't get enough attention, even though I also don't think they will play out as written, are Hubinger et al.'s [Risks from Learned Optimization](#) (AI acquires an "inner objective", somewhat similar to deceptive alignment), and Part I of Paul Christiano's [AI failure story](#) (the world becomes very complicated and AI systems create elaborate Potemkin villages for humans).

Paul Christiano's story in particular has made me more interested in understanding how reward hacking interacts with the sophistication of the supervisor: For instance, how much more readily do neural networks fool humans who have 5 seconds to think, vs. 2 minutes or 30 minutes? I more generally want to understand how reward hacking

depends quantitatively on both supervision quality and model capacity (qualitatively, we expect higher quality → less hacking and higher capacity → more hacking).

Understanding this quantitative relation would help ground Paul's story, since he imagines a world where humans have built extremely sophisticated systems for supervising ML models, but eventually the ML models become even more powerful and game the supervision signal anyways.

## What To Do About Weird Emergent Failures

When thinking about how to handle emergent risks, I often reflect on the example of uranium. For context, an atomic bomb is pretty much just a bunch of uranium put together---once you get enough, the reaction becomes self-sustaining---making it a good example of More Is Different.

The first nuclear reaction (not a bomb, but a [pile of uranium](#) in an abandoned football stadium in Chicago) was engineered by Enrico Fermi. The reaction required 12,400 pounds of uranium metal piled 57 layers high. Left unsupervised, a 57-layer pile would consume itself within two hours and kill everyone in the vicinity. On the other hand, a 56-layer pile would do nothing.

Fermi had a good understanding of nuclear physics and understood, from careful monitoring and underlying theory, that the pile would pass the critical threshold between layers 56 and 57. He also knew that cadmium rods would absorb neutrons and strongly inhibit the reaction. These rods were set up and the entire apparatus was carefully controlled to go only slightly supercritical. He brought the reaction to half a watt for several minutes before shutting it back down (see [The Making of the Atomic Bomb](#), pp. 524).

With AI, we currently lack both Fermi's conceptual understanding of the underlying risk factors and his ability to continuously measure them. We have neither a cadmium rod nor a measure of reaction criticality. But I think we can get there, by combining these weird thought experiments with [carefully chosen empirical experiments](#), which will be the topic of the next post.

---

1. Things are more complicated in reality, since  $\theta_t$  is updated even when  $a_t$  is optimal (due to noise in the training process). However, we'll ignore this for purposes of the example. [←](#)
2. Of course, there is still some distribution shift, since the agent can observe whether it is being trained or deployed. But this is a relatively minor and unintuitive shift compared to what is typically studied. [←](#)
3. Of course, emergence doesn't mean that we can just predict whatever we want---we'd need some reason to expect these specific capabilities to emerge. Long-term planning and environmental awareness are both useful for a wide variety of tasks, making them likely to emerge when training powerful models on a diverse data distribution. [←](#)

# **Luna Lovegood and the Fidelius Curse**

## **- Part 2**

"Then suddenly all the Mirror of Atlantis showed were the reflections of me and Harry Potter," said Luna.

Fay glared jealously over at Harry James Potter-Evans-Verres.

"He doesn't remember me," said Luna.

"What a horrible life you live. I feel so sorry for you," teased Fay. She bonked Luna with her golden plate.

"There must be a perfectly supernatural explanation for your condition. We'll see what information the library has on it," said Luna.

"But the Hogwarts Library doesn't have a section on Repelling Charms. I even looked in the Restricted Section. That's what I got the second detention for," said Fay.

Luna rolled her eyes at Fay. Luna wished she had brewed a pot of Comed-Tea.

"Oh," said Fay.

There was a new teacher at the Head Table. The new teacher wore dueling armor. Otherwise, her robes followed every Hogwarts rule, standard and regulation. The witch moved slowly and steady like a turtle. A snapping turtle.

Tink, tink, tink. The Headmistress called for quiet. There were all the usual announcements and then....

"There will be no Defense Against the Dark Arts teacher this year," announced Headmistress McGonagall.

The students gasped. Luna inventoried the teachers. Nobody was missing from last year besides the handsome brave gallant Professor Lockhart. He was not afraid to die, brave oh brave Professor Lockhart.

Fay sat on Luna's right. Luna whispered to Fay, "What do you think of the new Defense Professor?"

"Did you not hear what the Headmistress just said?" said Fay.

Luna shrugged. She pointed to the High Table. "Right there. There's a new teacher between Professor Sinistra and Lady Yue. No one is missing so she must be the new Defense Against the Dark Arts professor" said Luna.

No one (besides Fay) wanted to sit next to Luna. Colin Creevey sat two seats Luna's left. "Who are you whispering to?" he said to Luna.

"My friend," said Luna.

Colin Creevey took a careful look of Luna as if Fay's seat was empty. "There's nobody there," he said.

Padma Patil sat to the left of Colin Creevy. "Luna has an imaginary friend," Colin reported to Padma.

Fay flicked a Brussels sprout at Colin.

"I was just kidding!" objected Colin.

Luna's imaginary friend rolled her eyes.

---

Luna's first Defense Against the Dark Arts class was the next morning. Luna tried to ask the other students what they thought of the new Defense Professor. Their answers were a tangle of contradictions. The students all agreed that there was no Defense Professor even though they all had Defense Against the Dark Arts class.

The Defense Professor tapped her podium.

"Hello everyone. My name is Martina Memnuela but I don't expect any of you to remember that. You may refer to me as 'the Defense Professor' even though the thoroughest defense is a preemptive offence," said Defense Professor Memnuela.

Fay raised her hand.

"Reality is precarious. The Forces of Evil work tirelessly to undermine consensus reality. Most of you will live your lives in blissful indifference of the complex systems protecting you. A few of you, the Best of the Best, will join the first, last and only line of defense against existential threats. The security we provide is what Defense is about. Everything else is just paperwork," said the Defense Professor.

Xena Smith raised her hand. Professor Memnuela called on her.

"Are you an Unspeakable?" said Xena.

"I can neither confirm nor deny whether I work for the Department of Mysteries," said Professor Memnuela, "But I hope someday to introduce a few of you to the secrets contained therein."

There were scattered chuckles. Fay's hand was still up. The Defense Professor continued to not respond to it.

"I have a question," Fay said.

"Wait until you are called upon," said Professor Memnuela.

"You're never going to call on me," Fay said.

"Five points from Ravenclaw," said Professor Memnuela.

"I am asking a question. Asking questions is what Ravenclaws are supposed to do," said Fay.

"Detention," said Professor Memnuela.

Fay rolled her eyes. She passed a note to Luna. *Why can we remember her name when nobody else can?* Luna raised her hand. Professor Memnuela called on Luna immediately. Luna almost asked Fay's question before she realized that the Defense Professor was a

dangerous person, possibly evil, and that it would be imprudent for Luna to reveal that she knew the true name of an Unspeakable who had teachers' privileges to Hogwarts' wards.

"Did you have a question?" said Professor Memnuela.

Luna blurted the second thing off the top of her mind. "Where can I find information about Repelling Charms?" said Luna.

Professor Memnuela paused as if composing a riddle. Luna deduced that the Unspeakable had an answer memorized and was just pretending to deliver it impromptu. It was a trap. Luna resolved to ignore the Unspeakable's misdirection.

"When you stare into the void, the void stares back," said the Unspeakable.

*Sorry. I'll explain later.* Luna wrote to Fay. Professor Memnuela noticed Luna writing a note to herself. That got the Unspeakable's attention. The Unspeakable approached her slowly. Luna averted her eyes.

"Are you hiding something, young one?" said the Unspeakable.

Luna's eyes darted around the room, looking anywhere but at the Unspeakable's face. This would be easier if she were wearing her Spectrespecs.

"Look at me or I will compel you to do so," said the Unspeakable.

Luna stared defiantly into the Unspeakable's eyes. The Legilimens' fingers held Luna's eyelids open. There would be no official recourse. There was no Defense Against the Dark Arts Professor this year.

"Wow. I am impressed. I have never seen such a perfect Occlumens so young," said Professor Memnuela.

"I'm not an Occlumens," said Luna.

"You sincerely expect me to believe that Wrackspurts and Crumple-Horned Snorkacks are real, that you discovered the Lost Diadem of Ravenclaw and that you joined forces with the handsome Gilderoy Lockhart to face down You-Know-Who?" said Professor Memnuela.

Luna had gotten distracted by what she thought might be Wrackspurts floating in an upper corner of the classroom.

"Miss Lovegood?" said Professor Memnuela.

"No. Professor Lockhart and I joined forces with You-Know-Who too. The three of us worked together," said Luna. The Wrackspurt slipped out of her ear.

"You are an impressive actor. I admire your imagination, your sense of humor and your dedication to the role. That persistence will serve you well. But, please. I have a serious class to teach," said Professor Memnuela.

Luna was too stunned to retort. She sat in her chair until long after everyone else except Fay had left the classroom.

# **Excessive Nuance and Derailing Conversations**

## **Overview**

I've observed my interlocutors—and sometimes myself—applying excessive nuance to irrelevant points during some discussions. This misplaced nuance results in the derailing of conversations towards rabbit holes and dead ends.

## **High-Level Description**

When I bring a new supporting but tangential idea into a discussion (e.g. mention a widely accepted scientific theory), my interlocutor applies excessive nuance to this new idea at the expense of the salient points in the discussion. This misplaced and excessive nuance takes the form of nitpicking or questioning the underlying framework underpinning that newly introduced idea.

This is rarely fruitful because almost always, none of the participants is at the cutting-edge of the relevant fields to add any new insights. I'm all for rational inquiry and open discourse, but nitpicking oftentimes comes across as pretentious and unconstructive.

The effect of this conversation style is the unchecked growth of the stack of topics. We often forget to go back down the stack to revisit the original points that instigated the discussion unless someone cares sufficiently to intentionally apply enough pressure to steer the conversation back down the stack. Such discussions usually quickly end up in rabbit holes, and no one gets anything out of them.

## **Example**

Suppose that you are having a conversation about "what it feels like when your worldview is shattered." Your interlocutor is a fellow rationalist and mentions that they read a book promoting climate denial and describes to you what it felt like to almost have their worldview shattered by a professional motivated-skeptic and evidence-cherrypicker. They explain how difficult it is for us mere mortals to notice this black magic being applied to our minds and how easy it is to be deceived by a professional charlatan.

To add to this discussion, you contribute the idea that this feeling is similar going the other way. For example, a profoundly religious person reading a science book on biological evolution will feel similar to how you—a rigorous rationalist—felt reading a book promoting climate denial, anti-vax, or some other pseudoscientific theory.

Both of you agree that evolution is a widely accepted theory with plenty of evidence going for it. Still, your interlocutor decides that they want to show you how much they know about the philosophy of science, so they say, "Playing devil's advocate, evolution is still just a theory." In the back of your mind, you would like to stay on

topic, but rationalist verve takes over you, and you engage with all your heart in a fraught debate.

Now, suddenly, you find yourself debating the truth of biological evolution. You're both embroiled in the weeds of the philosophy of science, and other participants in the conversation are becoming visibly uninterested. You are no longer talking about "what it feels like when your worldview is shattered."

## Conversational Etiquette

### Conversation and Social Status

I read "The Elephant in the Brain" recently, and at some point, it talks about human conversation. Hanson and Simler explore why we humans are so eager to talk even when keeping information to ourselves gives us a competitive advantage.

We're so eager to speak, in fact, that we have to curb our impulses via the norms of conversational etiquette. If speaking were an act of giving, we would consider it polite for people to "selflessly" monopolize conversations. But in fact, it's just the opposite. To speak too much or "hog the mic" is considered rude, while the opposite behavior—inviting someone else to take the floor, or asking a dinner guest about one of her hobbies—is considered the epitome of good manners.

Improving conversational etiquette has a very high return on investment because we converse with others all the time. Giving others space to talk, listening when others speak, clearly articulating ourselves and avoiding nitpicking when inappropriate in context are all valuable skills. These skills develop naturally for some people but are subtle and difficult to cultivate for others. I have found that inviting friends to offer me constructive feedback on my conversational style has helped me improve these skills.

I find that there are three overarching types of conversationalists:

- High-status conversationalists:
  - They invite others to speak.
  - They check if others would like to switch topics, go back to a previous topic or end the conversation, so they don't miss their train.
  - When they speak, they are precise and pertinent. They are OK with letting go and don't need to explain every point exhaustively.
  - They often get invited by others to take centre-stage, and people are all ears when they speak.
- Low-status conversationalists:
  - They make every point as if they are answering a test, going into excessive detail that is usually not relevant in context.
  - They are excessively contrarian, even about minor or tangential points, thus being inimical to the progress of a conversation.
  - They don't give other people space to talk and don't make an effort to include newcomers.
  - They complain and forcefully change conversation topics at the expense of others when a particular subject doesn't suit their fancy.
- Non-conversationists:
  - They don't like to converse.

## **Stack Overflow**

A conversational stack overflow occurs when the stack gets so large and messy that everyone involved in a conversation forgets [how we got here](#). I argue that in many situations, traversing back down the stack is an essential social skill because there is at least someone in the conversation who cares about some primary point that was made earlier in the discussion.

## **When Stack Overflow and Nuance are OK**

There are specific conversations where allowing the conversation to meander and allowing the conversational stack to grow and blend are OK. For example, you are having beers next to a campfire, or when you explicitly ask other participants in the conversation if they would like to revisit an earlier point, switch topics or end the conversation, and they say they are happy to let the conversation flow.

Another context in which exploring many topics is desirable is when getting to know someone new. In this context, traversing a breadth of topics quickly and having flexibility in which topics to discuss will maximise the chances of finding a common interest.

Rigour is desirable in many contexts, such as writing an academic paper, reviewing academic work or explicitly engaging in a debate about a particular topic. Rigour is different from nitpicking. Nitpicking takes the form of applying excessive nuance in order to miss the point intentionally, or to show off how much you know about some irrelevant topic.

## **Concluding Remarks**

The constant shifting of conversation topics due to the application of excessive nuance to tangential points can be counterproductive in many contexts. There is a suitable place and time for nuance, and some discussions don't call for nuance about particular matters. Hogging the mic to make a point exhaustively is a poor social skill. Improving conversational social skills is a meta-skill that enhances all aspects of our lives, and it is worth intentionally cultivating.

# Apply for research internships at ARC!

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

**(Update: we are no longer accepting applications for interns.)**

In addition to [hiring full-time researchers](#), ARC is also hiring for 1-3 month internships. We're open to interns for summer 2022 in particular, but if a different time works for you that's fine too.

Internships are appropriate for undergraduates and graduate students, or for anyone considering a career change.

To apply for an internship, you can use [this application](#) and check the "internship" box. If you submitted an application previously there's no need to resubmit; we'll clarify if there is ambiguity.

Salary for interns is \$15k/month. We encourage interns to work from our Berkeley office especially at the start of their internship.

(The rest of this post is copied from our [previous hiring post](#).)

## What is ARC?

ARC is a non-profit organization focused on theoretical research to align future machine learning systems with human interests. We are aiming to develop alignment strategies that would continue to work regardless of how far we scaled up ML or how ML models end up working internally.

Probably the best way to understand our work is to read [Eliciting Latent Knowledge](#), a report describing some recent and upcoming research, which illustrates our general methodology.

We currently have 2 research staff (Paul Christiano and Mark Xu). We're aiming to hire another 1-2 researchers in early 2022. ARC is a new organization and is hoping to grow significantly over the next few years, so early hires will play a key role in helping define and scale up our research.

## Who should apply?

Most of all, you should send in an application if you feel excited about proposing the kinds of algorithms and counterexamples [described in our report on ELK](#).

We're **open to anyone who is excited about working on alignment even if you don't yet have any research background** (or your research is in another field). You may be an especially good fit if you:

- Are creative and generative (e.g. you may already have some ideas for potential strategies or counterexamples for ELK, even if they don't work).
- Have experience designing algorithms, proving theorems, or formalizing concepts.

- Have a broad base of knowledge in mathematics and computer science (we often draw test cases and counterexamples from these fields).
- Have thought a lot about the AI alignment problem, especially in the limit of very powerful AI systems.

Hiring will be a priority for us in early 2022 and we don't mind reading a lot of applications, so please err on the side of sending in an application even if you're not sure you'll be a fit!

# Nudging My Way Out Of The Intellectual Mosh Pit



As part of my ongoing [effort](#) to [improve](#) my cost/benefit ratio on social media, I'm nudging myself away from intellectual mosh pit platforms like Facebook and Twitter, and towards blog posts, articles, videos essays, etc. *Really* longform consumption (e.g. books) remains about the same, my limitations on that are mostly my insistence on fact-checking nonfiction and very narrow tastes in fiction, this post is about changing what I reach for when I'm bored in a line. Towards that goal I have made a few changes, which I list here roughly in ascending order of how much effort they were:

- Put all of my screens in greyscale. If you only try one thing on this list, make it this one. It takes five seconds to test once you have instructions, and the relief for me was immediate and has lasted days so far. Every time I have to turn color on to look at graphs I resent it.
  - [Instructions](#) for OS X.
  - [Instructions](#) for iPhone and older Android (untested).
  - [Instructions](#) for modern Android.
  - Windows users: you're on your own I only use Windows for games.
- Discovered the existence of [Focus Mode](#) for android, which allows you to use the internet but disables notifications.
  - All of the Focus Mode instructions require several clicks into a deep menu. You can access it more easily by enabling the relevant fast-access card, using the instructions for older Android phones above.
- Moved all my short-OODA-loop apps off my phone home in favor of long-OODA-loop apps.
  - The newly defavored apps include obvious candidates like Twitter, but also all messaging apps and Chrome itself.

- The new stars include my article aggregator, as well as very long-form content like Kindle, Audible, and PodCatcher.
- The home screen continues to hold non-content-consumption apps I want to access quickly, like Maps, Calendar, etc.
- Resumed use of a read-it-later tool, which lets me save cool articles I see on social media to be consumed when I'm in a better head space (I can't switch between reading microblogs and regular blogs quickly – different headspace ).
- Switched to an RSS reader that lets me read things out of order without marking earlier articles as Read.
- Combined my RSS feeds, email newsletters, and saved articles in a single service (that lets me read in any order) so I can choose from all my essay-length options at once. This was a stupid amount of effort and [yak shaving](#) and it's not pretty but I got it working. Most of this post will be about that.

## How to Combine All Article-Length Content

The following instructions work with both Inoreader and Feedly. I eventually chose Inoreader but it was a close call and your mileage may vary.

1. [Import](#) your existing RSS feeds to Inoreader.
  1. [Feedly](#) instructions.
2. You have two options for saving arbitrary content to your new aggregator.
  1. Inoreader has a built-in feature to do this with a Chrome plugin, but there's no way to see saved articles and RSS articles in the same list. So if that's important for you...
  2. Sign up for [Pocket](#), the leading read-it-later app, and configure Pocket to put [your saved articles on an RSS Feed](#), then add that RSS feed to Inoreader per normal.
    1. Install the [save to Pocket](#) extension in your browser to make it easy to add to the feed as you come across things you want to read at some point (note: not available for Android, so I have to manually copy the URL and open the Pocket app).
      1. All articles in the saved-to-Pocket feed will show in Inoreader as having the same author ("my content feed") and they won't have the body of the article, only the title and header image. I *really* care about having all of my articles in the exact same place, but if you don't, just using save-to-Inoreader will save you several steps.
3. You also have two options for newsletters.
  1. Set up forwarding using Inoreader's [built-in system](#).
    1. This is a Pro feature, so it's \$6/month.
      1. Feedly [version](#) (\$12/month)
    2. You will either need to set up a new email for every subscription (which Inoreader supports, although you'll need to upgrade past 20 subscriptions) or they will all be listed as having the same author.
  2. Or use a newsletter-to-RSS conversion tool like [kill-the-newsletter.com](#), and add the resulting RSS feed to Feedly. Unless substack lets you configure separate emails for each subscription this will still require you to set up forwarding.

Tada! With a mere 20 minutes of work and a small monthly fee you have a system that combines all of your article-length-content in one place. Inoreader and Feedly both support Youtube channels and podcasts as well, although I haven't tested those out.

## Rejected Options

## Pocket

Pocket bafflingly doesn't support RSS input. You can hack it with [IFTTT](#) or Zapier, but each RSS feed counts as its own applet so you reach the \$10/month plan very quickly. This is slightly cheaper than the Feedly pro plan but more than Inoreader, plus paying Inoreader gets rid of ads.

Pocket allows you to [forward individual emails](#) to it, but because Gmail requires forwarding address confirmation you can't automate forwarding to pocket from Gmail. You could fix this with a newsletter-to-RSS converter and then IFTTT/Zapier, but that's a lot of work.

## Feedly

I originally settled on Feedly before making one final sweep and choosing Inoreader instead. The driving concern was that Feedly required me to be on tier 3 (\$12) rather than Inoreader's tier 2 (\$6). Inoreader's browser plugin was also better, letting you subscribe while on a blog's homepage, where Feedly requires you to not only go to its own page to add feeds, but track down the actual feed URL rather than figuring it out from the blog's homepage URL (which is surprisingly hard because RSS is out of favor and most readers can guess, so the RSS feed is rarely displayed prominently).

I do like the Feedly Android app a little more; Inoreader has not adjusted to Pixel's lack of a back button and eats the replacement gesture, but I liked the webapp more so stuck with Inoreader.

Feedly Pro+ (required to get newsletter forwarding) boasts an AI assistant I assume is terrible. Inoreader has easily accessible filters and prioritization rules that I haven't tested. Between the two of these I predict I get more value out of Inoreader, but I could be wrong.

Lastly, Feedly promised me a trial and immediately charged me for an annual subscription, so they can die in a fire.

## Everyone Else

[This is a spreadsheet](#) where I went through every read-it-later service I could find, looking for RSS native support. None of them had it. It is possible there is another RSS reader with better bookmarking or newsletter support, but I am exhausted and Inoreader is working so I stopped looking.

# Regularization Causes Modularity Causes Generalization

*Epistemic Status: Exploratory*

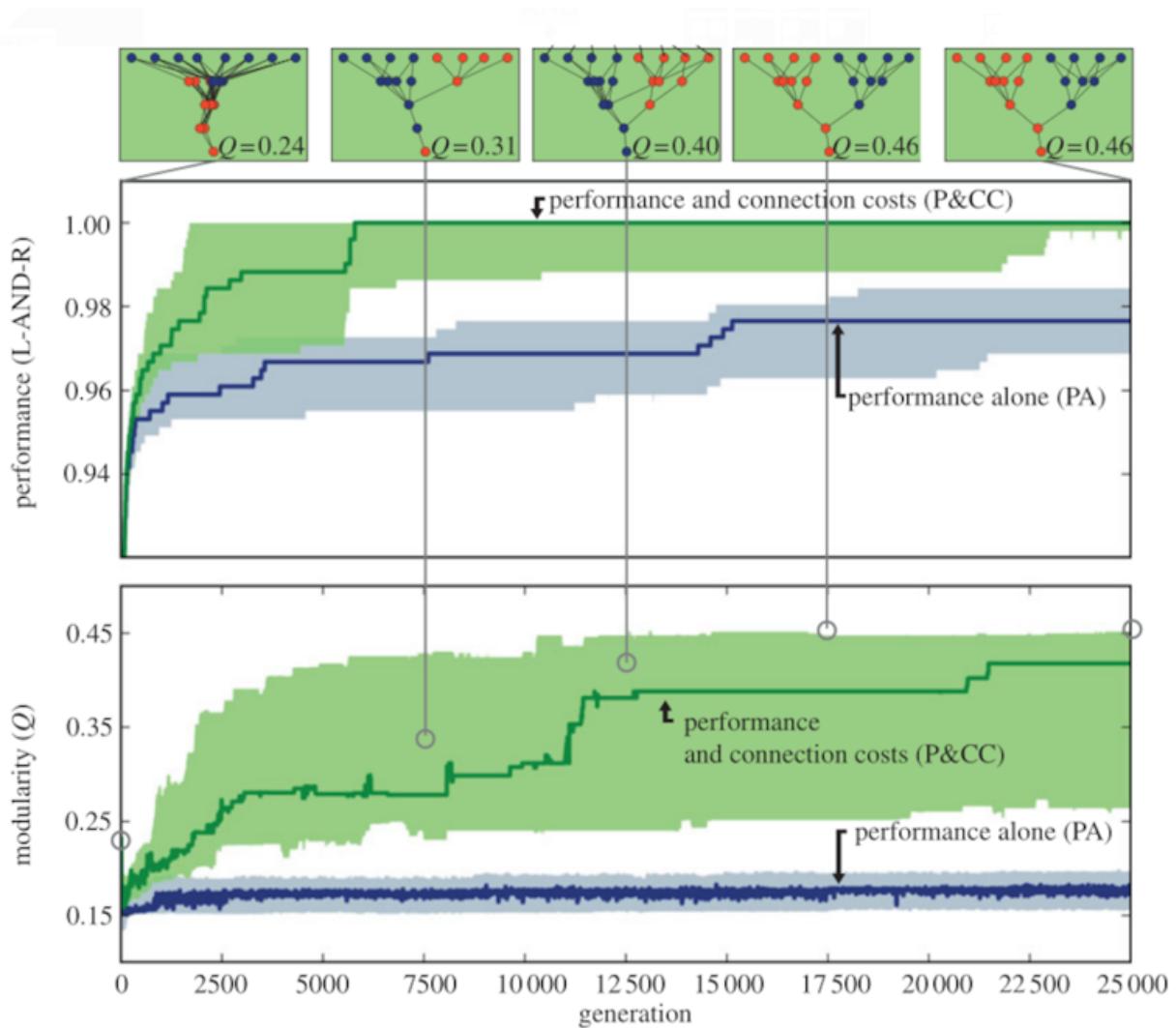
## Things That Cause Modularity In Neural Networks

**Modularity** is when a neural network can be easily split into several modules: groups of neurons that connect strongly with each other, but have weaker connections to outside neurons. What, empirically, makes a network become modular? Several things:

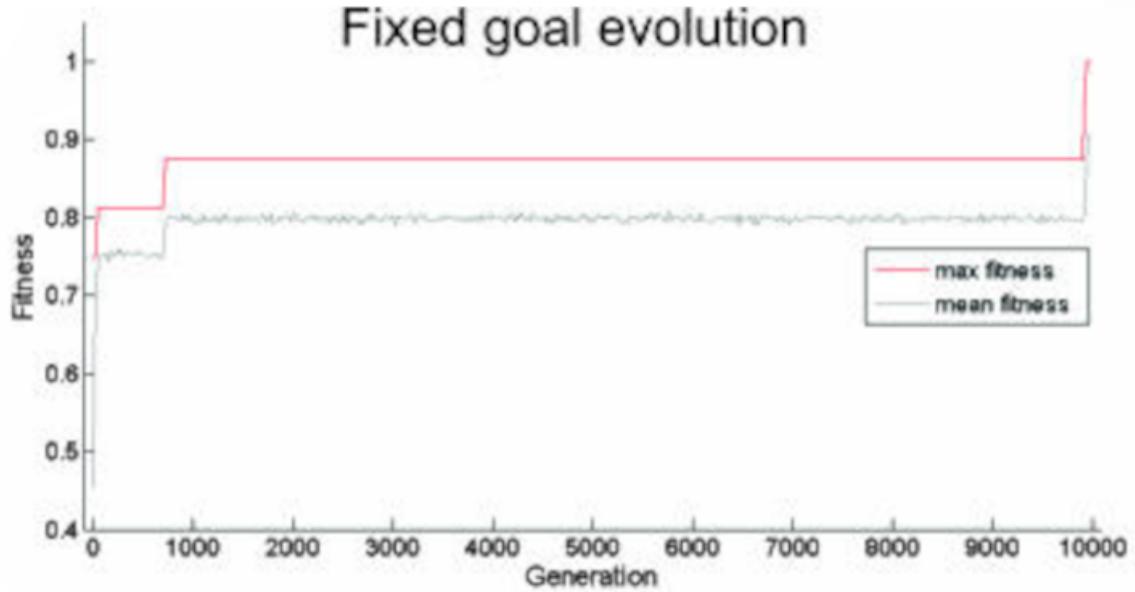
- [Filan et al.](#)<sup>[1]</sup>:
  - Training a model with dropout
  - Weight pruning
  - L1/L2 regularization
- [Kashtan & Alon](#): Switching between one objective function and a different (but related<sup>[2]</sup>) objective function every 20 generations
- [Clune et al.](#): Adding penalties for connections between neurons

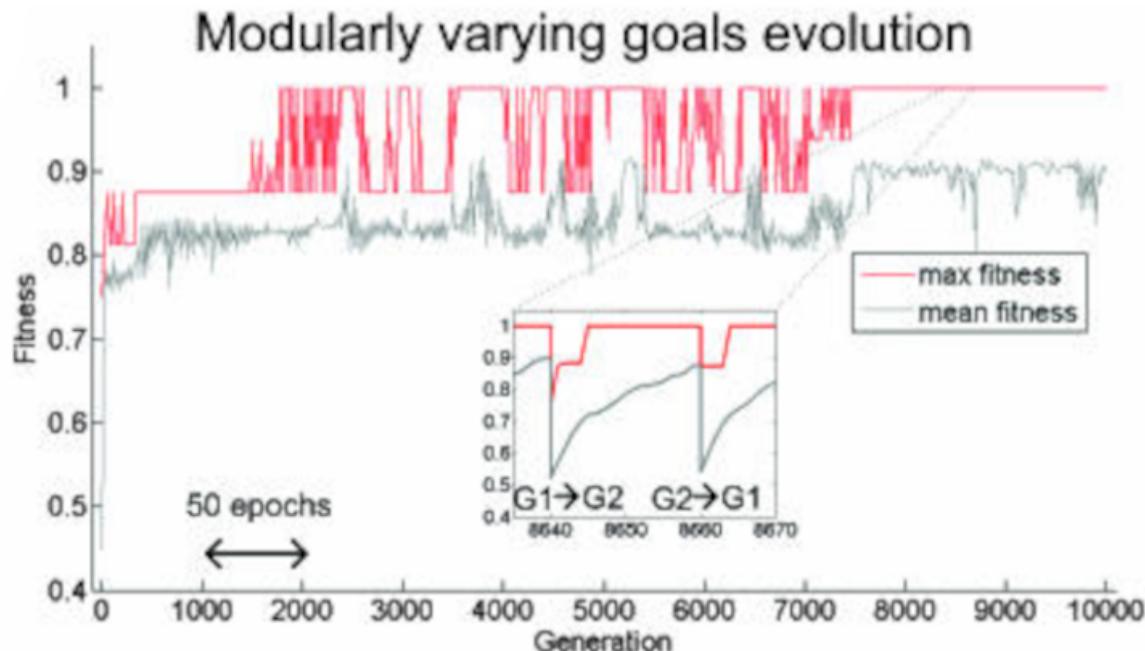
## Modularity Improves Generalization

What good is modularity? Both [Clune et al.](#) and [Kashtan & Alon](#) agree: more modular networks are more adaptable. They make much more rapid progress towards their goals than their non-modular counterparts do:



### Fixed goal evolution





Modular neural networks, being more adaptable, make faster progress towards their own goals. Not only that, but their adaptability allows them to rapidly advance on related [2:1] goals as well; if their objective function was to suddenly switch to a related goal, they would adapt to it much quicker than their non-modular counterparts.

In fact, modular neural networks are so damn adaptable that they do better on related goals *despite never training on them*. That's what generalization is: the ability to perform well at tasks with little to no previous exposure to them. That's why we use L1/L2 regularization, dropout, and other similar tricks to make our models generalize from their training data to their validation data. These tricks work because they increase modularity, which, in turn, makes our models better at generalizing to new data.

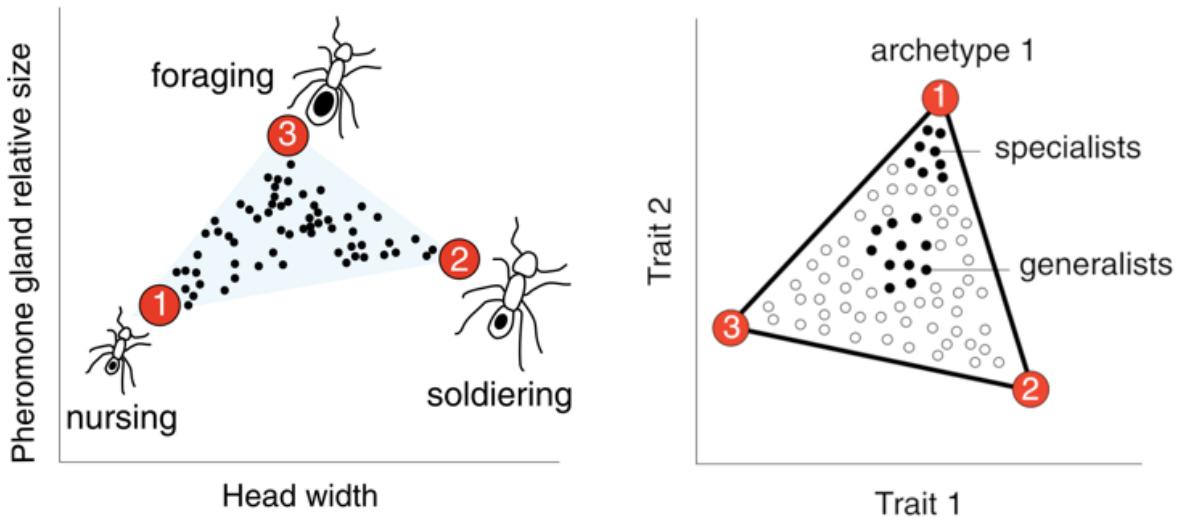
## How Dropout Causes Modularity

What's true for the group is also true for the individual. It's simple: overspecialize, and you breed in weakness. It's slow death.

—Major Kusanagi, *Ghost in the Shell*

Training with dropout is when you train a neural network, but every neuron has a chance of 'dropping out': outputting zero, regardless of its input. In practice, making 20-50% of your model's neurons spontaneously fail during training usually makes it much better at generalizing to previously unseen data.

Ant colonies have dropout. Ants die all the time; they die to war, to famine, and to kids with magnifying glasses. In response, anthills have a high [bus factor](#). Not only do anthills have specialist ants that are really good at nursing, foraging, and fighting, they also have all-rounder ants that can do any of those jobs in an emergency:



Dropout incentivizes robustness to random module failures. One way to be robust to random module failures is to have modules that have different specialties, but can also cover for each other in a pinch. Another way is to have a bunch of modules that all do the exact same thing. For a static objective function, from the perspective of an optimizer:

- If you expect a really high failure rate (like 95%), you should make a bunch of jack-of-all-trades modules that're basically interchangeable.
- If you expect a moderate failure rate (like 30%), you should make your modules moderately specialized, but somewhat redundant. Like ants!
- If you expect no failures at all, you should let modules be as specialized as possible in order to maximize performance.
  - Do that, and your modules end up hyperspecialized and interdependent. The borders between different modules wither away; you no longer have functionally distinct modules to speak of. You have a [spaghetti tower](#).
  - Why would modules blur together? "Typically, there are many possible connections that break modularity and increase fitness. Thus, even an initially modular solution rapidly evolves into one of many possible non-modular solutions." —*Design Principles of Biological Circuits* ([review](#), [hardcopy](#), [free pdf](#))

Dropout is performed on *neurons*, not "modules" (whatever those are), so why does this argument even apply to neural networks? Modules can have sub-modules, and sub-modules can have sub-sub-modules, so (sub-)<sup>n</sup>-modules are inevitably going to be made up of neurons for some value of n. The same principle applies to each level of abstraction: redundancy between modules should increase with the unreliability of those modules.

So dropout incentivizes redundancy. How does that boost modularity? A system built from semi-redundant modules is more, uh, modular than an intricately arranged spaghetti tower. Not all functionally modular systems have redundant elements, but redundant systems have to be modular, so optimization pressure towards redundancy leads to modularity, which leads to generalization.

## How L1/L2 Regularization Causes Modularity

L1/L2 regularization makes parameters [pay rent](#). Like dropout, L1/L2 regularization is widely used to make neural networks generalize better. L1 regularization is when you add a term to the objective function that deducts points proportional to the sum of the magnitudes of all of a model's parameters. L2 regularization is the same thing, but you square the parameters first, and take the square root at the end.

The primary effect of L1/L2 regularization is to penalize connections between neurons, because the vast majority of neural network parameters are weights, or connections between two neurons. Weight pruning, the practice of removing the 'least important' weights, also has a similar effect. As we know from Filan et al., L1/L2 regularization and weight pruning both increase the modularity of neural networks.

Connection costs don't just increase the modularity of artificial neural networks. They increase modularity for biological neural networks too! From [Clune et al.](#):

The strongest evidence that biological networks face direct selection to minimize connection costs comes from the vascular system and from the nervous systems, including the brain, where multiple studies suggest that the summed length of the wiring diagram has been minimized, either by reducing long connections or by optimizing the placement of neurons. Founding and modern neuroscientists have hypothesized that direct selection to minimize connection costs may, as a side-effect, cause modularity.

The authors of this paper then go on to suggest that *all* modularity in biological networks is caused by connection costs. Whether or not that's true<sup>[3]</sup>, it's clear that optimizers that penalize connections between nodes produce more modular networks. Natural selection and ML researchers both happened upon structures with costly connections, and both found them useful for building neural networks that generalize.

---

1. One other thing that increases modularity is just training a neural network; trained networks are more modular than their randomized initial states. ↩
2. When I say "related goals", I mean goals that share subgoals / modular structure with the original goal. See [Evolution of Modularity](#) by johnswentworth. ↩ ↩
3. I highly doubt it. ↩

# CICO is a weird reference class

People tend to form what I guess I'm going to call default-hypothesis-reference-classes. This is what they see as the "default" hypothesis that new hypotheses have to compete against. Obviously this isn't how a perfect bayesian reasoner would act, but we're only human.

These weird reference classes can be tricky, because you can get "sucked into them" and stop noticing that they're weird.

I started thinking of this when thinking about the CICO model of weight gain/loss. Calories-in-calories-out sounds obviously true under thermodynamics. There's about  $2 \times 10^{10}$  Calories in a gram of matter, and humans need about 2000 Calories per day, so make sure not to eat more than a tiny fraction of a gram per day and you're sure to lose weight.

Well, not quite. In my opinion, the most "natural" default reference class for weight gain or loss is mass-in-mass-out (MIMO), which is obviously true, if not super useful. This has to be true - everything we know about physics tells us it's impossible for your body to gain or lose mass unless it comes from or goes somewhere.

What people mean when they say CICO is that different foods have different amounts of bioavailable energy, and weight gain is dependent on difference between the bioavailable energy you eat and the amount of heat your body produces. This doesn't seem like it necessarily has to be true! People gain weight when they grow, and that's partially because they're eating actual minerals that are deposited into their bones. (And these minerals largely don't go anywhere once they're deposited.) So that's a straightforward counterexample to CICO.

We could imagine the same thing happening with fat gain or loss. Your body could store energy by accumulating fat, but not remove the actual fat molecules from your body when their chemical energy is consumed. (I'm not saying that this actually happens, just that it's not forbidden by thermodynamics.) Or our model for what kinds of energy are bioavailable could be wrong, and if so we'd give wrong predictions when trying to apply CICO even if the theory itself were true.

What I'm trying to say is, if your previous "default" reference class for models of weight gain was MIMO, CICO would seem weird and not that special to you. CICO is very likely more useful than MIMO, but it wouldn't intuitively seem like it should be the default theory all other theories are at best a special case of.

Once I started looking for weird reference classes, I started seeing them all the time. For example, I once saw some people discussing a study that examined the effect of new housing on "the asking price for 3-bedroom condos". What seemed weird to me is that the study chose a reference class, "the asking price", which intuitively seems like not the thing we usually want to talk about, which is the sale price.

I think this one was discussed more broadly, but at the beginning of the pandemic I saw a lot of people saying "this country had good outcomes, but their population density is very low" or things like that. But population density is an imperfect metric for what we really care about, which is how clustered a population is. Canada's population density is very low, but you could delete 90% of Canada and probably only lose 10% of their population. This fictional decimated-Canada would probably have

had similar covid outcomes to real-canada, but with a 9x higher population density. So the reference class of "the difficulty of managing covid in a country is dependent on the country's population density" seems like a sketchy choice for a default reference class of covid-management-difficulty prediction generators.

# Covid 1/20/22: Peak Omicron

The Omicron and Weekly posts are now combined, so this includes the last day's Omicron developments, which will be how it works going forward. Next week's will include the probability updates section as well. When there is urgent news, I'll supplement as appropriate.

This week was the peak. From here, it would be highly surprising if Omicron cases didn't start declining. The other news is mostly more of the same, with several developments worth analyzing: The Supreme Court throws out one of Biden's mandates but upholds the other, Djokovic gets deported [a second time](#), a great [proposal to expand Manhattan](#) (which got its own post), and other neat stuff like that.

## Executive Summary

1. If you live in USA and haven't done so yet, [order your free Covid-19 tests](#).
2. We are at or past the peak in Omicron infections in the United States and UK, but that doesn't mean it's over yet, next few weeks still rough.
3. UK lifts all Covid-19 restrictions, midnight is perhaps temporary after all.

Let's run the numbers.

## The Numbers

### Predictions

Prediction from last week: 5.6mm cases (+13%) and 15,000 deaths (+30%).

Results: 4.9mm cases (+0%) and 12,606 deaths (+7%).

Prediction for next week: 4.4mm cases (-10%) and 14,500 deaths (+15%).

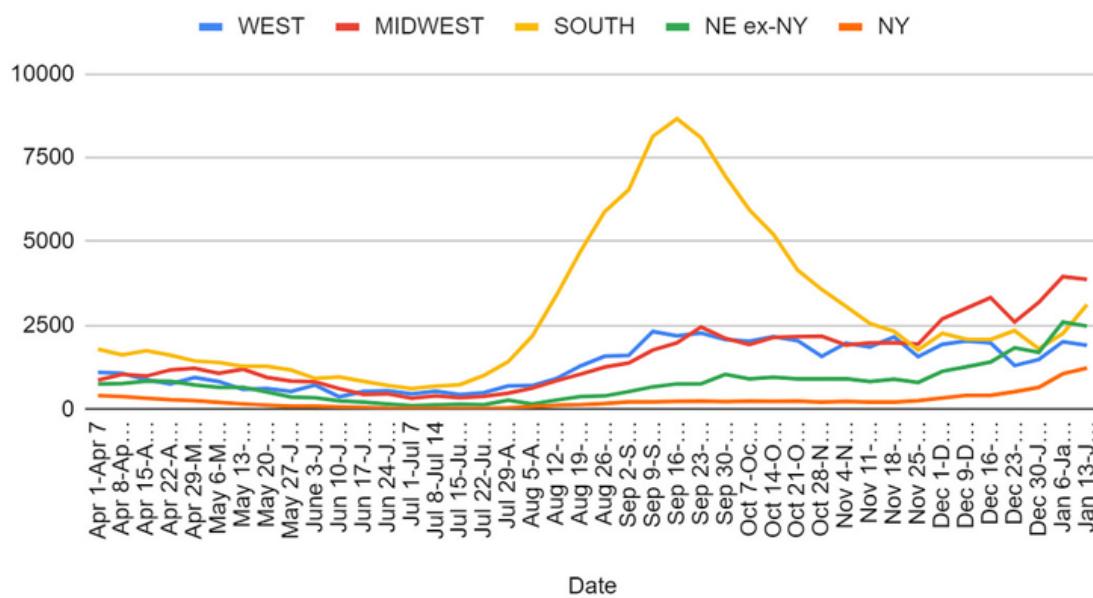
The deaths number is confusingly low this week. Cases increased by a *huge* amount with the three week lag, yet the deaths number isn't increasing much. There's about to be a *much bigger* three-week-lagged increase in cases, but the pattern is clear so my estimated size increase needs to adjust. I am fully prepared to miss low reasonably often here, if I wasn't the estimate would be wrong.

For cases, it seems clear that we are peaking, so we should expect things to begin to decline. The question is how fast, with some areas not yet peaked, and the Northeast already having a much smaller share of total cases. The decline could easily be much faster than this, but things could also stay mostly steady for one more week.

### Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Nov 25-Dec 1	1568	1940	1768	1053	6329
Dec 1-Dec 8	1933	2693	2264	1465	8355
Dec 9-Dec 15	2036	3014	2088	1675	8813
Dec 16-Dec 22	1978	3327	2080	1818	9203
Dec 23-Dec 29	1302	2602	2342	2353	8599
Dec 30-Jan 5	1481	3185	1805	2343	8814
Jan 6-Jan 12	2013	3955	2258	3656	11882
Jan 13-Jan 19	1904	3873	3121	3708	12606

## Deaths by Region

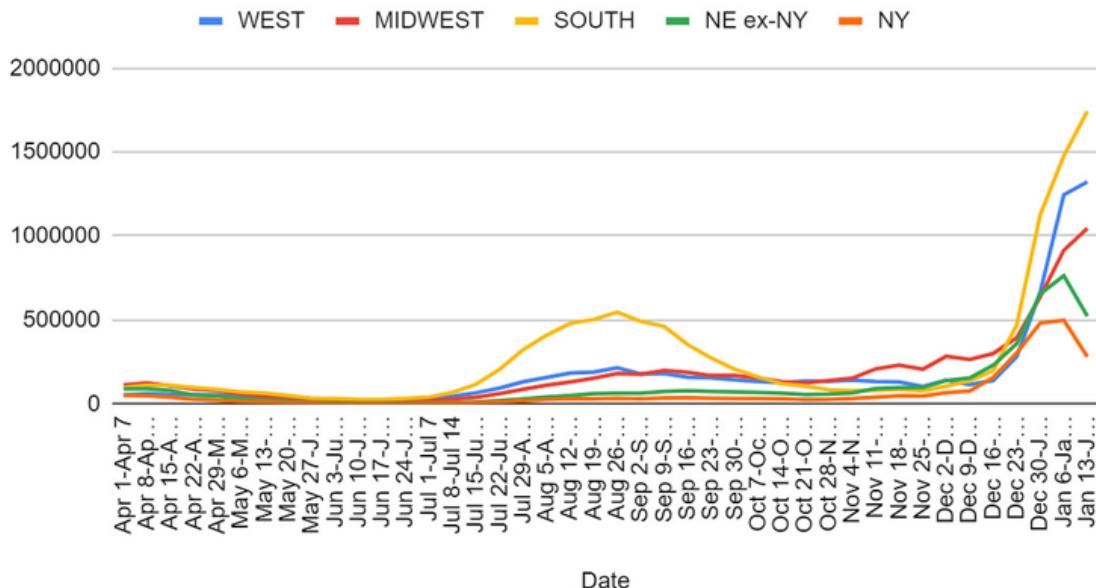


Even with Omicron being milder, the lack of more deaths is very good news here. The next week is the time when deaths should by all rights go much higher, so if they stay under control this week, we're fully in the clear.

## Cases

Date	WEST	MIDWEST	SOUTH	NORTHEAST	TOTAL
Nov 25-Dec 1	102,685	206,259	76,590	143,243	528,777
Dec 2-Dec 8	140,943	283,433	105,544	206,259	736,179
Dec 9-Dec 15	114,266	264,402	136,589	231,791	747,048
Dec 16-Dec 22	140,012	300,125	198,037	388,788	1,026,962
Dec 23-Dec 29	281,703	394,422	472,090	663,080	1,811,295
Dec 30-Jan 5	667,872	639,499	1,128,524	1,137,319	3,573,214
Jan 6-Jan 12	1,245,702	915,402	1,477,626	1,261,405	4,900,135
Jan 13-Jan 20	1,322,131	1,046,289	1,743,659	803,597	4,915,676

## Positive Tests by Region



The Northeast already peaked, as probably did a number of other large states like Illinois, Florida and California. The South might have one more week left in aggregate, it also might not.

[Cases at Yale in similar freefall to rest of east coast.](#)

[Thread pointing out that cases are declining in many places.](#)

[Trevor Bedford thread on us having peaked nationally but often not locally, including good state-by-state graphs.](#) Expects about a symmetrical march down to the path up.

[Ashish Jha thinks we have peaked.](#)

## Vaccinations

US announces ([MR secondary source](#), [StatNews article](#)) that...

...if the Food and Drug Administration decides to update Covid-19 vaccines to take better aim at Omicron or other variants, it is unlikely to go it alone.

Instead, a senior FDA official told STAT, the agency expects to take part in an internationally coordinated program aimed at deciding if, when, and how to update Covid-19 vaccines. The approach would ensure decisions are not left solely to individual vaccine manufacturers.

"We can't have our manufacturers going willy-nilly [saying], 'Oh well, the EMA decided they wanted this composition, but FDA wanted that composition,'" the official said, referring to the European Medicines Agency. "So we are very much of the mind that we would like to be part of a more global process in helping to come to what vaccine composition there should be now."

Tyler is kind:

Designed for flexibility and speedy response? I guess we'll see. Here is [the full StatNews article](#). And obviously, the entire public health community is up in arms about this...

The entire public health community being up in arms about it is at least *some* good news, although I haven't seen any other evidence of their arms being up in this, so perhaps arms are not so up as I would like to see. Any version of this that they *weren't* up in arms about would almost be *guaranteed* to involve sufficient delays as to render any updates so slow as to be useless.

Then again, the system for determining flu vaccines that this is modeled on works reasonably well as far as I can tell, so perhaps it will be fine, and also the alternative was perhaps a true nightmare? If comparing to 'the manufacturers do what they think is best and we agree to be OK with that' this is obviously a painfully slow nightmare, but that does not seem to be the relevant counterfactual, except insofar as this is *enshrining* that we will never, ever do anything that sensible. To that extent, this is a loss. To the extent that we'd already accepted that we'll never do anything sensible, this is probably an expected value win.

## Vaccine Effectiveness

[Vaccines do good work even shortly after being infected \(paper\)](#).



**Benjy Renton** @bhrenton · Jan 16

...

A preprint using Israeli health data finds that people vaccinated immediately following an exposure (unintended post-exposure-prophylaxis) can help reduce the chance of death from COVID-19, demonstrating that it is never too late to get vaccinated.

**Results** Among Israeli residents identified PCR-positive for SARS-CoV-2, 11,690 were found positive on the day they received their first vaccine injection (BNT162b2) or on one of the 5 days thereafter. **In patients over 65 years**, 143 deaths occurred among 1413 recently injected (10.12%) compared to 280 deaths among the 1413 unvaccinated (19.82%), **odd ratio (OR) 0.46 (95% confidence interval (CI), 0.36 to 0.57; P<0.001)**. The most significant reduction in the death toll was observed among the 55 to 64 age group, with 8 deaths occurring among the 1322 recently injected (0.61%) compared to 43 deaths among the 1322 unvaccinated control (3.25%), OR 0.18 (95% CI, 0.07 to 0.39; P<0.001).

It can still be too late to get vaccinated, of course, as the videos of intensive care patients can attest to, but it looks like the vaccine very much starts doing good work right away, and your risk of death is reduced even before the vaccine has time to do its work. This makes physical sense as does the effect size, so I'm inclined to believe it.

## Vaccine Mandates

[Mandates are popular](#), but this does change somewhat when the proposed details are *completely insane*. Somewhat.



– Fifty-nine percent (59%) of Democratic voters would favor a government policy requiring that citizens remain confined to their homes at all times, except for emergencies, if they refuse to get a COVID-19 vaccine. Such a proposal is opposed by 61% of all likely voters, including 79% of Republicans and 71% of unaffiliated voters.



- Nearly half (48%) of Democratic voters think federal and state governments should be able to fine or imprison individuals who publicly question the efficacy of the existing COVID-19 vaccines on social media, television, radio, or in online or digital publications. Only 27% of all voters – including just 14% of Republicans and 18% of unaffiliated voters
  - favor criminal punishment of vaccine critics.

So yes, there are *some* limits. Almost half of Democratic voters want to fine or imprison those who publicly question vaccines, but that means that the other half don't. And forty percent of Democrats *don't* want to place the unvaccinated under semi-permanent house arrest.

You say sixty-percent empty, I say forty-percent full.

I've seen objections to this poll on the grounds that Rasmussen Reports is a right-wing polling organization. It's definitely that, so you should take the information with that in mind. My understanding is that they will have a partisan lean in their horserace polls and might replicate such a lean in whatever direction is convenient in other polls, and they certainly *chose to ask these questions* because they think the questions are to their benefit, but I don't believe that they are making things up. These numbers are at least *approximately* correct – maybe 52% instead of 59% of Democrats want to place the unvaccinated under house arrest, that's different, but it's not *fundamentally* different.

I also find that reaction interesting, because it implies that every agrees that *this is a very bad take* and it's politically advantageous to Republicans if they can show that a lot of Democrats do in fact have this terrible take, so Democrats are accusing Republicans of making it up.

This implies that [the people on the left applauding such tough measures](#) are scoring in-group points but not helping their broader cause.



Alex Tabarrok @ATabarrok · Jan 15

...

Occupy Democrats stands strong for firing workers!



Occupy Democrats ✅ @OccupyDemocrats · Jan 14

If you're a Democrat who is thankful to Nike for standing strong and announcing today that they will be firing all unvaccinated employees by January 15, despite relentless Republican attacks on its company mandate, please RT and follow our account for more breaking news!

Also, [mandates work.](#)

As the global leader in live entertainment, Live Nation has been doing shows around the world throughout the summer where we put vaccine and testing protocols into practice, including Lollapalooza, Latitude, Camp Bestival and Wilderness. Data has continued to show the approach successfully allows everyone to enjoy live music, while also inspiring more people to get vaccinated. For example, at Lollapalooza 90% of the festival's 400,000 attendees showed up fully vaccinated, with 12% of those fans reporting that Lollapalooza was their motivation for getting vaccinated.

That sounds pretty great to me. A big music concert is a highly optional experience that provides a lot of value. Vaccine mandates make the concert more desirable for a lot of people, so having one provides more value to customers and hence more concerts and music. Also you convince almost fifty thousand people to get vaccinated in order to attend Lollapalooza alone.

So of course this was the comment that came attached.



Kerry 🇺🇸 @kerry62189 · 5h

...

This from Live Nation is wretched. I can't believe any one defends it. I mean, I don't know many who do, but I don't understand why more people don't see this as wildly inappropriate. The employer ones are even worse.

On the other hand, her other example seems different, with [Brookline taking dining mandates to the next level](#). There continues to be precious little middle ground where physical circumstances and other details *actually matter* to what people support.

**Brookline's mandate**, similar to Boston's, will begin to roll out on January 15, 2022. One key difference is that Brookline's mandate includes restaurants' outdoor dining areas, not just indoors. Some restaurants have already put their own requirements in place, like **Mahaniyom**, which began asking for proof **in late December 2021**.

[She also shares some details about the protests going on in Boston](#), which I found fascinatingly local and quirky.

[This thread by Andy Slavitt highlights the mindset](#). He refers to DeSantis and others as 'paying people to not get vaccinated' because they *refuse to make exceptions* and deny those people payments they would normally be entitled to, and consider vaccinating and boosting a personal choice. Letting those fired for not getting vaccinated get unemployment? Paying them. Hiring people from other states to come work in Florida and offering signing bonuses, because more officers are needed? Paying for not getting vaccinated. Asking Dr. Fauci questions that don't genuflect? He has no idea what anyone could be asking of this 'humble civil servant' who keeps a portrait of himself in his office and may have been involved with gain of function research.

We see this a lot, including outside the context of Covid-19 or vaccination. Any refusal to cut ties and payments to those engaging in X is seen as *paying for X*, or as support for X, as opposed to thinking there is more to life than X.

[Whereas here is a better take](#) we can hopefully all agree upon: Using that amazing Telemundo enthusiasm to make clear that you think anti-vax people are morons, and letting the morons think to themselves 'no, you're a moron' and then we all move on.

(An aside: For some reason, YouTubeTV decided that my Monday Night Football play-by-play for the wildcard game (but not the onscreen graphics or between-drives comments) were going to be handed over to the Spanish language announcers (but *not the closed captions*, which meant I couldn't use this as a chance to learn some Spanish, although I did pick up a few words like 'touchdown' and 'linebacker.'). I couldn't understand most of what they were saying, but I am confident that if I *could* have understood them they would have made the game a lot more fun, because those people were *so excited* to be there and clearly *loved* their football the same way they love their futbol, you love to hear it.)

## NPIs Including Mask and Testing Mandates

[Anarchy in the UK! Woo-hoo!](#)



**Interpolations** @interpolated · 15h

BREAKING: England removes all COVID restrictions including passes, mandates, and WFH requirements.

73

569

4.2K



...

If you thought this would never happen and these restrictions were going to be permanent, you need to update.

[It is delivery, but it's no DiGiorno.](#)



Brian Stelter ✅ @brianstelter · Jan 14

Oliver was so impressed with his pizza delivery speed, he texted me and said "The federal government should consult with Dominos on how to deliver Covid tests"

...



Oliver Darcy ✅ @oliverdarcy · Jan 14

The speed in which Dominos delivers in NYC is truly astounding.



217



62



604



Domino's *is not pizza*. It is a substance almost, but not quite, entirely unlike pizza. At a minimum, *within the confines of New York City* it is *definitely not pizza*. [This is how to tell your pizzas](#), which are things Domino's *does not produce*. [Here are some places in New York City that do produce pizzas, high quality ones at that](#).

The thing is, here, that's an advantage. Domino's is in the non-pizza delivery business. That's exactly what we need, to deliver non-pizzas as quickly and efficiently as possible. Capitalism solves this. Speed premium, ahoy.

The good news is it turns out we did solve this one, at least on the ordering side. By all reports, [the ordering and delivery service](#) is giving us four free masks, with sub-minute ordering times and no hassles whatsoever. It even went live a day early. We finally got one right. Let's stop to ask *how we got this one right*, so we can get future ones right as well. Then let's check to see if the deliveries show up on time.

There's also another delivery coming soon, [they're sending us N95 masks](#). That's also a very good idea. It comes from the Strategic National Stockpile, which hopefully we will rapidly replenish. It seems like we frequently draw from some huge National Stockpile in time of need, but we never hear about the announcement of the expansion or creation of a National Stockpile.

As government interventions go, creating lots more and bigger Strategic National Stockpiles of goods that are highly valuable in a pinch seems like an excellent plan, and a 1-out-2-in general rule might even be justified. If we'd had enough masks in reserve in March 2020, things would have turned out very differently. The pandemic would have been better contained and institutional credibility wouldn't have been incinerated lying about masks.

If left to its own devices Capitalism Solves This because private actors would invest in stockpiles to sell at a profit, but we have a variety of official and unofficial price controls that prevent this. Everyone is terrified to raise prices lest they piss people off, and when higher prices are most needed to ensure incentives and proper allocations, we start arresting the people who charge them. So, same as it ever was, once we ban the market from solving the problem we need to solve it collectively. There's no reason we can't have a reserve of billions of masks.

[The question was asked](#), if masks are for one-day use, what's the point of sending each person *one mask*? The response was that masks save lives, that's the important thing, we want to give people the tools they need to protect themselves and their families. Which is both the right answer – some supply is better than no supply – and also completely ignoring the actual question.

A better answer (in addition to 'this is what we can do right now and it's a start') would be that people can use the N95s in the situations where they matter most, since risks follow

power laws, so they still do a lot of good, and also they send a strong message to switch over to N95s in general. An even better message would be to admit that almost no one treats N95s as the disposable items they're supposed to be, and instead reuses the same N95 for quite a while, and mostly This Is Fine, it's still way better than cloth masks (that you also probably reuse for weeks or months on end).

It also sheds light on the initial mask shortage back in March 2020.

## I'm Sorry, Sir, This is The Centers for Disease Control and/or a Wendy's

It should be the job of the Centers for Disease Control to *control disease*.

Or at least, [that's one theory, sure](#).

[It turns out that controlling disease means doing things when they might matter](#).



Nate Silver @NateSilver538 · 22h

I'm sympathetic to the CDC, but making "rapid decisions based on scant evidence" isn't some esoteric function... it's how life works in any crisis!

...

And even if you're only in a crisis 5% of the time, what you do in that 5% is probably more important than the other 95%.



Apoorva Mandavilli @apoorva\_nyc · Jan 17

NEW: The coronavirus has forced the CDC to make rapid decisions based on scant evidence. For a bureaucratic agency staffed primarily by meticulous medical professionals, the change has not been easy.



Nate Silver @NateSilver538 · 22h

...

Replies to [@NateSilver538](#)

It's sort of like saying "Sure, the National Weather Service can tell you the probably that it will rain tomorrow or what the high temperature might be. But hurricanes and tornadoes? Forget it."

That's not to say that there is no useful meticulous science to be done, very slowly and carefully, around controlling disease. That, too, has its time and place. But if we hope to control disease in a crisis when it matters most, your [OODA loop](#) has to be measured in days or hours [rather than months or years](#). I've still got further discussion of that thread on CDC publications on my stack, but I want to do it justice and it no longer seems time-sensitive.

Thinking probabilistically and making decisions quickly under uncertainty *is their job*. Even more than it is the job of most other people as well.

Thus, I was thrilled to see America's Finest News Source report that [they're finally taking correct and bold action of exactly the right type](#).

# CDC Announces Plan To Send Every U.S. Household Pamphlet On Probabilistic Thinking

ATLANTA—Stressing that the effort represented the best chance of ensuring American make responsible choices around the pandemic, the Centers for Disease Control and Prevention announced Thursday that it planned to send every U.S. household a pamphlet on probabilistic thinking and decision-making. “What we’re hoping to do is give every American a quick refresher on how to use statistical analysis to assess their priors and make Bayesian inferences, thereby ensuring they overcome their innate psychological biases—simple stuff, but important nonetheless,” said CDC director Rochelle Walensky, estimating that the pamphlets’ lessons on the baseline fallacy alone would save far more lives than mask-wearing, handwashing, and the Covid-19 vaccine combined. “Obviously, most everyone in the nation has already at least skimmed the seminal studies of psychologist Danny Kahneman on the fallibility of human thinking. So the pamphlet will skip over some of the basic stuff like the availability heuristic and head straight to prospect theory. Hopefully none of this feels too patronizing.” Walensky added that if Americans took away one easy lesson from the pamphlet, she hoped it would be  $P(H|E) = (P(E|H) * P(H)) / P(E)$ .

Definitely that, completely unironically.

## Supreme Court

The Supreme Court upheld the federal vaccine mandate on health care workers, but rejected Biden’s attempt to use OHSA to impose a similar mandate on all large employers.

The mandate for federal workers was being imposed without the proper authority to do so, and while I am not a lawyer, I’m weird in that I still believe that should matter when things go before a court of law. The mandate on health care workers had explicit authorization in the statute, so it was justifiably upheld.

[Scott Gottlieb pointed out that the administration was doing little to enforce the rule](#), thus the rule was symbolic and more about giving employers cover to impose mandates they wanted to impose anyway, and so the change makes little difference.

Left-wing voices generally seemed to be of the opinion that the Supreme Court has no business deciding what is and is not allowed under the law, and instead their Very Serious People should decide via Expert Opinion. Or on a more basic level, they cared about the *practical effect* of the restrictions the President wanted to impose, but didn’t seem to care *at all* about the rule of law or whether the President had the legal right to impose the restrictions in question.

[This is the reasonable version, understanding there is a law but not caring for it.](#)



**Ashish K. Jha, MD, MPH**  @ashishkjha · Jan 13

As a physician and public health person

...

I find this ruling very disappointing

So many Americans have been poisoned with vaccine misinformation

Mandates work to protect them and those around them

This ruling will mean more suffering, more deaths, and more overwhelmed hospitals



**World News Tonight**  @ABCWorldNews · Jan 13

The Supreme Court issued a stay of the OSHA vaccine-or-test requirement on private businesses of 100 or more workers.  
[abcn.ws/31UkJ2a](http://abcn.ws/31UkJ2a)

[Here's the pure version, what is the law?](#)



**Oni Blackstock MD MHS** 

@oni\_blackstock

...

Unclear why, during a public health emergency, no less, public health policy is being decided by people without public health expertise.

3:49 PM · Jan 13, 2022 · Twitter Web App

**2,179** Retweets

**124** Quote Tweets

**13.5K** Likes

[And here's the opinion piece version \(WaPo\).](#)



**Kirsten Bibbins-Domingo**  @KBibbinsDomingo · Jan 13

...

“Think about that for a moment. 4 justices, 1 vote short of a majority, would have ruled that, in the midst of a brutal pandemic that has killed millions around the world, the government can't even require health-care workers to be vaccinated”

## Opinion: The frightening philosophy driving the Supreme Court's new vaccine mandate rulings

And in a development that was in some ways even more alarming, [the court upheld](#) a vaccine mandate that applied only to health-care workers in facilities that receive federal funds — but four of the justices voted against it.

...

But four justices — Clarence Thomas, Samuel A. Alito Jr., Neil M. Gorsuch and Amy Coney Barrett — [said just that](#). Legislative language that authorizes the Department of Health and Human Services to regulate the “health and safety” of medical facilities, they said, isn’t enough, since there is no specific language authorizing precisely this kind of mandate.

“If Congress had wanted to grant [the government] authority to impose a nationwide vaccine mandate, and consequently alter the state-federal balance, it would have said so clearly,” Thomas wrote for the others. “It did not.”

The alarming philosophy in question, that a minority of justices endorsed here, is that if the Federal Government wants to impose a law, it must first *pass one*. Scary, I know.

I happen to disagree on the merits here, because I do think that Congress authorized this when it said it could require things that guard the health of Medicare and Medicaid recipients, so if you want to object you’d have to object to the whole broader structure, and they’re not doing that here.

There is one other note I am very sympathetic to here, which is that this rule of law kick only seems to properly and fully apply when a Democrat is occupying the White House. That does seem to be a thing, but the error should be fixed where there is an error.

The note I am mostly *not* sympathetic to is that laws shouldn’t have to be passed because the other side is refusing to pass any laws. If that’s so terrible, as one could reasonably argue, shouldn’t you go and win elections by pointing this out? It’s not like this no-laws policy is new or surprising.

The liberal justices did not appear to be in as good standing as members of the reality-based community as one might have hoped. [They said some things that are not.](#)

For example, this very much is not.

Sotomayor, who did not join her eight colleagues on the bench for the arguments, but opted to take part remotely, was the worst offender. At one point, the Bronx-born jurist claimed that implementing the requirement for businesses was necessary because “Omicron is as deadly as Delta … we have hospitals that are almost at full capacity with people severely ill on ventilators.”

“We have over 100,000 children, which we’ve never had before, in serious condition,” Sotomayor added, “and many on ventilators.”

Not only did Sotomayor’s statement contradict experts who say Omicron is less severe than Delta, but it defied data from the Department of Health and Human Services showing a total of 3,342 confirmed pediatric hospitalizations with COVID-19 across the US as of Friday — making the justice’s math off by a factor of nearly 30.

Here’s Breyer, also saying that which is not.

When Breyer waded into the fray, he suggested the OSHA rule was needed because “hospitals are full almost to the point of maximum” and that “750 million new cases” had been reported in the US yesterday — despite the fact that the population of America is around 330 million.

More disturbingly, [no, seriously, what is the law?](#)

Perhaps more disturbingly, Sotomayor said at another point in the argument that “I’m not sure I understand the distinction why the states would have the power” to institute a rule like the one being pursued by the Biden administration, “but the federal government wouldn’t.”

And there’s this Zen koan, [how is a human not like a workplace machine](#), to be modified to the specifications required for safe operation in the factory? And what does blood have to do with this given that Covid isn’t even blood-borne? Why does everything have to be so creepy?

JUSTICE SOTOMAYOR: So what’s the difference between this and telling employers, where sparks are flying in the workplace, your workers have to be — wear a mask?

MR. KELLER: When sparks are flying in the workplace, that’s presumably because there’s a machine that’s unique to that workplace. That is the —

JUSTICE SOTOMAYOR: Why is the human being not like a machine if it’s spewing a virus, blood-borne viruses?

And of course there was the whole ‘well more people might quit if we don’t force them to get vaccinated so coercion is fine’ argument, from Breyer.

And they said, in our view, hmm, yeah, that’s right, some people may quit, maybe 3 percent. But more may quit when they discover they have to work together with unvaccinated others because that means they may get the disease.

[Here is the actual decision.](#)

The Secretary of Labor, acting through the Occupational Safety and Health Administration, recently enacted a vaccine mandate for much of the Nation’s work force. The mandate, which employers must enforce, applies to roughly 84 million workers, covering virtually all employers with at least 100 employees. It requires that covered workers receive a COVID-19 vaccine, and it pre-empts contrary state laws. The only exception is for workers who obtain a medical test each week at their own expense and on their own time, and also wear a mask each workday. OSHA has never before imposed such a mandate. Nor has Congress. Indeed, although Congress has enacted significant legislation addressing the COVID-19 pandemic, it has declined to enact any measure similar to what OSHA has promulgated here. Many States, businesses, and nonprofit organizations challenged OSHA’s rule in Courts of Appeals across the country. The Fifth Circuit initially entered a stay. But when the cases were consolidated before the Sixth Circuit, that court lifted the stay and allowed OSHA’s rule to take effect. Applicants now seek emergency relief from this Court, arguing that OSHA’s mandate exceeds its statutory authority and is otherwise unlawful. Agreeing that applicants are likely to prevail, we grant their applications and stay the rule.

The counterargument seems to be that OSHA is tasked with safety, this is safety, [therefore OSHA is tasked with it:](#)

In a [dissent](#), Justices Stephen Breyer, Sonia Sotomayor, and Elena Kagan argued that OSHA was acting within its broadly defined limits. “The administrative agency charged with ensuring health and safety in workplaces did what Congress commanded it to: It took action to address COVID-19’s continuing threat in those spaces.”

There was no specific command from Congress for OHSA to address Covid-19, so the implication is that *anything* that impacts health is fair game in the workplace and can be regulated without further authority from Congress. I am not a lawyer, but as I read this, since almost everything impacts safety, this effectively means we would no longer have law with respect to large businesses, only regulations that the President can change at will.

[Here is a legal argument](#) that the OSHA rules do in fact apply to any 'grave danger' and that the court was wrong to draw a distinction between workplace and non-workplace harms, but that the decision was still correct because Covid-19 does not impose a 'grave danger.' Without one, Congress needed to delegate the necessary authority to do this because it is a major question, and they did not do so. Post agrees with my instinct that a decision the other way would have given OSHA control over essentially all workplace conditions.

The interesting question is, how different would physical conditions need to be, in order to change the answer under the law and justify such rules? What if Covid-19 was a lot deadlier, and vaccines protected against transmitting to others but didn't protect you yourself? I'd certainly then *want to impose the mandate*, and I'd hope Congress would be willing to pass a law to do that, but if they refused then 'do it anyway' isn't obviously how the law works.

The concrete issue is moot at this point in any case. The corporations that *want* to impose a mandate already did. The ones that don't want to aren't going to start now either way. It's done. What matters is potential future regulatory overreach, on which the difference has been split.

There was also a to-do about whether justices asked each other to wear masks. There was a report they were in conflict over this, [then it turned out they weren't](#). No idea what happened beyond a clear underlying demand to find a controversy that wasn't there.

## Hospitals

[The 'circuit breaker' dashboard continues.](#)

Here we are in micro.

Benjy Renton @bhrenton · 8h

Tonight's Circuit Breaker Dashboard update:

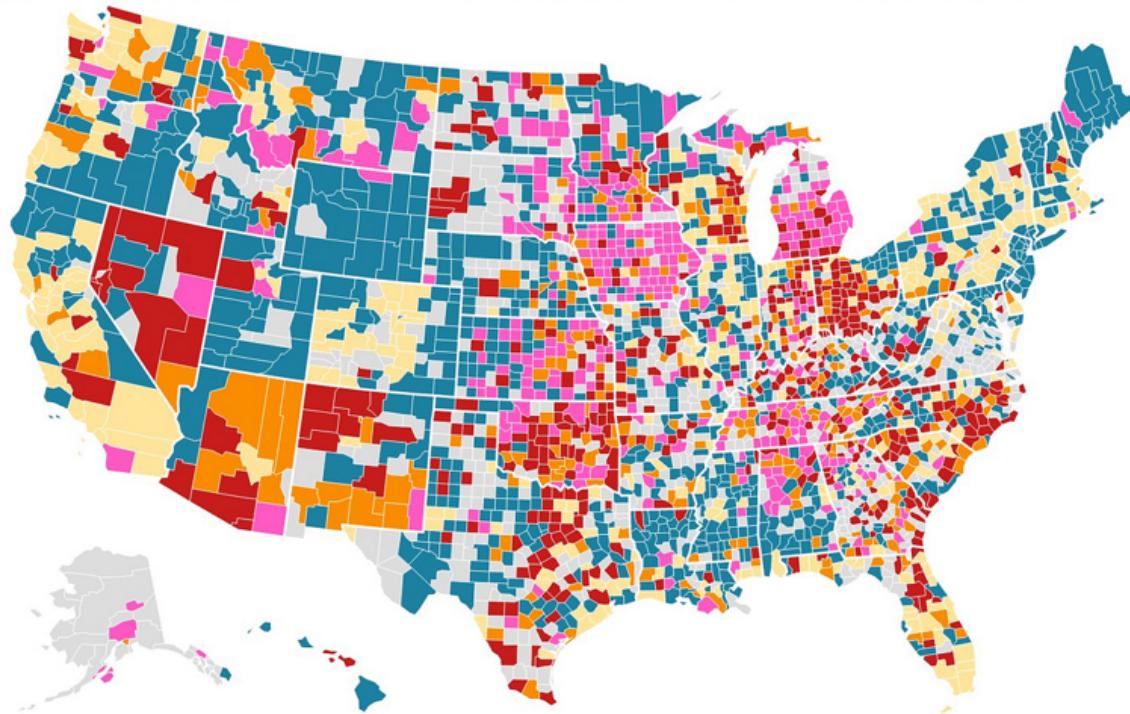
- 478 counties projected to be at hospital capacity (up from 383 yesterday)
- 294 counties are at high risk of hitting capacity within the next 7 days (up from 247 yesterday)
- 364 as "unsustainable"

Explore: [alexanderjxchen.github.io/circuitbreaker/](https://alexanderjxchen.github.io/circuitbreaker/)

## COVID-19 Hospital Capacity Circuit Breaker Dashboard – County

View the circuit breaker dashboard at [alexanderjxchen.github.io/circuitbreaker](https://alexanderjxchen.github.io/circuitbreaker). Data as of January 19, 2022.

■ At or Over Capacity ■ Forecasted to Exceed Capacity ■ Unsustainable ■ Continued Risk ■ Has Capacity



At Capacity - Hospitals are projected to be at or above 100% capacity given a Nowcast with recent COVID-19 case trends (assuming non-COVID care is stable); Forecasted to Exceed Capacity - Given recent COVID-19 case trends, counties are at high risk of exceeding 100% hospital capacity in the next 7 days; Unsustainable - The single day case count yesterday is on a trend to the area reaching a circuit breaker; Continued Risk - Area has been at or forecasted to exceed capacity in the past 7 days; Has capacity - Hospital capacity not at high risk of being exceeded

Map: Jeremy Faust, Bill Hanage, Benjy Renton, Kristen Panthagani, Alexander Chen

As a reminder, here's a translation guide:

1. 'Has capacity' means there's no doubt everything is fine and will stay fine.
2. 'Continued risk' means there is *some* way to doubt everything will stay fine.
3. 'Unsustainable' means things are getting worse so might not stay fine.
4. 'Forecasted to exceed capacity' means there's actual math saying things are going to not be fine.
5. 'At or over capacity' means things are, to some extent, not fine.

We are overall at the peak, but that doesn't mean *any given area* is at the peak, so things that are currently fine could easily end up not fine, but I'd mostly treat anything in categories 1 and 2 as fully fine, and anything in category 3 as likely to be sustained and highly unlikely to be serious levels of not fine. The counterargument to that is that they're continuing to move things from category 3 to category 4, but I still think that the marginal moves are likely to only go marginally over capacity at this point.

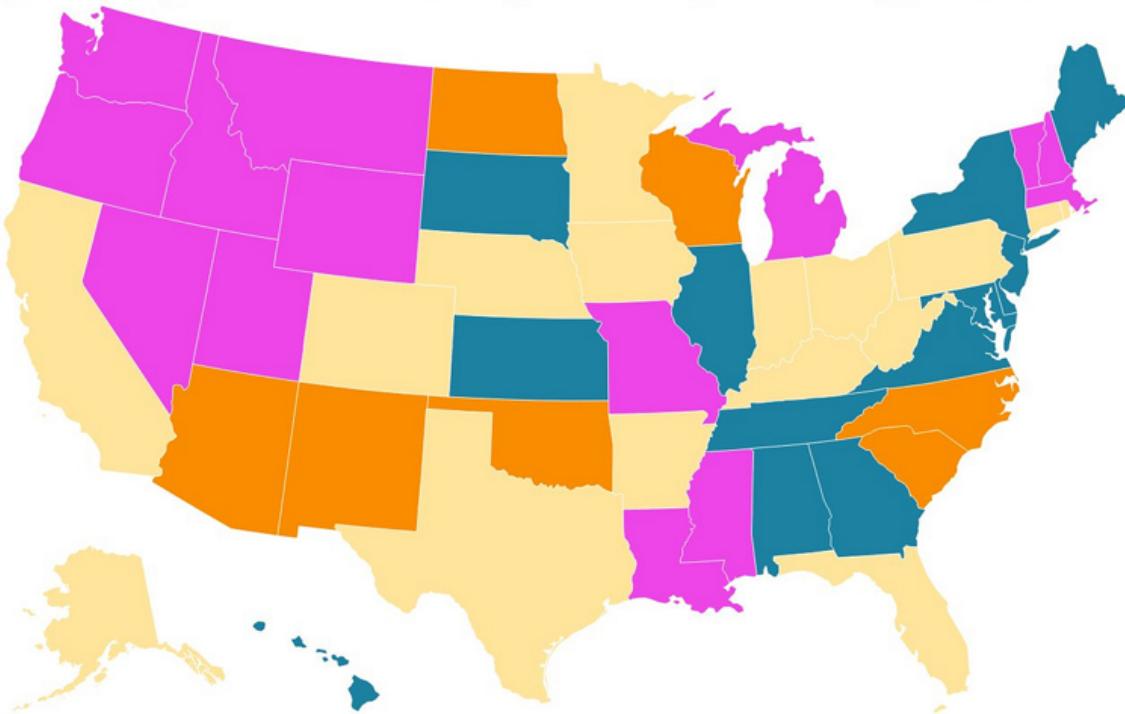
That does still leave a bunch of scary areas, although their combined populations aren't as high as they look due to the rural-urban divide.

Here's the update by state.

## COVID-19 Hospital Capacity Circuit Breaker Dashboard – State

View the circuit breaker dashboard at [alexanderxchen.github.io/circuitbreaker](https://alexanderxchen.github.io/circuitbreaker). Data as of January 19, 2022.

■ At Capacity ■ Forecasted to Exceed Capacity ■ Unsustainable ■ Continued Risk ■ Has Capacity



*At Capacity - Hospitals are projected to be at or above 100% capacity given a Nowcast with recent COVID-19 case trends (assuming non-COVID care is stable); Forecasted to Exceed Capacity - Given recent COVID-19 case trends, states are at high risk of exceeding 100% hospital capacity in the next 1-7 days; Unsustainable - The single day case count yesterday is on a trend to the area reaching a circuit breaker; Continued Risk - Area has been at or forecasted to exceed capacity in the past 7 days; Has capacity - Hospital capacity not at high risk of being exceeded*

Map: Jeremy Faust, Bill Hanage, Benjy Renton, Kristen Panthagani, Alexander Chen • Created with Datawrapper

Notice that Ohio, despite being mostly red in the first picture, is only in yellow here, which illustrates the population densities, and also that local conditions look worse than average conditions and there's some amount of transfer capacity. Where things are over capacity, they're not *wildly* over, and our worst fears are not going to come to pass.

I don't say this to make light of the situation. [As per reports like this](#), conditions in many hospitals are terrible and overwhelming. There's still a world of difference between this and the types of conditions that were feared as recently as a few weeks ago.

## Some Kind of Djokovic

The saga of Djokovic at the Australian Open, [which will presumably be the first episode of Netflix's new tennis documentary, yes I can tell you they definitely got that](#), did not exactly increase the extent to which people's outer coverings contained glory.

First, there's Djokovic himself. As I understand it, he did the following.

1. Was already a player nobody liked.

2. Got Covid-19 early in the pandemic.
3. Refused to get vaccinated on grounds he'd already had Covid-19.
4. Refused to get vaccinated even though this was clearly going to cause endless trouble due to his need for constant international travel.
5. Took a generally strong anti-vaccine stance.
6. Reported he'd tested positive for Covid-19 again at exactly the time necessary to allow him to compete in both the Australian Open and the French Open, uh huh, yeah.
7. Was seen maskless shortly after the test among fans.
8. Made a mistake on his visa application and failed to coordinate the paperwork properly.
9. Got refused entry to Australia.
10. Had his family issue threats of at least symbolic physical violence.
11. Won a court case letting him enter Australia.
12. Also, oh yeah, he lied about his travels on his visa application.
13. Got deported from Australia.
14. Missed the Australian Open and his chance to set historical records.

So, yeah, whatever else is going on, let's be clear, f\*\*\* that guy. He played with fire while also pissing off the fire department, and then his house burned down.

That doesn't mean that it's fine for the government to go burn down his house. If it *does* head over with a casket of lighter fluid and burn the place down, we can notice that this is not an acceptable process, while still acknowledging that the fire-playing also was not cool and we're not that unhappy about the particular result.

How much will this matter? That depends how this happened, but more than that it depends on how people [think that it happened](#). Which narrative will take hold?



**President Dawg** @PresidentDawg · Jan 16

...

Benny don't surf.

Or think kangaroos and koalas are cute.



**Ben Domenech** ✅ @bdomenech · Jan 16

I feel like Australia underestimates how this Novak Djokovic incident is going to tar the entire nation as fundamentally anti-freedom, and a place you should not visit, for a long time to come.

[Show this thread](#)



**President Dawg** @PresidentDawg · Jan 16

...

Replies to [@PresidentDawg](#)

If a Salvadoran overstates her fear of MS-13 gang rape applying for an asylum visa in Mexico by all means send her back, but if a multimillionaire star athlete who has immigration lawyers at his beck and call fucking lies outright, it's "anti-freedom" to deport him.



**Ben Domenech**  @bdomenech · Jan 16

...

Replying to @bdomenech

For a nation that depends so much on tourism, it's a choice to embrace regime-wide authoritarianism that will definitely have negative economic consequences.

575

344

1.6K



**Ben Domenech**  @bdomenech · Jan 16

...

"If they'll do that to the best tennis player in the world, what do you think they'll do to you?"

4.4K

1.3K

2K



[My good friend Seth Burn thinks this is simple enough \(and yes, it stands\).](#)



**Seth Burn** @SethBurn · Jan 16

...

I'm sure there will be more legal action, but if this stands, I have two thoughts.

1. Holy shit.

2. Djokovic did this to himself and got what he deserved.



**Jon Wertheim**  @jon\_wertheim · Jan 16

Wow...."Australia confirms that Novak Djokovic banned from the country for THREE YEARS " | Daily Mail Online  
[dailymail.co.uk/news/article-1...](http://dailymail.co.uk/news/article-1...)



**Ben Swinerton** @Swinertontb · Jan 16

...

Replying to @SethBurn

What did he do to get banned from country? Not getting Vance's should do that. Shows tyrannical attitude of the country.

1

1



**Seth Burn** @SethBurn · Jan 16

...

It was a series of steps that got him banned, but the short version is he wasn't vaccinated and he lied on his visa application.

There certainly exist worlds where it is that simple, in which case banning him seems totally fine. [But do we live in such a world \(WaPo\)?](#) It's complicated.

After a weekend of hurried court hearings, a panel of three Australian federal justices unanimously upheld the immigration minister's decision to cancel Djokovic's visa on the grounds that his presence in the country might incite anti-vaccine sentiment and "civil unrest." The decision cleared the way for the government to deport him, ending his hope of competing in the Australian Open.

Although it would have been possible for Djokovic to appeal the ruling to Australia's High Court, the timing of Sunday's decision — roughly 24 hours before Djokovic was due to take to the tennis court — made another challenge unfeasible.

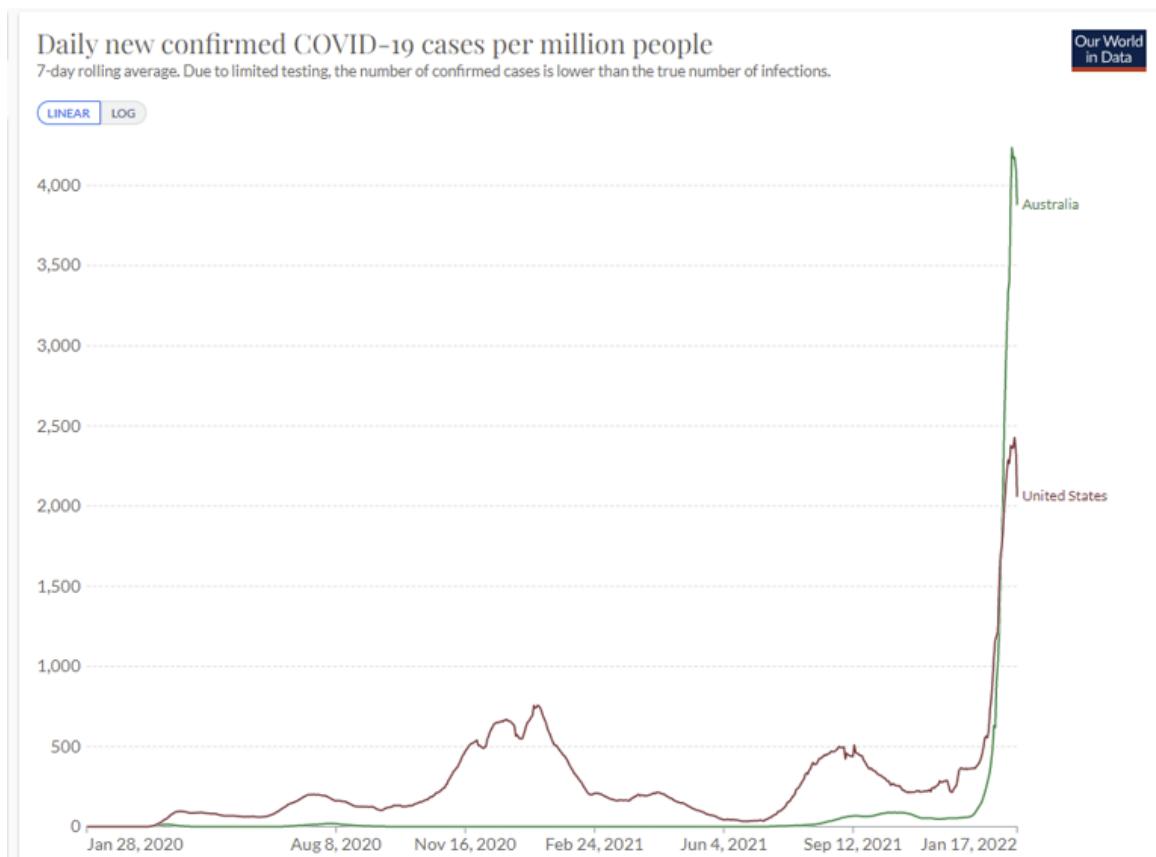
The timing is suggestive, but the correct focus here is on the grounds used: That his presence in the country might 'incite anti-vaccine sentiment and "civil unrest."

That is *very different* from saying that Djokovic lied on his application, or that he might get infected and spread the virus. The danger they're pointing to is his *symbolic meaning*.

Australian Prime Minister Scott Morrison praised the decision. "Australians have made many sacrifices during this pandemic," he said in a statement, "and they rightly expect the result of those sacrifices to be protected."

As usual, politicians praise courts when they get the results they like, and don't care about the rule of law or the principles involved. Ignoring the sunk cost fallacy issue, in what way is Djokovic's exile 'protecting those sacrifices?'

First off, I'm going to leave this graph here for no particular reason.



Whatever they were protecting, it wasn't the Australian people being protected from Covid infections. All that hard work over two years didn't fully go to waste, but Omicron is very much out of the box.

What they are protecting Australians from is something else.

Adding to the drama was Hawke's decision to put the unvaccinated player's personal beliefs on trial, arguing that his past anti-vaccine statements, behavior and huge platform as one of the world's biggest sports stars meant his continued presence in the country could incite anti-vaccine sentiment and "civil unrest."

That claim immediately came under criticism from Djokovic's attorneys on Sunday, who said the minister had unreasonably failed to consider what deporting the tennis star would do for civil unrest.

"It was just quite obvious that that in itself might be apt to generate anti-vax sentiment," Nick Wood said.

So at this point, the government is arguing that letting the world's best tennis player into the country to play tennis would *incite anti-vaccine sentiment* because they wouldn't punishing everyone equally for their failure to comply. The response *that seemed best to Djokovic's lawyer* was not to point out that this was absurd or a violation of rights, but rather to *question which decision would have the better symbolic meaning*.

Meanwhile, what is Djokovic doing that is so provocative or likely to incite sentiment? Seems mostly like it's 'not getting vaccinated.'

Wood also argued that the government had relied on old and selectively used quotes from Djokovic to describe his position on vaccines. Djokovic had been playing tennis around the world — including in Australia — for the past year of the pandemic without inciting unrest, he said. And the only evidence of a connection between Djokovic and anti-vaccine protests was the reaction to the government's own decision to deport him.

Stephen Lloyd, arguing for the immigration minister, said that Djokovic had done nothing to retract or change his stance, including in an [Instagram statement](#) last week, and that the fact he remained unvaccinated spoke volumes.

Indeed, it seems like there is very strong evidence that 'letting tennis player play some tennis' doesn't have the impact they are claiming, at least not of much magnitude, unless you decide to make a big deal out of it.

The judges clearly showed what matters in this Australian court, and it wasn't the rights of individuals or the rule of law.

Two of the three justices appeared concerned with whether the minister had fully weighed the potential outcome.

"One could see a situation where it was plain to anyone with common sense that canceling the visa would cause overwhelming public discord and risks of transmission through very large public gatherings," Allsop said.

Then there's this obvious nonsense.

"Given Mr. Djokovic's high profile status and position as a role model in the sporting or broader community, his ongoing presence in Australia may foster similar disregard for the precautionary requirements following receipt of a positive covid-19 test," he wrote.

So basically, failures to obey authority are likely to lead to arbitrary other failures to obey authority, so authority must always be obeyed no matter whether the rules in question make any sense, since some of the rules make sense.

Australia could have made a principled decision that Djokovic lied on his visa application, and those who lie on their application get deported. They could have made a principled

decision that there is no exemption for recently having Covid-19, or that they didn't believe his claim that he'd had it. Or they could claim that they admit people into the country as inherently political decisions at their sole discretion, and been straight about it.

They didn't do those things. Instead, they made absurd consequentialist claims that entail the right to enforce arbitrary punishments for failures to kowtow to authority, as the guiding principle of the state, as a matter of principle. They *want* this to be common knowledge.

What we do about it is up to us.

On the bright side, [there's at least two things we can all agree on](#).



**Harry Shearer** @theharryshearer · 20h

Novak Djokovic is an intransitive Serb.

64

168

1.1K

...

And this bit of trivia.

**NOVAK DJOKOVIC IS THE FIRST  
PLAYER TO BE KNOCKED OUT OF  
A GRAND SLAM TOURNAMENT  
*AFTER MISSING ONLY TWO SHOTS.***



OCCUPY DEMOCRATS [f](#) [i](#) [t](#)

## Think of the Children

So, I can confirm that this happened this week.

1. 75% of the staff of a school calls in sick due to Covid-19 (including quarantines).
2. This happens in the morning, so the school says it is *technically* open since some kids have already left and have nowhere to go, but there's no actual school happening, tells everyone to please stay home.
3. 15% of parents actively ask in the morning about whether, if their kid doesn't show up, *they will be marked as absent*.
4. Because, you see, absences might go on your *permanent record*, and also make it more difficult to get *reimbursed* when the time comes to *sue the city* to reclaim the massive tuition bill you spent on the assumption you'd mostly be able to recover it, but which the city forces you to sue over in the hopes they might not have to pay or you'll give up and send your kid to a normal public school instead.
5. This raises the question of whether they would *send their kids to this non-school* in order to avoid being marked as absent, if it came to that.

As a reminder, this is a big deal, [if you don't have an official 'remote option' or other way to avoid these 'unexcused absences'](#) then the system is effectively threatening to derail a student's life if they don't show up for class, in addition to any financial issues. Report to serve your time, at the time we say, *or else*, Covid exposures be damned.

[The good news is that Tuesday was the first week that those 'learning from home' could be marked as attending without a positive Covid test in the NYC school system](#). Which led to a recent high of almost 82% attendance thanks to the changing definition. That still seems quite low.

What is school about, again?

[Freddie deBoer tells it like it is](#), while explaining one of the core problems with closing schools.

Here's a basic point I've been making for at least a dozen years, including in my book, and will now do again: the educational function of public schools, while certainly of prime importance, is the secondary function of public schools. The first function is giving children warm, safe places where they can be stimulated and looked after, and where they can access cheap or free meals if they need them. The humanitarian good of this function dwarfs that of the education function.

I would also challenge exactly what the 'educational function' is teaching children, but either way it is good to remember what is going on here, and that this is a place we require children to be or else at some times, then closed down at others.

And I'd challenge the 'fine, then keep *your kids* home, then, if you want' attitude because the school system very much *does not* generally offer any kind of flexibility to children. You show up, where you are told to show up, and you are here delivering all the proper passwords to the proper authorities, *or else*. Even when the 'or else' is a subtle one, it's very much there, and carries many days.

[No, professor, your students are not primarily there to be 'vectors of disease.'](#) (WaPo). Hell of a story.

[Nor is it about obsessing over Covid-19 case rates in the places most protected against Covid](#), but those places don't seem to agree, yet they often don't share their case data, which is probably good since the value of information is strongly negative.

In other news, we learned via MR that [children as young as nine are launching DDoS attacks on their schools](#), with rates doubling from 2019 to 2020. They blame games where 'disrupting opponents is viewed as just another way to win' and worry about this leading to a life of crime. I would instead be happy the kids are practicing useful skills, and notice that if they're turning their attention to messing up the school's internet connection more often, maybe there are a large number of obvious reasons why they might want to do that.

## **Permanent Midnight**

[A theory of permanent masking.](#)



Andrew Lee @andrewleedr · Jan 14

...

Been asked how long we need to use face masks for. My answer -think of it like an umbrella which you take out to use on certain occasions (high risk settings, lots of community infection etc). It's raining heavily just now.



72

814

3.9K



One can interpret this in a number of ways. In the sensible metaphor, you keep masks *in your house* and take them out *when you are sick* or when there is *another pandemic*. That makes perfect sense, and I support this idea, so we won't have a mask shortage if and when this happens again. I do think from the details here that this is the intention.

One can also interpret this type of thing as a *request to always carry a face mask, forever, and use them in some places like transport, forever*. That makes a lot less sense.



science lover @PolyEyebrow · Jan 15

...

Replying to @andrewleedr

Except that carrying around an umbrella doesn't appreciably lower one's quality of life. And isn't mandated.

6

1

20



Brook Valentine @brookmeister · Jan 15

...

No, but being outside without an umbrella, as rain clouds gather or while it's actually chucking it down in a storm, make you look naive, unprepared & a bit silly. Live your life, keep a mask in your pocket or bag, pop it on in "stormy" places. Easy.

It's interesting how well metaphors like this reveal thinking. Does being outside without an umbrella *as rain clouds gather* reflect poorly on you? Absolutely not, and unless it's going to be major storm, anyone going only a short distance is probably making the right decision.

I'd also note that *has anyone tried carrying around an umbrella all the time?* Or actually thought about the decision on whether to carry one on a given day? It's highly annoying. Unless you're already using a backpack or other convenient carrying device and have one that folds up nicely, it means one of your hands is busy and you have another thing to remember all the time. It's a non-trivial cost, which is why people often get caught without an umbrella. A universal mandatory-umbrella-carrying social norm would be rather expensive and stupid.

Also, as my father often said, do you know what happens when you get wet? You dry off.

And that's why we mostly *don't* mock people who get caught without an umbrella. And why those who *do* mock such people are what we like to call *assholes*.

## Long Covid

Some of these deserve more detailed treatment, I intend to address that in another post when I have time. Hopefully next week.

[A request for anyone who had Long Covid then got monoclonal antibodies](#), your experiences could help justify a new clinical trial. If this is you, please do help.

[A new paper explores potential mechanisms for neurological Long Covid symptoms.](#)

[Long Covid patients lack native T and B cells](#) according to a new study, suggesting a mechanism.

[Report from Israel](#) that breakthrough cases of Covid-19 after double vaccinations are not more likely to lead to Long Covid symptoms than not having been infected at all. So weird to say ‘we know it can happen after mild infection’ when there’s a ‘baseline rate’ of it happening *without an infection at all*, even in the worlds where it is mostly the result of Covid-19.

Katja Grace, whose thinking I generally respect, [has a post making the case for taking Long Covid more seriously](#). You have to consider the post’s arguments *in light of it being a steelman case for taking Long Covid seriously*, from a reference class of thinkers I’ve observed to be likely to be inclined to take such things seriously. One does still have to also update on her decision to write the post.

## In Other News

[Scott Gottlieb suggests an obvious improvement](#), to combine the FDA and CDC advisory panels. There’s no reason to have two such veto points. At most we want one.

He also reports that [people are getting rid of their pet hamsters](#) after Hong Kong ordered its hamsters killed. Please don’t do this, especially if you have kids. The Covid-19 risk here is at most minimal.

[And he is doubtful that China suspending almost all international travel can be sustained for that long](#). I’m in the ‘they can sustain this for a while’ camp for this particular intervention, to me the question is whether it will be enough, and also whether they’re willing to make that permanent.

[Various views of DeStantis’s Covid-19 record in Florida](#), it’s not clear who is trying to claim they did what and why but it’s going to be fun watching them claim all of it.

Via MR, [an archived \(now deleted\) Twitter thread](#) with claims about the combination of remote work, complete lack of social connections outside of work, over the top pandemic fear and a highly pro-woke corporate culture resulting in deeply dysfunctional and unhappy employees at a major tech company. Source has obvious agenda but no reason I can see to be faking the reports of general Covid-19 paranoia and resulting deep unhappiness.

This madness must end. Life must resume, and soon. As much as I’m sympathetic to ‘this is not the week to do that’ one must reply with [some version](#) of ‘[if not now, when?](#)’ and have a *concrete answer*. Next week or two weeks from now is *totally acceptable* as an answer to when, *if you mean it*. Under a hundred thousand cases a day is too conservative but understandable. Real Soon Now is not.

We also need to stop the madness of [not letting people see their dying loved ones](#). I do not know how common this is, but it is mind-boggling that even now this is not a decision you are allowed to make for yourself. Damn right I’d demand to see my parents if they were dying of Covid-19. If I get it, I get it.

[Report from a nursing home where testing was reportedly actively sabotaged to prevent positives and keep staff coming in.](#) If true, someone please nail these bastards.

[BBC overview of what happened with Ivermectin.](#)

A hospital reported to have used [a rather more aggressive approach to race-based care allocation](#) than the current system where you need any excuse at all and the right racial identification counts but so do lots of other things. In this case, you needed a total of 20 points, which was a real bar to clear, and racial identification could net you 7, whereas most conditions max out at 6. They caved under legal pressure.

[Thread on ethics of sharing anecdote.](#) Doctor who has concerns contrary to party line (and contrary to my math calculations) points out that sharing anecdotal information that supports party line ('Omicron isn't mild') is widely praised, whereas information opposed to party line ('boosters can have rare side effects') is widely condemned.

This is because what people care about is the [Simulacra-Level](#)-2 results of information sharing combined with the associated Level-3 affiliations. If you are shown Scary Medial Picture, that's presumably because You Should Be Scared Of This Thing, and that's the message a lot of people get, full stop. There's also a Level-1 concern about people's resulting models becoming inaccurate. Sharing *rare* things can effectively be misleading to humans who use the availability heuristic and can't handle base rates, whereas sharing examples of common things is helpful.

These are all valid things to be concerned about. We also need to be concerned about using such concerns to entirely suppress information that goes against the desired narrative, especially if rare concerns are impactful enough that they may dominate the calculus. By default these concerns get weaponized in exactly this way, forcing all information sharing to serve authority.

In *this particular case*, I basically classify frequently sharing and harping on the particular information in question as Not Helping. I understand the source 'means well' but source is doing what I can only call *harping* on concerns where the math very much does not justify the observed level of harp, either in general or especially by this source. I sympathize with the condemnations. I view the effort as persuasion via 'getting around' the math rather than providing information that lets us do math better. Of course, my preferred intervention, should this get sufficiently out of hand, is fully covered by the unfollow button.

## Not Covid

[Bryan Caplan does not place enough bets, and bets too small.](#)

[New paper shows](#) that EBV (the virus that causes "mono") is the *leading cause* of multiple sclerosis ([paper](#), gated). I haven't read the paper but everyone seems to think it is the real deal, and the basic case seems ironclad. This is huge. A million Americans have MS, and this suggests such mechanisms are common elsewhere too. We should update our priors about how bad it is to catch viruses in general, and of course EBV in particular, and also have new hope we may be able to prevent or cure MS down the line. If nothing else, a vaccine for EBV would also prevent MS, and hey look, we have this new nifty mRNA tool for making vaccines.

[The latest Very Serious Person message](#) in its correct form.



Jonathan Newman @NewmanJ\_R · Jan 14

...

All you people have been tricked into thinking mass formation psychosis is a real thing.

92

63

583



[FedEx asks permission to add anti-missile lasers to their planes.](#) Our cyberpunk future is that much closer, but if it were here FedEx wouldn't ask for permission.

Music consumption is focusing more on older music. This is good in the sense that music that is older tends to be better because age filters the good from the bad, and also because the exhaustion of the search space combined with various forms of over-optimization and over-commercialization means that new music tends to start out worse as well. It is bad to the extent that it means we are under-investing in new music. Since artists gather only a small portion of gains from new music, we should be suspicious we are under-investing, but also artists vastly over-invest in new music creation compared to their potential profits due to intrinsic motivations, social payoffs and delusional aspirations, so this could go either way. Mostly, it seems like the amount of new music isn't down much whereas older music consumption is up a lot, and in the long run the potential for long-run profits should provide more incentive to create new music, so it all seems fine. My bigger concern is that it seems as if the range of old music is narrowing, but perhaps this is indeed welfare-maximizing.

[Eye emoji.](#)



Al Yankovic ✅ @alyankovic · 1h

...

So... I'm making a movie.

1.2K

1.5K

16.1K



# Thought Experiments Provide a Third Anchor

Previously, [I argued](#) that we should expect future ML systems to often exhibit "emergent" behavior, where they acquire new capabilities that were not explicitly designed or intended, simply as a result of scaling. This was a special case of a general phenomenon in the physical sciences called More Is Different.

I care about this because I think AI will have a huge impact on society, and I want to [forecast what future systems will be like](#) so that I can steer things to be better. To that end, I find More Is Different to be troubling and disorienting. I'm inclined to forecast the future by [looking at existing trends](#) and asking what will happen if they continue, but we should instead expect new qualitative behaviors to arise all the time that are not an extrapolation of previous trends.

Given this, how can we predict what future systems will look like? For this, I find it helpful to think in terms of "anchors"---[reference classes](#) that are broadly analogous to future ML systems, which we can then use to make predictions.

The most obvious reference class for future ML systems is current ML systems---I'll call this the **current ML anchor**. I think this is indeed a pretty good starting point, but we've already seen that it [fails to account](#) for emergent capabilities.

What other anchors can we use? One intuitive approach would be to look for things that humans are good at but that current ML systems are bad at. This would include:

- Mastery of external tools (e.g. calculators, search engines, software, programming)
- Very efficient learning (e.g. reading a textbook once to learn a new subject)
- Long-term planning (e.g. being able to successfully achieve goals over months)

Models sufficiently far in the future will presumably have these sorts of capabilities. While this still leaves unknowns---for instance, we don't know how rapidly these capabilities will appear---it's still a useful complement to the current ML anchor. I'll call this the **human anchor**.

A problem with the human anchor is that it risks anthropomorphising ML by over-analogizing with human behavior. Anthropomorphic reasoning correctly gets a bad rap in ML, because it's very intuitively persuasive but has a [mixed at best](#) track record. This isn't a reason to abandon the human anchor, but it means we shouldn't be entirely satisfied with it.

This brings us to a third anchor, the **optimization anchor**, which I associate with the "Philosophy" or thought experiment approach that I've [described previously](#). Here the idea is to think of ML systems as ideal optimizers and ask what a perfect optimizer would do in a given scenario. This is where Nick Bostrom's colorful description of a [paperclip maximizer](#) comes from, where an AI asked to make paperclips turns the entire planet into paperclip factories. To give some more prosaic examples:

- The optimization anchor would correctly predict [imitative deception](#) (Lin et al., 2021), since a system optimized to produce high-probability outputs has no intrinsic reason to be truthful.

- It also would observe that power-seeking is instrumentally useful for many different goals, and so predict that optimal policies (as well as sufficiently powerful neural networks) will [tend to do so](#) (Turner et al., 2021).

Ideas produced by the optimization anchor are often met with skepticism, because they often contradict the familiar current ML anchor, and they don't benefit from the intuitive appeal of the human anchor. But the differences from these other two anchors are precisely what make the optimization anchor valuable. If you (like me) feel that both the current ML and human anchors paint an incomplete picture, then you should want a third independent perspective.

The optimization anchor does have limitations. Since it abstracts ML into an ideal optimizer, it ignores most on-the-ground facts about neural networks. This can lead to underconstrained predictions, and to ignoring properties that I think will be necessary for successfully aligning ML systems with humans. I'll say more about this later, but some particularly important properties are that neural networks often generalize in "natural" ways, that we can introspect on network representations, and that training dynamics are smooth and continuous. Researchers focused on the optimization anchor don't entirely ignore these facts, but I think they tend to underemphasize them and are overly pessimistic as a result.

## The Value of Thought Experiments

The optimization anchor points to the value of thought experiments more generally. While it poses the thought experiment of "What if AI were a perfect optimizer?", there are many other thought experiments that can provide insights that'd be hard to obtain from the ML or human anchors. In this sense thought experiments are not a single anchor but a generator for anchors, which seems pretty valuable.

One thought experiment that I particularly like is: *What happens if most of an agent's learning occurs not during gradient descent, but through in-context learning*<sup>[1]</sup>? This is likely to happen eventually, as ML agents are rolled out over longer time horizons (think artificial digital assistants) and as ML improves at in-context learning. Once this does happen, it seems possible that agents' behavior will be controlled less by the "extrinsic" shaping of gradient descent and more by whatever "intrinsic" drives they happen to have<sup>[2]</sup>. This also seems like a change that could happen suddenly, since gradient descent is slow while in-context learning is fast.

It would be great if we had a community of researchers making thought experiments with clearly stated assumptions, explaining in detail the consequences of those assumptions and ideally connecting it to present-day research.

## Other Anchors

There are many other anchors that could be helpful for predicting future ML systems. **Non-human animal behavior** could provide a broader reference class than humans alone. **Evolution** and **the economy** are both examples of powerful, distributed optimization processes. I am most excited about better understanding **complex systems**, which include biological systems, brains, organizations, economies, and ecosystems and thus subsume most of the reference classes discussed so far. It seems to me that complex systems have received little attention relative to their

germaneness to ML. Indeed, emergence is itself a concept from complex systems theory that is useful for understanding recent ML developments.

## Limitations of Thought Experiments

I've focused so far on *predicting* problems that we need to address. But at some point we actually have to *solve* the problems. In this regard thought experiments are weaker, since while they often point to important big-picture issues, in my view they fare poorly at getting the details right, which is needed for engineering progress. For instance, early thought experiments [considered a single AI system](#) that was much more powerful than any other contemporary technologies, while in reality there will likely be many ML systems with a continuous distribution of capabilities. [More recent](#) thought experiments impose discrete abstractions like "goals" and "objectives" that I don't think will cleanly map onto real ML systems. Thus while thought experiments can point to general ideas for research, even mapping these ideas to the ontology of ML systems can be a difficult task.

As a result, while we can't blindly extrapolate empirical trends, we do need a concerted empirically-based effort to address future ML risks. I'll explain why I think this is possible in a later post, but first I'll take us through an example of "taking a thought experiment seriously", and what it implies about possible failure modes of ML systems.

- 
1. In-context learning refers to learning that occurs during a single "rollout" of a model. The most famous example is [GPT-3](#)'s ability to learn new tasks after conditioning on a small number of examples. [←](#)
  2. While this statement borders on anthropomorphizing, I think it is actually justified. For instance, depending on the training objective, many agents will likely have a "drive" towards information-gathering, among others. [←](#)