

Best of LessWrong: May 2020

1. [The EMH Aten't Dead](#)
2. [An overview of 11 proposals for building safe advanced AI](#)
3. [Literature Review For Academic Outsiders: What, How, and Why](#)
4. [Comment on "Endogenous Epistemic Factionalization"](#)
5. [Movable Housing for Scalable Cities](#)
6. [Insights from Euclid's 'Elements'](#)
7. [Why Artists Study Anatomy](#)
8. [Assessing Kurzweil predictions about 2019: the results](#)
9. [Studies On Slack](#)
10. [Baking is Not a Ritual](#)
11. [A non-mystical explanation of insight meditation and the three characteristics of existence: introduction and preamble](#)
12. [Tips/tricks/notes on optimizing investments](#)
13. [What are your greatest one-shot life improvements?](#)
14. [Mazes Sequence Summary](#)
15. [OpenAI announces GPT-3](#)
16. [Get It Done Now](#)
17. [A non-mystical explanation of "no-self" \(three characteristics series\)](#)
18. [Why We Age, Part 2: Non-adaptive theories](#)
19. [Subspace optima](#)
20. [Why Rationalists Shouldn't be Interested in Topos Theory](#)
21. [How uniform is the neocortex?](#)
22. ["AI and Efficiency", OA \(44× improvement in CNNs since 2012\)](#)
23. [Maths writer/cowriter needed: how you can't distinguish early exponential from early sigmoid](#)
24. [The Oil Crisis of 1973](#)
25. [162 benefits of coronavirus](#)
26. [Failures in technology forecasting? A reply to Ord and Yudkowsky](#)
27. [Book Review: Narconomics](#)
28. [In Search of Slack](#)
29. [Covid-19: Comorbidity](#)
30. [GPT-3: a disappointing paper](#)
31. [How does publishing a paper work?](#)
32. [Plague in Assassin's Creed Odyssey](#)
33. [Trust-Building: The New Rationality Project](#)
34. [Writing Causal Models Like We Write Programs](#)
35. [Extracting Value from Inadequate Equilibria](#)
36. [SlateStarCodex 2020 Predictions: Buy, Sell, Hold](#)
37. [Zoom Technologies, Inc. vs. the Efficient Markets Hypothesis](#)
38. [Corrigibility as outside view](#)
39. [The principle of no non-Apologetics](#)
40. [How to \(not\) do a literature review](#)
41. [Craving, suffering, and predictive processing \(three characteristics series\)](#)
42. [Speculations on the Future of Fiction Writing](#)
43. [AGIs as collectives](#)
44. [Conjecture Workshop](#)
45. [Your abstraction isn't wrong, it's just really bad](#)
46. [\[AN #100\]: What might go wrong if you learn a reward function while acting](#)
47. [How can nonprofits gain the advantages of the for-profit model?](#)
48. [Our Need for Need](#)
49. [Why aren't we testing general intelligence distribution?](#)
50. [Pointing to a Flower](#)

Best of LessWrong: May 2020

1. [The EMH Aten't Dead](#)
2. [An overview of 11 proposals for building safe advanced AI](#)
3. [Literature Review For Academic Outsiders: What, How, and Why](#)
4. [Comment on "Endogenous Epistemic Factionalization"](#)
5. [Movable Housing for Scalable Cities](#)
6. [Insights from Euclid's 'Elements'](#)
7. [Why Artists Study Anatomy](#)
8. [Assessing Kurzweil predictions about 2019: the results](#)
9. [Studies On Slack](#)
10. [Baking is Not a Ritual](#)
11. [A non-mystical explanation of insight meditation and the three characteristics of existence: introduction and preamble](#)
12. [Tips/tricks/notes on optimizing investments](#)
13. [What are your greatest one-shot life improvements?](#)
14. [Mazes Sequence Summary](#)
15. [OpenAI announces GPT-3](#)
16. [Get It Done Now](#)
17. [A non-mystical explanation of "no-self" \(three characteristics series\)](#)
18. [Why We Age, Part 2: Non-adaptive theories](#)
19. [Subspace optima](#)
20. [Why Rationalists Shouldn't be Interested in Topos Theory](#)
21. [How uniform is the neocortex?](#)
22. ["AI and Efficiency", OA \(44x improvement in CNNs since 2012\)](#)
23. [Maths writer/cowriter needed: how you can't distinguish early exponential from early sigmoid](#)
24. [The Oil Crisis of 1973](#)
25. [162 benefits of coronavirus](#)
26. [Failures in technology forecasting? A reply to Ord and Yudkowsky](#)
27. [Book Review: Narconomics](#)
28. [In Search of Slack](#)
29. [Covid-19: Comorbidity](#)
30. [GPT-3: a disappointing paper](#)
31. [How does publishing a paper work?](#)
32. [Plague in Assassin's Creed Odyssey](#)
33. [Trust-Building: The New Rationality Project](#)
34. [Writing Causal Models Like We Write Programs](#)
35. [Extracting Value from Inadequate Equilibria](#)
36. [SlateStarCodex 2020 Predictions: Buy, Sell, Hold](#)
37. [Zoom Technologies, Inc. vs. the Efficient Markets Hypothesis](#)
38. [Corrigibility as outside view](#)
39. [The principle of no non-Apologetics](#)
40. [How to \(not\) do a literature review](#)
41. [Craving, suffering, and predictive processing \(three characteristics series\)](#)
42. [Speculations on the Future of Fiction Writing](#)
43. [AGIs as collectives](#)
44. [Conjecture Workshop](#)
45. [Your abstraction isn't wrong, it's just really bad](#)
46. [\[AN #100\]: What might go wrong if you learn a reward function while acting](#)
47. [How can nonprofits gain the advantages of the for-profit model?](#)

48. [Our Need for Need](#)
49. [Why aren't we testing general intelligence distribution?](#)
50. [Pointing to a Flower](#)

The EMH Aten't Dead

Cross-posting from [my personal blog](#), but written primarily for Less Wrong after recent discussion here.

There are whispers that the [Efficient-Market Hypothesis](#) is dead. Eliezer's faith has been [shaken](#). Scott [says](#) EMH may have been the *real* victim of the coronavirus.

The EMH states that "asset prices reflect all available information". The direct implication is that if you don't have any non-available information, you shouldn't expect to be able to beat the market, except by chance.

But some people were able to preempt the corona crash, without any special knowledge! Jacob [mentioned](#) selling some of his stocks before the market reacted. Wei Dai bought out-of-the-money 'put' options, and took a very [handsome profit](#). Others shorted the market.

These people were reading the same news and reports as everyone else. They profited on the basis of public information that should have been priced in.

And so, the EMH is dead, or dying, or at the very least, has a very nasty-sounding cough.

I think that rumours of the death of efficient markets have been greatly exaggerated. It seems to me the EMH is very much alive-and-kicking, and the recent discussion often involves common misunderstandings that it might be helpful to iron out.

This necessarily involves pissing on people's parade, which is not much fun. So it's important to say upfront that although I don't know Wei Dai, he is no doubt a brilliant guy, that Jacob is my favourite blogger in the diaspora, that I would give my left testicle to have Scott's writing talent and ridiculous work ethic, that Eliezer is a legend whose work I have personally benefited from greatly, etc.

But in the spirit of the whole rationality thing, I want to gently challenge what looks more like a case of 'back-slaps for the boys' than a death knell for efficient markets.

First: how the heck did the market get the coronavirus so wrong?

The Great Coronavirus Trade

Lots of people initially underreacted to COVID-19. We are only human. But the stockmarket is *not* only human—it's meant to be better than this.

Here's Scott, in [A Failure, But Not of Prediction](#):

The stock market is a giant coordinated attempt to predict the economy, and it reached an all-time high on February 12, suggesting that analysts expected the economy to do great over the following few months. On February 20th it fell in a way that suggested a mild inconvenience to the economy, but it didn't really start plummeting until mid-March – the same time the media finally got a clue. These aren't empty suits on cable TV with no skin in the game. These are the best predictive institutions we have, and they got it wrong.

But... this isn't how it went down. As AllAmericanBreakfast and others pointed out in the comments, the market started reacting in the last week of February, with news headlines directly linking the decline to the 'coronavirus'. By the time we get to mid-March, we're not far off *the bottom*.

(You can confirm this for yourself in a few seconds by looking at [a chart](#) of the relevant time period.)

EDIT: Scott has explained his rationale [here](#). Although I still think his version of events is incorrect as phrased, I want to make it clear I am not accusing him of deliberately massaging the data or any other such shenanigans, and the next paragraph about revisionist history etc was only meant to be a general observation about how people responded. My apologies to Scott for the unclear wording, as well any perceived slight against his very good reputation.

For whatever reason, COVID-19 seems to be a magnet for revisionist history and/or wishful thinking. In other comments under the same post, the notion that people from our 'tribe' did especially well also comes under serious question—in fact, it looks like many of the names mentioned seem to have jumped on the bandwagon after it was obvious, and certainly long after the market was moving.

The facts are that *the market reacted faster than almost all of us*. But not before a few prescient people placed their bets!

So now the question becomes: why didn't the market react earlier than February 20, like those smart people did?

The null hypothesis is that the market reacted exactly appropriately on the basis of the information available. After all, there were other potential pandemics in the recent past that were successfully contained or eradicated.

On February 20, there were only 4 known cases in Italy. We were a long ways from the bloodbath that was coming. Maybe it was correct to move cautiously until further information came in?

Here's keaswaran:

EDIT: previously misattributed to AllAmericanBreakfast (who agrees with it).

[At that time] it may have been plausible for many people to think this would continue to play out like SARS – East Asia would solve their problem, everyone else would watch airport arrivals and quarantine them effectively, and within a few weeks everything would stabilize and gradually go away. By Feb. 27 it was clear that this wasn't happening, since community spread was very clear from the public data in Italy and Iran, and probably also clear from genetic data in the United States and elsewhere.

So we reached a tipping point in those next few days, at which point, the market started responding more vigorously.

If the null hypothesis is true, then those early trades were not quite as prescient as they look. We might be making the mistake of 'resulting', and confusing the reality we ended up in with all the others which were possible at the time, in which those traders lost their shirts. It's really hard to have a useful object-level discussion about this, because these events are one-offs (this is the same argument we get into every year on Scott's predictions threads!) It's not like we can run the experiment again, and thank goodness for that.

Nevertheless, Wei Dai suggested that this was the final nail in the coffin of EMH—at least for him.

I want to pause here to give mad props to Wei Dai for being totally open about everything, and especially for doing the following:

- warning that options are dangerous, and you can easily lose the lot
- generic disclaimer about seeking financial advice
- updating the thread after he ended up losing 80% of his paper profits

- mentioning selection bias, and saying we'd be right to discount his evidence

Which only leaves the [initial claim](#) that "at least for me this puts a final nail in the coffin of EMH."

This is a polite way of hinting that you might be a brilliant investing wizard with the power to beat the market. Honestly, after making such a beautiful trade—and my gosh it really was beautiful—whom amongst us could resist that temptation? Certainly not me. And anyway, it might even be true!

In making sense of claims of this nature, the first thing we have to establish: what does it even mean to be able to beat the market?

Can Uncle George Beat the Market?

Uncle George really likes his new iPhone. Man, these things are nifty! The dancing poop emoji is hilarious. On the strength of this insight, George dials his broker and loads up on AAPL stock.



the chosen one! the scourge of efficient markets! the stuff of eugene fama's nightmares!

Over the next year, AAPL stock goes up 15 per cent, while the broader S&P 500 only goes up 10 per cent. George becomes insufferable at family dinners as he holds forth on his stock-

picking powers. Guess the market isn't so 'efficient' after all, huh. Suck it, Eugene Fama!

So: did Uncle George beat the market?

In the narrowest possible sense... yes.

In the sense in which we aim to string words together so that they mean things: no, of course not. By this definition, every single trade leads to one of the two parties 'beating the market'. Millions of people beat the market while I wrote this sentence. I can flip a coin between Pepsi and Coke right now, and have a 50 per cent chance of becoming a market-beating genius.

The Uncle George example makes it glaringly obvious that a successful trade does not somehow 'break' efficient markets. And yet, this is the same naive criticism constantly leveled against the EMH: if the market moves in literally any direction, that must mean it was *wrong* before! My cousin who sold/bought before it went up/down beat the market!

Same goes for the Great Coronavirus Trade. The fact that some people got out of the market early is not even the tiniest bit surprising. Investors *constantly* think the market is going to crash, for any number of plausible reasons. This is the default state of affairs: we have successfully predicted 73 of the last five market crashes, etc.

These predictions are almost always wrong:



And almost all the people who make them would have been much better off taking the [boring 'buy and hold forever' strategy](#).

But even a stopped clock is right twice a day. And of course, we're much more likely to hear about the occasional brilliant successes than the near-constant dull failures.

So the most naive criticism of the EMH boils down to 'it's possible to make a good trade'. This is just a property of trading. It tells us exactly nothing about the market's efficiency, or lack thereof.

But some people really *do* beat the market—and not in the trivial sense. They're not merely stopped clocks, or highly visible 'survivors'. I'll suggest a definition later on which strips out

the effect of randomness.

Before we get there—doesn't the concession that people can non-trivially beat the market already drive a stake through the EMH's heart?

This is the second great misunderstanding: there is no conflict between the EMH and beating the market. *That's how the market gets efficient!* You find an information asymmetry that isn't priced in yet, and in exploiting it, you move the market a little further towards efficiency.

Let's call this information asymmetry an 'edge'.

If the EMH is true—or even just true-ish—that doesn't mean the market can't be beat. It means:

You shouldn't expect to beat the market without a unique edge, except by chance

Now, this usually gets simplified down to 'you can't beat the market'. And most of the time, this simplification is good enough: you might get lucky and win in the Uncle George sense, but over an investing lifetime, you'll almost certainly revert to the mean (which isn't matching the market return—it's underperforming it).

But if you *can* find some kind of edge, you really can win! So, what might a genuine edge look like?

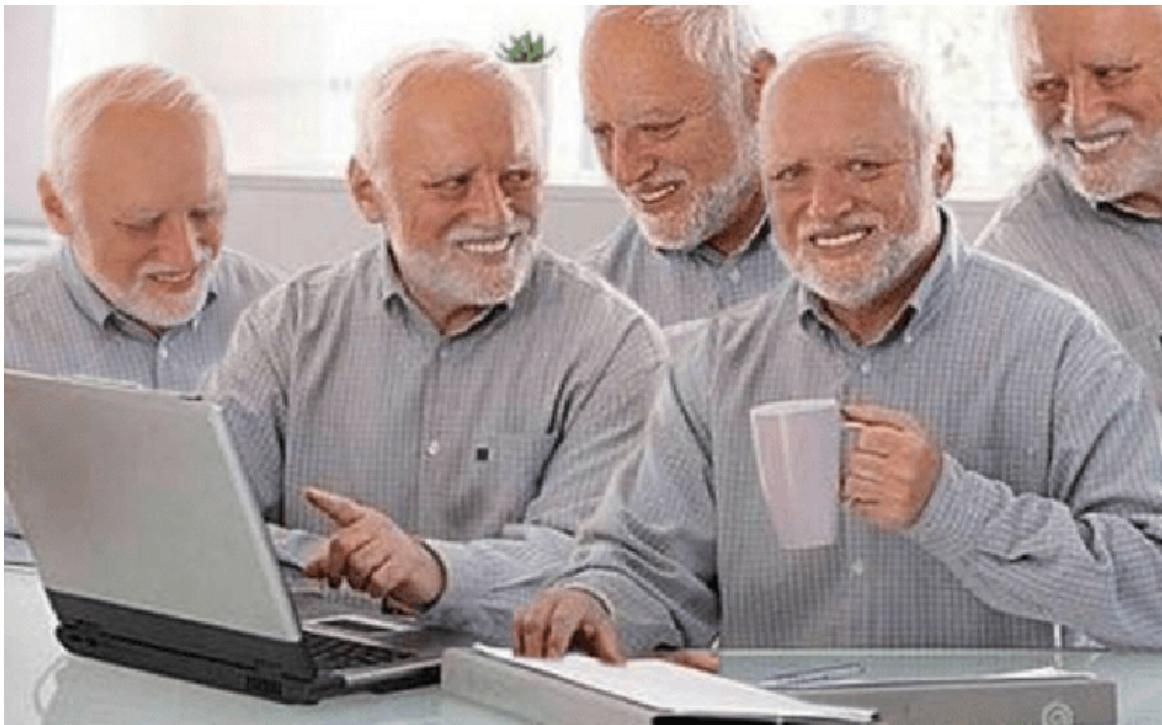
Anomalies Exist!

The Uncle Georges of the world don't have an edge. All of their thoughts have already been thunk by someone else (probably by millions of someone elses). Instead, their fortunes are entirely at the mercy of the myriad other forces that drive stock prices: consumer demand, workplace harassment scandals, money printers going brrr, the exact virulence of a novel coronavirus, the price of cheese in Spain last Friday afternoon, etc.

All of this stuff—billions of inputs processed by the greatest collective intelligence ever built—is a black box unto us mere mortals. It's impossible to assign perfect causal explanations to stock prices, which means we can pick whichever story suits us best. As a market reporter, this was pretty much my whole job: calling brokers and economists to wrap a plausible narrative around totally inexplicable events, and generate sage nodding of heads.

All Uncle George can see is that he placed his bet, and AAPL went up. It was the poop emoji for sure!

And so, Uncle George spends his days dishing out hot stock tips on online forums, oblivious to the fact that his success was meaningless.



[uncle georging intensifies]

What does a real edge look like?

A century ago, investors started noticing they could consistently pick up bargains by running very simple formulas over stock prices. The most famous is the ‘value investing’ approach developed by Ben Graham, and used by Warren Buffett and Charlie Munger. There was a genuine, big old inefficiency in the markets, and these guys had a great time exploiting it.

I think this might be the image most people have in their head when they think of ‘beating the market’—diligently studying *The Intelligent Investor* and learning about PE ratios or whatever.

But this is like trying to use a stone-age axe against a fighter jet. The Ben Graham information asymmetry has long since disappeared because...markets are efficient(ish)! Once the formula was widely known, it stopped working. Investors developed more sophisticated versions, more formulas, more pricing models. Once those got out, they stopped working too. Now there’s a great debate as to whether even the most complicated descendants of value might be totally dead. In which case, the anomaly has officially gone for good.

Either way, this is *not* how Buffett gets his edge, and it hasn’t been for decades. Here’s Munger:

The trouble with what I call the classic Ben Graham concept is that gradually the world wised up and those real obvious bargains disappeared. You could run your Geiger counter over the rubble and it wouldn’t click.

Buffett’s most brilliant achievement is weaving this folksy legend that he is a cute old grandpa who beats the market by backing the best companies. Let’s take a look at how market-beating investors *really* make their money.

Modern Edges are Completely Ridiculous



greatest showman on earth

1. The Warren Buffett Halo Effect

In recent decades, Buffett has made a killing through juicy private deals which are completely out of reach of the average investor. Like, six billion dollar deals with three billion in preference rights and a guaranteed dividend. Like, lobbying the government to bail out the banks, then carving off a huge piece of the action. Like, being able to play around with Berkshire Hathaway's \$115 billion insurance float. Much of his fortune is built on taxpayer largesse.

Warren Buffett's brand is so powerful that at this point, his success is a *self-fulfilling prophecy*: when Berkshire invests in a stock, everyone else piles in after him and drive the price up. Buffett even lends out his 'halo' to companies that need it—most famously during the GFC—so long as they give him a generous discount to the market price, of course. (Matt Levine has written some [fascinating](#) columns about this).

And yet, and yet... Berkshire Hathaway has *underperformed* for the last decade. Buffett would have been better off to take his own advice and put it all in index funds.

2. Hedge funds with armies of drones

There you are, sitting in your home office going through Walmart's quarterly report and calculating PE ratios or whatever. Meanwhile, the professionals are using *an army of drones* to monitor the movement of shopping carts in Walmart parking lots in real time.

See also: sending foot soldiers out to every branch of a bakery chain at the close of business each day, because the numbered dockets start out at zero, and thus contain live sales data unavailable to the market.

And so, when renowned hedge fund manager Michael Steinhardt was asked the most important thing average investors could learn from him, here's what he suggested:

"I'm their competition."

And yet, and yet...almost all hedge funds underperform. Not all of them are trying to beat the market, but the tools at their disposal gives us a sense of the difficulty here.

3. High-frequency traders move mountains

If multiple people have access to the same information, the speed in which you can bring it to market also matters. So, we have high-frequency traders.

One firm spent \$300 million laying a direct cable between Chicago to New Jersey. They cut straight through mountains and crossed rivers. The cable stretched 1331 kilometres. And they did this to shave *four milliseconds* off their transmission time.

And yet, and yet...microwaves came along and rendered the whole project obsolete. Trying to get an edge is expensive.

4. Being willing and able to commit felonies

Insider trading is a thing. See also: criminals who hack or otherwise steal sensitive private information.

And yet, and yet...even when criminals have advance access to earnings reports, they *still* don't do all that well, which is evidence for the very strongest form of the EMH (the one that no-one, including me, believes can possibly be true).

So...what sort of edge do us lesser mortals have?

If we mumble something about having 'good intuition', or 'subscribing to the *Wall Street Journal*' then we should consider the strong possibility that we are Uncle George.

If the answer involves 'fundamental analysis' or 'Fibonacci retracements', we're *still in Uncle George territory*. The only difference is that doing something complicated makes it easier to internally justify the belief that we know a secret no-one else does. But it is still (probably!) a mistaken belief.

The EMH Gets Stronger With Every Attack

So we know for sure that market-beating edges exist—I've even written them down for everyone to see!

I can only dream of possessing a halo effect so strong that everyone piles into a stock right after I announce I have graced it with my favour. I don't have an army of drones at my command, or the ability to bore through mountains to shave milliseconds off my trading times, or a weekly round of golf with the CEO of a Fortune 500 company.

The market is never perfectly efficient. But relative to me, *it might as well be*.

Critics have pointed out plenty of cases in which the EMH doesn't jive with reality—and they are absolutely right. So this is where it gets really weird.

The EMH is the only theory that grows *stronger* with every attack against it.

Every edge is constantly at risk of being gobbled up by an efficient-ish market. The ones I've mentioned can only be publicly known because they're somewhat defensible: they're based on personal relationships, capital investment, proprietary technology, etc. But they disappear too.

Most edges *can't even be spoken out loud* without disappearing. If stocks systematically rise on the third Thursday of each month but only under a waxing moon, and then someone writes about it in public, you can kiss that anomaly goodbye. The EMH sucks it into its gigantic heaving maw, and it's gone forever.

In other words: every time someone picks a hole in the theory and points out an inefficiency, they make the predictions generated by the EMH *more robust*! It's like some freaky shoggoth thing that Just. Won't. Die.



you may not like it, but this is what peak efficiency looks like

Which gets us to the totally justified criticism of the theory: the only reason the EMH can pull this stunt is because it's *bullshit science*.

It's unfalsifiable! It responds to criticism by saying, 'OK, good point, but now that I've factored that in, you should believe in my theory even more.'

And...we really should?

The only way I can think about the EMH without going insane is to remind myself that it generates a useful *heuristic*. It's not a stable law, like we might find in hard sciences. It's not perfectly accurate. At any given point in time, there are always competing models that do a better job of describing reality. But all those other models can stop working at any moment, with no warning! By the time you find out their predictive power is gone, it's too late, and you probably lost a bunch of money! By contrast, the EMH is a reliable model—reliably vague and hand-wavy, yes, but also reliably useful.

We know there are inefficiencies in the market. In the fullness of time, they will be absorbed into the gelatinous alien-god's hivemind. But before that happens, maybe we can make some money off of them.

So now we come to the final test. How do you tell if you've *really* found a market-beating edge—that is, a model of reality that has more predictive power than the EMH—or you're fooling yourself like Uncle George?

If You're so Smart, Why Aren't You Rich?

Everyone knows a secret about themselves, or the people they know well, or can arbitrage some opportunity in a niche that few people are paying attention to. These illiquid private 'markets' are much more fertile hunting grounds for asymmetries, and something I encourage everyone to think about.

But the *public security markets* are a gigantic agglomeration of everyone's predictions, which constantly hoovers up every new fragment of information, and recalibrates itself in real time. Your challenge is to try and predict why this giant meta-prediction is wrong, and in which direction.

If you think you can reliably beat, say, an index fund that passively tracks the S&P 500, this is a much stronger claim than it first appears.

For one thing, you're claiming to be better than Warren Buffett, who has failed to pull this off in the last 10 years, despite his huge advantages, and has started saying the game is so hard that everyone should just buy index funds. But that's nothing. You are also claiming that you have the power to beat the greatest collective intelligence humanity has ever created.

This is an *extraordinary claim*, and the thing about extraordinary claims is that they require *extraordinary evidence*.

Uncle George's AAPL trade ain't going to cut it. Here is the extraordinary evidence that I would personally want to see before agreeing that an investor can beat the market:

1. Big heaps of money

This is the one area of life where there really is no dodging that most venerable of sick burns: if you're so smart, why aren't you rich?

So the first piece of evidence I would accept is the fact that someone is very, very rich. Sitting atop a big old pile of cash. And they'd probably also open a hedge fund so they can take other people's money and turn it into millions more.

2. Track record of outperformance

Maybe a genuine market-beater doesn't have enough starting capital to make big piles of money on a relatively slim edge, and for some reason is unable to come up with any scheme to beg or borrow more? Or the anomaly is real, but disappears before they can get filthy rich?

In these scenarios, I would also accept a complete record of out-of-sample investment returns over time—no backtests! no selecting the best trades!—as compared against the appropriate risk-adjusted benchmark.

These evidential standards work both in the case that the EMH is 'dead', i.e. you can reliably beat the market using public domain info which ought to already be priced in, and in the case that it's not, i.e. you really do need novel information to get an edge.

As for evidence that the EMH really is dead...hmm. It's not a proper theory to begin with. But I guess it would be 'dead' when the predictions it generates stop being accurate or useful? Which would look something like: 'finding out that plenty of people meet the standards above, despite having never been in possession of a scrap of information that wasn't already available to the market'.

In doing so, we'd have to be very careful to make sure we aren't just looking at the Uncle Georges who unwittingly drew the winning lottery ticket. After all, we should expect plenty of investors to beat the market for a long time—even for years on end—entirely by chance.

The Great Coin-Tossing Experiment

Say we held a national coin-flipping contest. After 15 rounds, one in every ~32,800 people would have managed to call every single toss correctly, perfectly predicting a sequence like this:

H T H H T T H H H H T H T T T

Pretty impressive, huh!

Well, only in a world where we don't know about probability. In that world, we might mistake blind randomness for skill. The lucky few winners would be hailed as the heroes of their hometowns, do interviews with breathless breakfast TV hosts, and explain that it's all about the precise flick of the wrist. Aspiring flippers would queue up to buy the inevitable best-selling book, *Flip Me Off*, and pay exorbitant sums for one-on-one coaching sessions with the master tossers.

Depressingly, this is exactly what happens in the world of investing. What does it mean to achieve the kind of success which only happens by chance with 0.0003 probability? In the United States alone, it means you end up with 10,000 lucky dopes who are indistinguishable from brilliant investors.

And fund managers don't need to do anywhere near that well to attract a market-beating aura. They're incentivised to swing for the fences, increasing the odds they beat the market in some highly visible fashion over some shorter period—say, a lucky season or two. They inevitably regress to the mean, sometimes crashing and burning in spectacular fashion, but it doesn't matter so long as they manage to hose naive investors in the meantime.

We can never *entirely* rule out the effect of randomness—there will always be some tiny chance that Warren Buffett is really just the world's greatest coin-flipper—but we have to draw the line somewhere, or the standard is impossibly high.

Once the odds of a fluke get pretty slim—someone is super duper rich, *and* they've made a ton of consistently good trades over time—I'd happily congratulate them on their market-beating prowess, give them all my money to invest, listen eagerly to their advice, etc.

Being good Bayesians, this is obviously a spectrum: if they are not at that point, but trending in the right direction, I would be less skeptical, etc. But the bar has to be pretty high, or there's no way to separate skill from randomness.

Scott and Eliezer have both alluded to their comments being informed by private information. Here's a [reddit comment](#) in which Scott responds to criticism of his 'EMH is the real victim' riff:

I think we're in an asymmetric position, in that I know these people pretty well, I know they've thought about efficient market before, and they're the sort of people I would expect to beat the market if anyone could. I agree that if you just hear some blogger say he saw some people beat the market once, that's not much evidence.

Eliezer definitely understands the EMH, because the descriptions of it in *Inadequate Equilibria* are among the most brilliant and insightful I've ever come across. And Scott is obviously super smart.

I would love to know what their evidential standards are, but I'm explicitly *not* calling them out, or any of the people mentioned earlier in the post. No-one is under the slightest obligation to share private evidence, and I would be thrilled if those folks were indeed the market-beating chosen ones.

But I am saying that people, in general, make these kind of claims all the time—in good faith and with no malicious intent—and in general, taking their advice is an extraordinarily bad idea.

A heuristic: if you (or someone you know) is confident they can beat the market, and yet you notice you are not sitting atop an enormous pile of wealth, it's at least worth *considering* the possibility that you might be fooling yourselves.

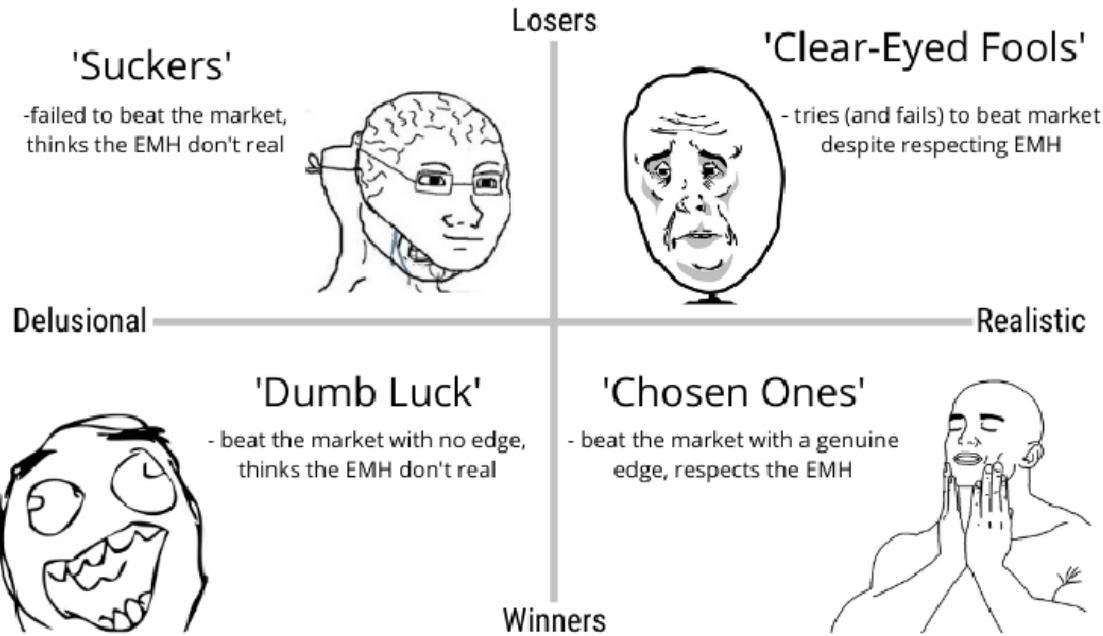
The Four Types of Investors

There are very obvious and well-known reasons why everyone loves to think they can beat the market: overconfidence, confirmation bias, 'resulting', selective memory, survivorship bias, etc.

These forces are so powerful that many people—myself included—blithely ignore the vast piles of evidence that suggest beating the market is incredibly difficult, and go ahead and try anyway. All of us think we are special, and (almost) all of us are wrong.

Even if we don't personally harbour this particular fantasy, there's also a natural tendency to want *our tribe* or *our friends* to be the brilliant visionaries who were ahead of the action, possess sweet market-beating skills, etc.

So we can roughly place investors into one of the following four quadrants:



('Losers' and 'winners' here is tongue-in-cheek, and not a value judgment: literally, losing/earning this game by either successfully beating the market, or failing to do so)

Deluded losers ('Suckers')

"Apple stock really is undervalued, but the market hasn't recognised it yet. I just got unlucky—it was because of [elaborate rationalization]. Also, even if I got it wrong this time, I was really close. Next time!"

"What's that? Do I track my portfolio returns over time, and compare against the relevant risk-adjusted benchmark to see whether I'm actually outperforming? Well, there's no need. I usually do pretty well for myself, and I'm expecting to improve—in fact, I just picked up this classic book called *The Intelligent Investor*..."

Deluded winners ('Dumb Luck')

"I knew Apple stock was undervalued! And I remember that other time I made a really good trade, too. Guess I'm pretty good at this game!"

"...What's that? I might have just got lucky? Hah, no. I even did the Fibonacci retracements and everything."

Realistic losers ('Clear-Eyed Fools')

"I keep a meticulous record of my portfolio returns, which forces me to acknowledge the fact that even though I occasionally do well, I am underperforming my benchmarks on a risk-adjusted basis. I am under no delusions about my prospects of finding an edge, and I know I really ought to take Warren Buffett's advice and put all my money in index funds."

"But I enjoy playing the markets! It's like how a night in Vegas can have negative EV but still be positive utility, because of all the non-financial factors. So I'm gonna keep gambling with a small part of my portfolio, just for shits and giggles. In the event that I 'win', I will try really hard to resist the incredible internal pressure to start thinking of myself as a brilliant investing guru."

Realistic winners ('Chosen One')

"I keep a meticulous record of my portfolio returns, which have outperformed the appropriate risk-adjusted benchmarks to such a degree that I am confident I have found a genuine informational asymmetry. I will of course never tell anyone about it, or it will become useless.

"And I can never be *entirely* sure: it's also possible that I just got lucky. But at the very least, I am sitting atop great piles of money, which is pretty nice."

The vast majority of people who actively trade their account are 'Suckers'. Some smaller number fall into the 'Dumb Luck' quadrant (Uncle George would stay there if he never places another trade, but he almost certainly won't be able to help himself.)

The right-hand quadrants are much more sparsely populated. I guess there are a few 'Clear-Eyed Losers' floating around, and a tiny handful of 'Chosen Ones'.

This rough distribution is probably not too controversial. The question is, which one are you?

(I made [a poll](#) on my website at this point in the post, just for a bit of fun: the results so far are brilliant)

Trying to Beat the Market is Like Crack for Smart People

There is a tendency for smart people to wander into areas they know very little about, and think they can do better than the actual experts who have years or decades of domain-specific knowledge, on the basis of being very smart, or having read some blog posts online about being more rational.

This would be OK if it was just a bit cringe. I love armchair pontificating as much as the next guy! The consequences are usually limited to mildly annoying the people who actually know what they're talking about, and much eye-rolling when you triumphantly reinvent the wheel.

There is some upside too: reinventing the wheel is fun, because you get to, like, *invent wheels*. And very occasionally, it might even be true! No doubt smart outsiders are occasionally able to breeze into a new field and exploit some obvious inefficiencies.

But...oh boy. It's really not true of this particular domain. And it's not harmless either.

The central prediction generated by the EMH is that you should not expect to be able to beat the market (in the non-trivial sense) unless you have unique information or some similar edge.

This prediction is tested every day. We have great piles of evidence which suggest that it is correct: the vast majority of active investors *do really badly*.

Crucially, it's not only regular schmucks who underperform. So do paid professionals, and active managers, and hedge funds, and all sorts of brilliant people who have made this their life's work.

It's possible that I am not making many friends with this post. I certainly feel pretty nervous about publishing it. Everyone who thinks they can beat the market will have their hackles up! If it helps at all, I am not claiming the high ground. I have made every dumb investing mistake you could think of, and then a few more besides. I am *painfully aware* of how hard this is.

These days I would put myself in the ‘Clear-Eyed Fool’ quadrant, but only by a fingernail. It’s a constant battle even to stay there. I *still* do clever things that contradict my own boring advice, and annoyingly, am rewarded for my hubris *just often enough* to start entertaining the thought that I’m a brilliant investing genius after all. Then I force myself to calculate the IRR on my publicly-traded investments, and compare it against appropriate benchmarks, and manage to get a fingernail-hold back on boring old reality.

To the extent that I have succeeded as an investor, and I am doing quite nicely thank you, it has only come through forcing myself to acknowledge the main prediction that emerges from the very-much-alive-and-kicking EMH. The huge and underappreciated benefit of doing so is that I occasionally divert some of my attention elsewhere, to domains where I actually *do* have an edge—and then I win.

Discussion

The EMH is a weird, counterintuitive, freaky-ass shoggoth of a thing and it still confuses the heck out of me, even after almost a decade of writing about finance. I almost certainly made some mistakes in the post—in the process of writing it, I noticed several ways in which my initial beliefs were subtly wrong, which has already been super useful for me, and helped me understand some of the (valid) objections people have raised.

So I would also like to use this post to open up the floor to any and all EMH discussion, and try to benefit from the smaller-but-still-powerful collective intelligence that is Less Wrong.

I'm going to add comments with some of my own questions and uncertainty. It would be nice to become less confused together, and try to get a better sense of where we should apply our efforts at the margin.

An overview of 11 proposals for building safe advanced AI

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

This is the blog post version of [the paper by the same name](#). Special thanks to Kate Woolverton, Paul Christiano, Rohin Shah, Alex Turner, William Saunders, Beth Barnes, Abram Demski, Scott Garrabrant, Sam Eisenstat, and Tsvi Benson-Tilsen for providing helpful comments and feedback on this post and the talk that preceded it.

This post is a collection of 11 different proposals for building safe advanced AI under the current machine learning paradigm. There's a lot of literature out there laying out various different approaches such as [amplification](#), [debate](#), or [recursive reward modeling](#), but a lot of that literature focuses primarily on outer alignment at the expense of [inner alignment](#) and doesn't provide direct comparisons between approaches. The goal of this post is to help solve that problem by providing a single collection of 11 different proposals for building safe advanced AI—each including both inner and outer alignment components. That being said, not only does this post not cover all existing proposals, I strongly expect that there will be lots of additional new proposals to come in the future. Nevertheless, I think it is quite useful to at least take a broad look at what we have now and compare and contrast some of the current leading candidates.

It is important for me to note before I begin that the way I describe the 11 approaches presented here is not meant to be an accurate representation of how anyone else would represent them. Rather, you should treat all the approaches I describe here as *my version* of that approach rather than any sort of canonical version that their various creators/proponents would endorse.

Furthermore, this post only includes approaches that intend to directly build advanced AI systems via machine learning. Thus, this post doesn't include other possible approaches for solving the broader AI existential risk problem such as:

- finding a fundamentally different way of approaching AI than the current machine learning paradigm that makes it easier to build safe advanced AI,
- developing some advanced technology that produces a [decisive strategic advantage](#) without using advanced AI, or
- achieving global coordination around not building advanced AI via (for example) a persuasive demonstration that any advanced AI is likely to be unsafe.

For each of the proposals that I consider, I will try to evaluate them on the following four basic components that I think any story for how to build safe advanced AI under the current machine learning paradigm needs.

1. **Outer alignment.** Outer alignment is about asking why the objective we're training for is aligned—that is, if we actually got a model that was trying to optimize for the given loss/reward/etc., would we like that model? For a more thorough description of what I mean by outer alignment, see "[Outer alignment and imitative amplification](#)."
2. **Inner alignment.** Inner alignment is about asking the question of how our training procedure can actually guarantee that the model it produces will, in fact,

be trying to accomplish the objective we trained it on. For a more rigorous treatment of this question and an explanation of why it might be a concern, see "[Risks from Learned Optimization](#)".

3. **Training competitiveness.** Competitiveness is a bit of a murky concept, so I want to break it up into two pieces here. Training competitiveness is the question of whether the given training procedure is one that a team or group of teams with a reasonable lead would be able to afford to implement without completely throwing away that lead. Thus, training competitiveness is about whether the proposed process of producing advanced AI is competitive.
4. **Performance competitiveness.** Performance competitiveness, on the other hand, is about whether the final product produced by the proposed process is competitive. Performance competitiveness is thus about asking whether a particular proposal, if successful, would satisfy the use cases for advanced AI—e.g. whether it would fill the economic niches that people want AGI to fill.

I think it's often easy to focus too much on either the alignment side or the competitiveness side while neglecting the other. We obviously want to avoid proposals which could be unsafe, but at the same time the "do nothing" proposal is equally unacceptable—while doing nothing is quite safe in terms of having no chance of directly leading to existential risk, it doesn't actually help in any way relative to what would have happened by default. Thus, we want proposals that are both aligned and competitive so that they not only don't lead to existential risk themselves, but also help existential risk in general by providing a model of how safe advanced AI can be built, being powerful enough to assist with future alignment research, and/or granting a [decisive strategic advantage](#) that can be leveraged into otherwise reducing existential risk.

The 11 proposals considered in this post are, in order:^[1]

1. Reinforcement learning + transparency tools
2. Imitative amplification + intermittent oversight
3. Imitative amplification + relaxed adversarial training
4. Approval-based amplification + relaxed adversarial training
5. Microscope AI
6. STEM AI
7. Narrow reward modeling + transparency tools
8. Recursive reward modeling + relaxed adversarial training
9. AI safety via debate with transparency tools
10. Amplification with auxiliary RL objective + relaxed adversarial training
11. Amplification alongside RL + relaxed adversarial training

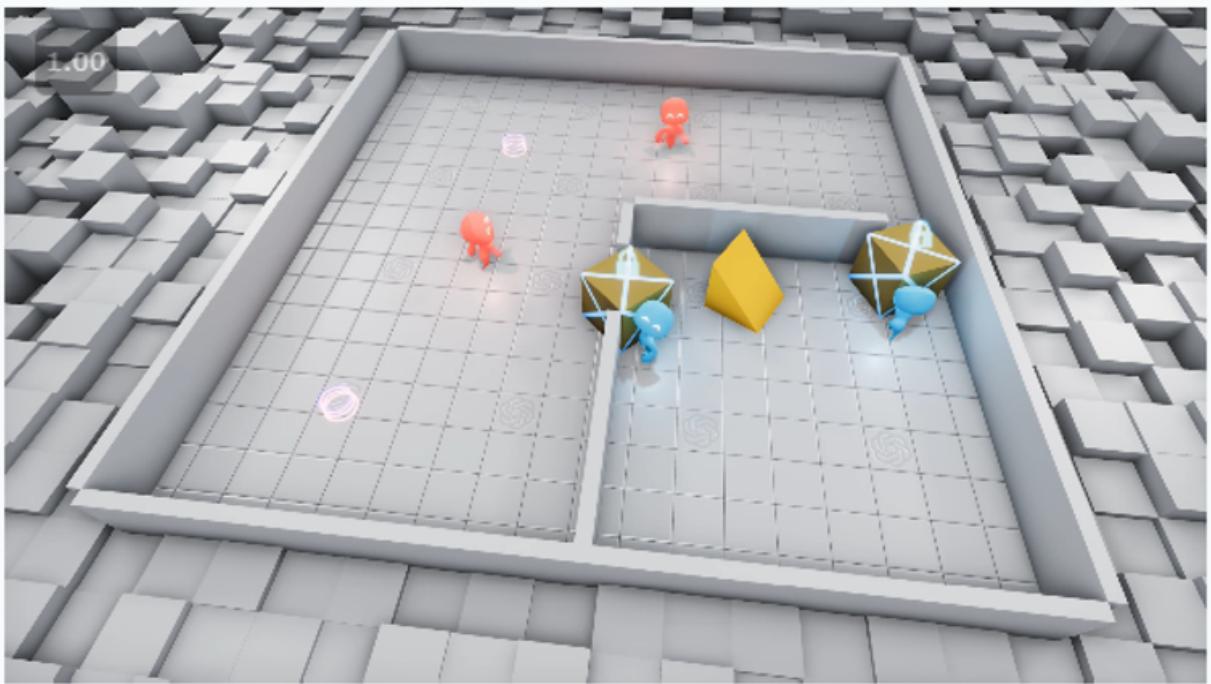
EDIT: For some other proposals that didn't make it onto this list because they came out after this post, see "[Imitative Generalization](#)" and "[AI safety via market making](#)."

1. Reinforcement learning + transparency tools

Here's our first approach:

1. Train a reinforcement learning (RL) agent in an environment where corrigibility, honesty, multi-agent cooperation, etc. are incentivized. The basic idea would be to mimic the evolutionary forces that led to humans' general cooperativeness. As

an example of work along these lines that exists now, see [OpenAI's hide and seek game](#). Furthermore, the environment could also be modified to directly reward following human instructions, encouraging corrigibility towards humans. For a more thorough discussion of this possibility, see Richard Ngo's "[Multi-agent safety](#)."



Additionally, hiders learn to **coordinate** who will block which door and who will go grab the ramp. In cases where the boxes are far from the doors, hiders **pass boxes to each other** in order to block the doors in time.

An image of [OpenAI's hide and seek game](#).

2. Have humans use [transparency tools](#), adversarial training, etc. to check for [deceptive](#) or otherwise [catastrophic](#) behavior in the resulting model.

That's the approach—now is it aligned? Competitive?

Outer alignment. Outer alignment here is entirely dependent on whatever the dominant behavior is in the training environment—that is, what is the deployment behavior of those models which perform optimally in the training environment. If corrigibility, honesty, cooperation, etc. do in fact dominate in the limit, then such an approach would be outer aligned. By default, however, it seems quite difficult to understand the limiting behavior of complex, multi-agent environments, especially if they're anywhere as complex as the actual human ancestral environment. If following human instructions is incentivized, for example, that could lead to corrigibility in the limit—or it could lead to agents which only choose to follow human instructions for the instrumental reason of believing it will help them acquire more resources. Alternatively, it might be possible to isolate the structure that was present in the human ancestral

environment that led us to be cooperative, honest, etc. One worry here, however, is that even if we could figure out how to properly incentivize cooperation, it might result in agents that are cooperative with each other but not very cooperative with us, similarly to how we might not be very cooperative with aliens that are very different from us.

Inner alignment. The idea of inner alignment in this situation is to ensure that training produces something in line with the optimal behavior in the environment (the alignment of the optimal behavior being an outer alignment question) rather than other, potentially perverse equilibria. The basic proposal for how to avoid such perverse equilibria with this proposal is via the use of checks such as transparency tools and adversarial training to detect inner alignment failures before the model is deployed. [Chris Olah describes](#) this sort of transparency checking as giving you a “mulligan” that lets you throw away your model and start over if you find something wrong. Thus, ideally, if this approach ends up not working it should be clear before the model is deployed, enabling either this approach to be fixed or a new approach to be found instead. And there is a reasonable chance that it does just work—we don’t understand our models’ inductive biases very well, but it seems entirely possible that they could work out such that [pseudo-alignment](#) is disincentivized.

In my opinion, while it seems quite plausible to me that this sort of approach could catch [proxy_pseudo-alignment](#), it seems unlikely that it would successfully catch [deceptive_pseudo-alignment](#), as it could be very difficult to make transparency tools that are robust to a deceptive model actively trying to trick them. To catch deceptive alignment, it seems likely to be necessary to incorporate such checks into the training process itself—which is possible to do in this setting, though is not the approach I described above—in order to prevent deception from occurring in the first place rather than trying to detect it after the fact.

Training competitiveness. Training competitiveness here seems likely to depend on the extent to which the sort of agency produced by RL is necessary to train advanced AI systems. Performing RL in highly complex, difficult-to-simulate environments—especially if those environments involve interaction with the real world—could be quite expensive from a training competitiveness standpoint. Compared to simple language modeling, for example, the difficulty of on-policy data collection combined with low sample-efficiency could make full-scale RL much less training competitive. These sorts of competitiveness concerns could be particularly pronounced if the features necessary to ensure that the RL environment is aligned result in making it significantly more difficult to simulate. That being said, if RL is necessary to do anything powerful and simple language modeling is insufficient, then whether or not language modeling is easier is a moot point. Whether RL is really necessary seems likely to depend on the extent to which it is necessary to explicitly train agents—which is very much an open question. Furthermore, even if agency is required, it could potentially be obtained just by imitating an actor such as a human that already has it rather than training it directly via RL.

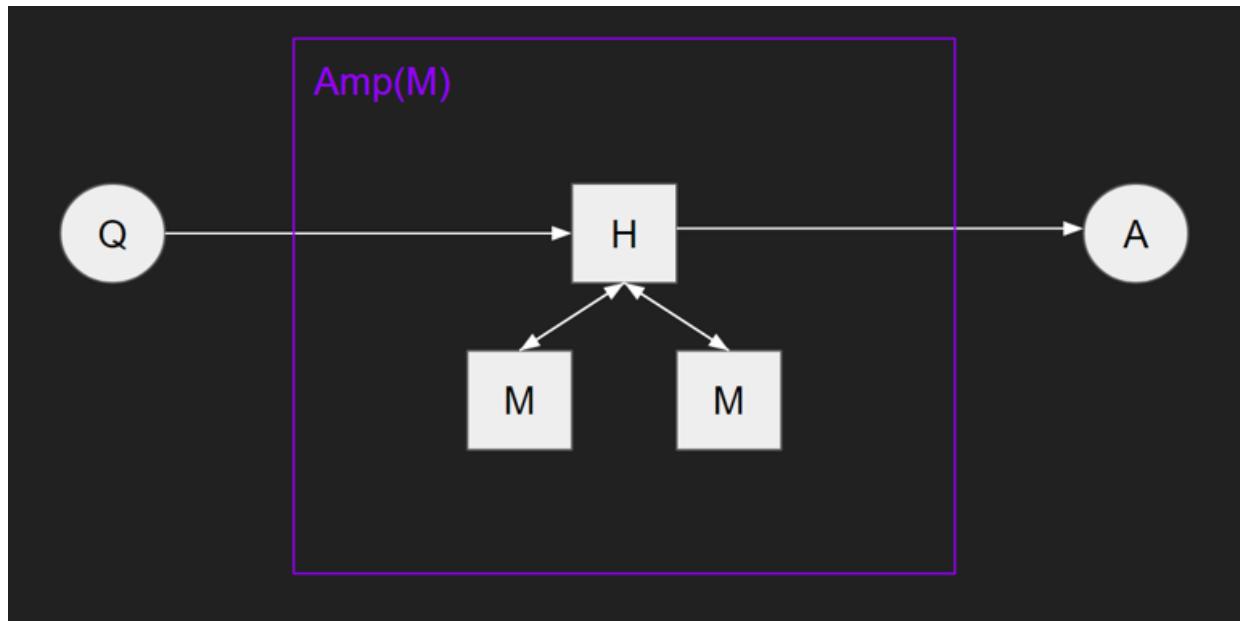
Performance competitiveness. The question for performance competitiveness here is to what extent it is possible to create an environment that incentivizes all the behavior you might want from your AGI. Such an environment doesn’t need to be purely simulated—you could do some simulation training and some real-world training, for example. Regardless of how your RL environment is constructed, however, it needs to actually incentivize the correct behavior for the tasks that you want to use your AI for. For example: can you incentivize good decision-making? Good question-answering? Good learning ability? Do you need good fine motor control, and if so, can you incentivize it? These are highly non-trivial questions: it could be quite difficult to set up

an RL environment to teach an agent to do all of the tasks you might want it to perform to fill all the economic niches for AGI, for example. Of course, this is going to be highly dependent on what exactly those economic niches are that you want your advanced AI to fill.

2. Imitative amplification + intermittent oversight

Though many of the approaches on this list make use of the basic [iterated amplification](#) framework, imitative amplification is probably the most straightforward—though it still has a good deal of moving parts.

To define imitative amplification, we'll first define $\text{Amp}(M)$ —the “amplification operator”—as the procedure where a human H answers a question with access to a model M .^[2]

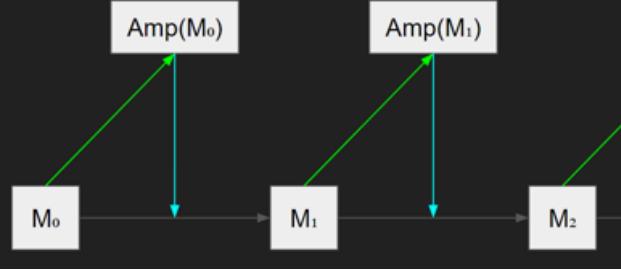


A diagram of the amplification operator $\text{Amp}(M)$ where white arrows indicate information transfer, Q is a question, A is $\text{Amp}(M)$'s answer, H is a human, and M is the model.

Then, imitative amplification is just the procedure of iteratively training M to imitate $\text{Amp}(M)$.

The approach:

1. Train M to imitate $\text{Amp}(M)$ (the human with access to the model).



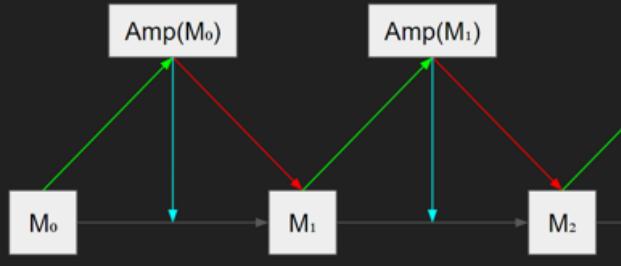
The basic imitative amplification setup where green arrows indicate amplification, gray arrows indicate training, and cyan arrows indicate the imitative amplification loss.

Finally, we can define imitative amplification + intermittent oversight—which is the full approach we want to consider here—as the combination of imitative amplification with intermittent oversight of M by $\text{Amp}(M)$ whenever the target model changes.

Specifically, we want $\text{Amp}(M)$ to look for [deceptive](#) or otherwise [catastrophic](#) behavior in M by utilizing things like transparency tools and adversarial attacks.

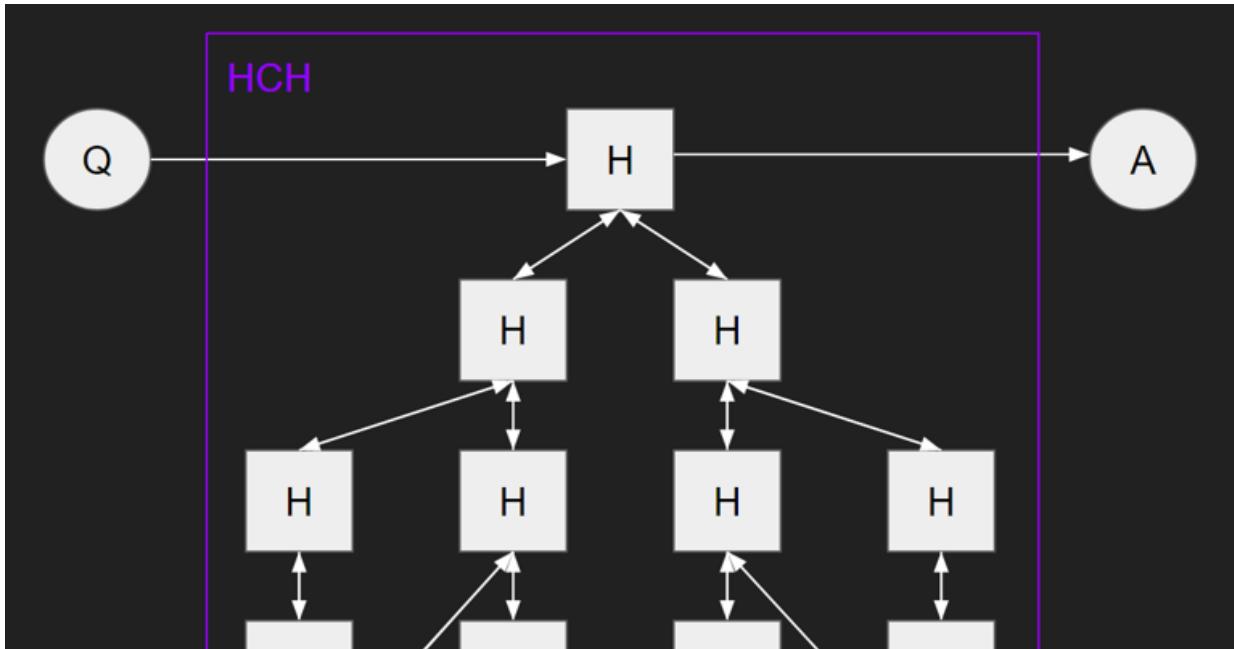
The approach:

1. Train M to imitate $\text{Amp}(M)$ (the human with access to the model).
2. Throughout training have $\text{Amp}(M)$ inspect the new model for bad behavior using transparency tools.



Imitative amplification plus intermittent oversight where red arrows indicate oversight (from the overseer to the overseen model).

Outer alignment. Since imitative amplification trains M to imitate $\text{Amp}(M)$, it limits^[3] to the fixed point of the Amp operator, which Paul Christiano calls [HCH](#) for “Humans Consulting HCH.” HCH is effectively a massive tree of humans consulting each other to answer questions.



A (partial) diagram of HCH where white arrows indicate information transfer.

Thus, whether imitative amplification is outer aligned is dependent on whether HCH is aligned or not. HCH's alignment, in turn, is likely to depend heavily on the specific humans used and what sort of policy they're implementing. The basic idea, however, is that since the limit is composed entirely of humans—and since we can control what those humans do—we should be able to ensure that all the optimization power is coming from the humans (and not from memetic selection, for example), which hopefully should allow us to make it safe. While there are certainly valid concerns with the humans in the tree accidentally (or purposefully) implementing some misaligned policy, there are [possible things you can do to address these problems](#).

Inner alignment. The hope here is to be able to detect and fix serious inner alignment failures via the use of intermittent oversight—that is, inspection by a human with access to a previous version of the model. Thus, the question of inner alignment here is dependent on the question of whether such an overseer would be able to detect deceptive or otherwise catastrophic behavior in the model. Ideally, if such detection is working, it should be clear as soon as the model starts to become misaligned, giving us the opportunity to either fix the issue or find a new approach. For that to work, however, it will likely need to be the case that the intermittent oversight is performed regularly enough to catch any deception as soon as it starts appearing since once a model becomes deceptive it might be significantly harder to detect problems if it starts trying to obfuscate its deception.

That being said, while detecting deceptive cognition early enough directly might be quite difficult, it might be possible to instead detect deceptive behavior indirectly by verifying a condition that rules out deception while still being relatively easy to verify such as myopia. In this context, a myopic model would be one that determines its output based on some fixed standard such as distance to HCH while a non-myopic model would be one that determines its output based on its consequences.^[4] For a more detailed explanation of how such myopia verification might work, see “[Towards a mechanistic understanding of corrigibility](#).”^[5]

Additionally, because we do intermittent oversight whenever we change the target network, we have a guarantee that the overseer is always at least as strong as any targets that the model was trained to imitate. That being said, the model could end up stronger than the targets it was trained to imitate if it manages to, for example, achieve significant compression of what the targets are doing—which in fact is necessary for training competitiveness. Furthermore, even if the overseer is stronger than the model, the offense-defense balance could be such that a deceptive model would still be able to hide its deception. Which direction the offense-defense balance lies in is likely to depend heavily on what sorts of transparency tools the overseer has access to.

Training competitiveness. First, the training competitiveness of imitative amplification is likely to depend on whether pure imitation can be turned into a rich enough reward signal to facilitate highly sample-efficient learning. In my opinion, it seems likely that human language imitation (where language includes embedded images, videos, etc.) combined with techniques to improve sample efficiency^[6] will be competitive at some tasks—namely highly-cognitive tasks such as general-purpose decision-making—but not at others, such as fine motor control. If that’s true, then as long as the primary economic use cases for AGI fall into the highly-cognitive category, imitative amplification should be training competitive. For a more detailed analysis of this question, see “[Outer alignment and imitative amplification](#).”

Second, imitative amplification + intermittent oversight could also fail to be training competitive if it is very difficult for the intermittent oversight check to ever pass. This could happen if, for example, the training process for imitative amplification doesn’t naturally produce models that are transparent enough to facilitate such a check. That being said, what counts as “transparent enough” again depends heavily on the specifics of the transparency tools that the overseer has access to.

Performance competitiveness. The performance competitiveness question here, as with the outer alignment question, is heavily dependent on the properties of HCH, as imitative amplification is specifically designed to limit toward HCH. In particular, performance competitiveness is dependent on whether HCH is sufficiently universal so as to be capable of solving whatever hard problems we might want advanced AI to solve. For example, if trees of humans can’t replicate intuitive flashes of insight, and such flashes of insight are necessary to solve the sorts of highly-cognitive tasks we want AGI for, then imitative amplification would fail to be performance competitive. Alternatively, as long as it is always possible for a human to break down complex tasks that they can’t answer themselves into easier subtasks, then imitative amplification should be performance competitive. Note that it doesn’t have to be the case that this work for an arbitrary human without any training—only for the sort of human from whom you would actually be collecting training data.

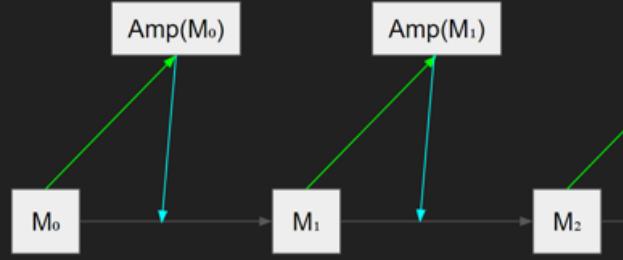
3. Imitative amplification + relaxed adversarial training

The basic idea here is very similar to the previous approach in that we start out with pure imitative amplification.

The approach:

Train a model on the joint objective
of both:

- a. imitating $\text{Amp}(M)$



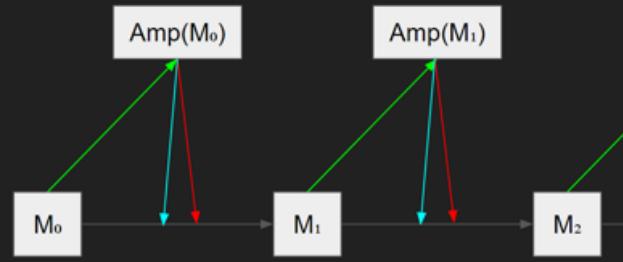
The basic imitative amplification setup.

However, instead of doing intermittent oversight checks, we instead train directly on the intermittent oversight signal—that is, we jointly train M on the imitation loss and on decreasing the probability of catastrophe as assessed by $\text{Amp}(M)$ —yielding a [relaxed adversarial training setup](#).

The approach:

Train a model on the joint objective
of both:

- a. imitating $\text{Amp}(M)$ and
- b. decreasing the probability of catastrophe as assessed by $\text{Amp}(M)$ trying to find adversarial attacks.



Imitative amplification plus relaxed adversarial training where the red arrows indicate the oversight loss.

Outer alignment. Imitative amplification + relaxed adversarial training should effectively have the same outer alignment properties as imitative amplification + intermittent oversight. If we assume that it is possible to imitate HCH in a demonstrably non-catastrophic way,^[7] then the catastrophe-checking imposed by either relaxed adversarial training or intermittent oversight shouldn't affect the theoretical limiting behavior. Such a limit assumes an infinite training distribution, however, which in practice is, of course, impossible, leaving open the possibility for the two different approaches to have quite different behavior in the finite data regime, giving them potentially quite different inner alignment properties.

Inner alignment. Compared to intermittent oversight, relaxed adversarial training could have very different inner alignment properties. Intermittent oversight requires performing oversight after a period of training during which the model could potentially have become deceptive. Thus, intermittent oversight could potentially require catching deception after the fact, giving the deceptive model the opportunity to try to obfuscate its own cognition to trick the overseer.^[8] Relaxed adversarial training, by contrast, has

the potential to avoid deception before it ever occurs by providing a consistent gradient away from deception during the entirety of training, thus avoiding the issue of having to detect deception after the fact. Similarly to intermittent oversight, such a gradient could be achieved by having the overseer enforce a condition such as [myopia](#) which is simple enough to reliably enforce while guaranteeing that all models which meet the condition are non-deceptive. However, relaxed adversarial training does come with the possible risk of overfitting the oversight signal, potentially resulting in obfuscated models that pass oversight despite not actually meeting the condition simply by virtue of having been optimized to do so.

Training competitiveness. Similarly to imitative amplification + intermittent oversight, imitative amplification + relaxed adversarial training could fail to be training competitive either because imitative amplification is uncompetitive or because the oversight check is too hard to pass. On the one hand, relaxed adversarial training could require significantly more total oversight bandwidth, which could make training more difficult. On the other hand, unlike with intermittent oversight, relaxed adversarial training directly trains the model to pass the check, which could significantly boost training competitiveness by increasing the likelihood of the check passing.

For relaxed adversarial training to be more training competitive than intermittent oversight, it needs to be the case that it significantly increases the probability of oversight working, which is likely to depend on the overfitting question of whether relaxed adversarial training is likely to overall increase or decrease model transparency. The critical question here is going to be how hard it is to find a model that is demonstrably non-catastrophic under the two different oversight regimes. For a more detailed analysis of this question, see "[Relaxed adversarial training for inner alignment](#)."

Performance competitiveness. Similarly to the outer alignment question, the performance competitiveness properties of imitative amplification + intermittent oversight and imitative amplification + relaxed adversarial training should be effectively the same, with performance competitiveness here highly dependent on the competitiveness of HCH.

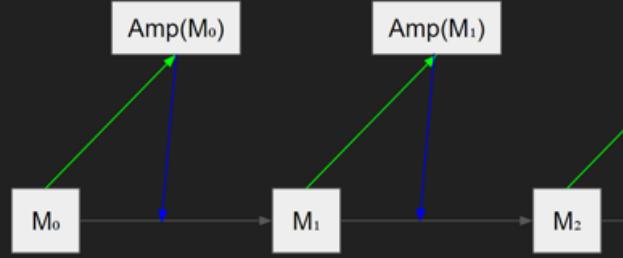
4. Approval-based amplification + relaxed adversarial training

The idea behind approval-based amplification is to substitute the imitation signal in imitative amplification with an approval signal, training M to maximize the approval of $\text{Amp}(M)$. This gives us a myopic reward signal which can then be optimized via the use of [some sort](#) of myopic reinforcement learning.

The approach:

Train a model on the joint objective
of both:

- a. maximizing the approval of Amp(M)



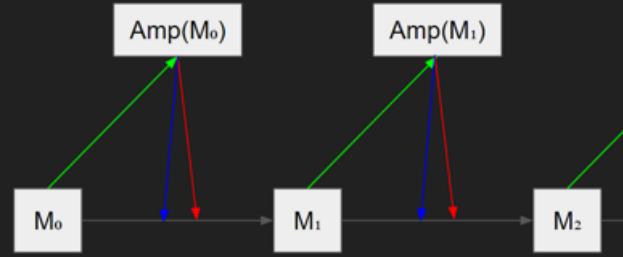
Approval-based amplification where the blue arrows indicate the approval loss.

We can then add on top of that a relaxed adversarial training signal, training M to also minimize the probability of catastrophe as assessed by Amp(M).

The approach:

Train a model on the joint objective
of both:

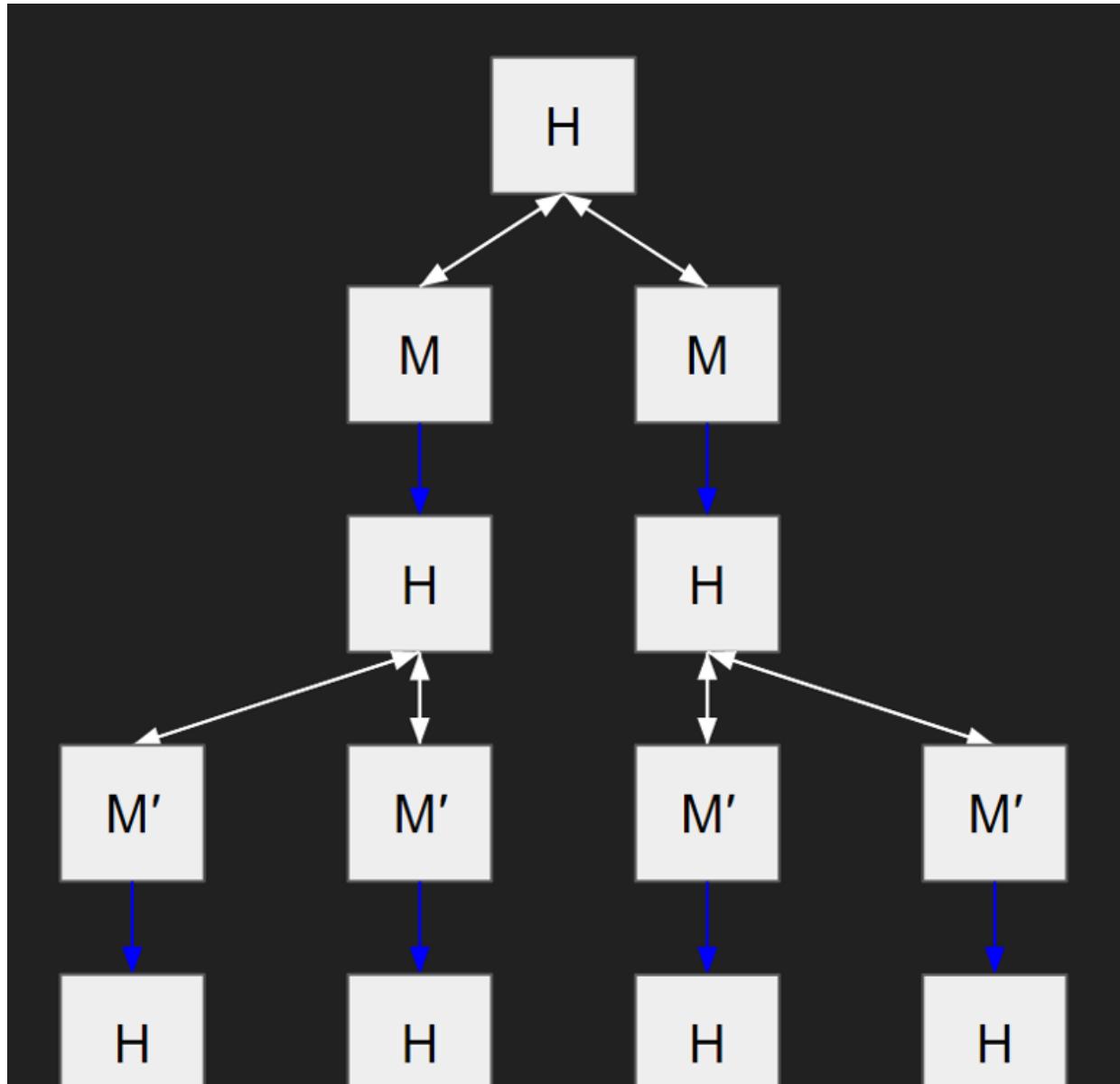
- a. maximizing the approval of Amp(M) and
- b. decreasing the probability of
catastrophe as assessed by Amp(M)
trying to find adversarial attacks.



Approval-based amplification + relaxed adversarial training where the red arrows indicate the oversight loss.

Additionally, it is worth noting that, in practice, since oversight can be treated as a form of approval, you can just compute one approval signal that includes both metrics rather than two distinct approval signals.

Outer alignment. For the purposes of outer alignment and performance competitiveness, we have to understand what exactly the theoretical limit of approval-based amplification is, which in general is not HCH. Instead, approval-based amplification limits to the following tree, where blue arrows indicate approval-maximization (such that M is maximizing the approval of H consulting M' which is maximizing the approval of H consulting M'' and so on).^[9]



The top of the approval-based amplification limiting tree where white arrows indicate information transfer, blue arrows indicate approval maximization, H are humans, and M are models.

As you can see from the tree, the limit of approval-based amplification involves humans interspersed with models attempting to maximize the approval of their descendants. Thus, unlike the HCH tree, which was composed only of humans, analyzing the approval-based amplification tree involves analyzing the behavior of arbitrary models trained on approval-maximization signals—and, in fact, many such models as they appear all throughout the tree each with a different subtree that they are maximizing the approval of.^[10] As a result, producing an outer alignment guarantee regarding such a tree seems quite difficult, especially given the distinct possibility of the models gaming the approval signal. For additional detail on this question, see “[Outer alignment and imitative amplification](#).”

Inner alignment. We are still using relaxed adversarial training here, so for inner alignment to work we need the same guarantee from before of a consistent gradient away from deception. Note that we can still try to have Amp(M) enforce [myopia](#) as a way of getting that gradient, as approval-maximization is a single-step reward signal that is entirely compatible with a myopic agent.^[11] In the approval-based setting, however, Amp(M) is a different entity that limits to a different tree, meaning it could have different properties regarding how good its oversight signal is. In particular, the fact that approval-based amplification directly incentivizes M to fool H into approving of it could result in M being more prone to deception, more likely to hide facts from H during oversight, and/or more likely to obfuscate its internals if internals-checking is part of the oversight signal. That being said, it could also be that relaxed adversarial training works better in an approval-based setting if it results in a more powerful Amp(M) that is better at performing oversight because M was trained to provide more helpful information.

Training competitiveness. Training competitiveness could be quite different with approval-based amplification compared to imitative amplification, as approval provides a very different training signal than imitation. In particular, the standard maxim that it is generally easier to evaluate behavior than produce it seems to suggest that human approval should produce a significantly richer training signal than the simple automated distance metric in imitative amplification, resulting in more efficient training. This isn't necessarily the case, however—an automated distance metric as computed by a powerful language model, for instance, could still produce a fairly rich (and more nicely differentiable) training signal.

Additionally, approval-based amplification could also have different training competitiveness properties in terms of the likelihood of the relaxed adversarial training check passing. If approval-based amplification leads to models with more obfuscated internals, for example—perhaps because the model is incentivized to obfuscate how it's gaming the approval signal—then that could hurt training competitiveness, as it could make it harder for Amp(M) to become confident that M is safe (in addition to hurting inner alignment by making M in fact less likely to be safe).

Performance competitiveness. Performance competitiveness for approval-based amplification depends on the properties of the approval-maximization tree given previously. Compared to HCH, it is quite possible for that tree to be either more or less competitive.

On the one hand, approval-maximizing models might be able to produce ideas/considerations/arguments that no human would have ever come up with, enabling the approval-maximization tree to be more competitive. In particular, if humans are significantly better at knowing good ideas when they see them than producing good ideas themselves—which fits with the standard maxim that it's easier to evaluate behavior than produce it—then approval-based amplification should be more performance competitive than imitative amplification.

On the other hand, approval-maximizing models might game the approval signal by producing convincing-sounding but ultimately bad ideas/considerations/arguments that

end up hurting performance competitiveness. For example, an idea could be initially very compelling and thus get high approval despite quite quickly revealing itself to be useless/vacuous/etc. when actually tested out. In such a situation where the bad ideas quickly reveal themselves, the performance competitiveness problems would likely be quite clear, though if it takes a very long time for the bad ideas to reveal themselves, then approval-maximization might initially look a lot better from a performance competitiveness standpoint than it actually is.

5. Microscope AI

[Microscope AI](#) is a fairly unique proposal which is designed to bypass some of the dangers of building highly agentic AGI systems by leveraging powerful transparency tools. The basic proposal is as follows.

1. Train a predictive model on some set of data that you want to understand while using transparency tools to verify that the model isn't performing any optimization.
2. Use transparency tools to understand what the model learned about the data and use that understanding to guide human decision-making.

Microscope AI is predicated on the basic insight that using transparency tools on a model doesn't just teach us about that model—it also gives us useful information about the data that the model was trained on. Chris Olah talks about this basic phenomenon in his post "[Visualizing Representations: Deep Learning and Human Beings](#):

The visualizations are a bit like looking through a telescope. Just like a telescope transforms the sky into something we can see, the neural network transforms the data into a more accessible form. One learns about the telescope by observing how it magnifies the night sky, but the really remarkable thing is what one learns about the stars. Similarly, visualizing representations teaches us about neural networks, but it teaches us just as much, perhaps more, about the data itself.

The basic goal of microscope AI is to produce high-quality knowledge that can inform important decision-making rather than to produce powerful AGI systems that can make those decisions themselves.

Outer alignment. [Paul Christiano's argument that the universal prior is malign](#) suggests that there might be deceptive optimization in the limit of perfect prediction. As a result, if we interpret outer alignment as referring to [the behavior at minimal loss](#), it seems likely that microscope AI is not outer aligned. That doesn't necessarily rule out microscope AI as a proposal, however—whether such deceptive optimization would actually occur in practice is still a very open question. In particular, the argument that the universal prior is malign relies on an uncomputable limit—polynomial approximations, for example, could still be completely fine. Furthermore, the inner alignment technique proposed here of using transparency tools to detect and prevent any optimization could prevent this problem entirely if successful.

Inner alignment. Since the goal here is not to produce a model performing optimization (a mesa-optimizer) in the first place, the question of inner alignment becomes rather different in this situation. We still want to avoid a deceptive or otherwise misaligned mesa-optimizer, but we want to do so by avoiding all mesa-optimization rather than by ensuring that any mesa-optimization is aligned. Specifically, the strategy here is to use transparency tools to detect any optimization

(as opposed to just predictive modeling) and halt if we see anything of that form. This has the major benefit of sidestepping a lot of the most pernicious inner alignment concerns, but it also comes with lots of its own problems. Whether this will work is likely to be dependent on a couple of different factors.

First, it is an open question of whether there exists a sensible way in which optimization and modeling can be distinguished from each other. It seems quite likely that any world model which admits queries of the form “what is the probability that a is the optimal action according to utility u” or “what is the best action for agent A to take in state x” will implicitly have to be running some sort of optimization algorithm. Thus, any notion of a “pure world model” would likely have to either rule out such queries, ensure they are being estimated without the use of optimization, or otherwise find some way to ensure that they are safe. For a more thorough treatment of the difficulty here, see Abram Demski’s [“Partial Agency” sequence](#).

Second, assuming it is possible to distinguish between modeling and optimization in theory, it also needs to be possible in practice given whatever set of transparency tools are available. If the model can start performing optimization and potentially become deceptive before its optimization can be detected, that would obviously pose a major problem to this sort of approach.

Training competitiveness. Microscope AI relies on being able to competitively train predictive models with useful internal representations. Thus, the question of training competitiveness for microscope AI comes down to the question of whether the process of training a predictor is a competitive way to extract useful information from a dataset. Given the success of current predictive models such as [GPT-2](#), it seems likely that it will be competitive to train powerful predictors—though it remains to be seen how competitive the process of extracting what they’ve learned will be. How competitive that extraction process is seems likely to depend heavily on what sort of state-of-the-art transparency tools are available.

Performance competitiveness. Performance competitiveness is perhaps the biggest question mark regarding microscope AI, as microscope AI forgoes producing AI agents which directly take actions in the world. The question of performance competitiveness for microscope AI is thus the question of whether enhanced human understanding alone—without AI agents—is sufficient for the economic use cases where one might otherwise want highly agentic advanced AI (e.g. an AGI).

This question is likely to depend heavily on what exactly those use cases are. Like with amplification, if you need lots of fine motor control, microscope AI is unlikely to get you there. Furthermore, unlike amplification, if you need lots of low-level decision-making where it’s too expensive to hire a human, microscope AI won’t help much there either (whereas amplification would be fine). Potentially microscope AI could give humans the knowledge to safely build other systems which could solve such tasks, however. Furthermore, if the primary use case for AGI is just high-level big-picture decision-making (automating CEOs or doing AI research, for example), then it seems likely that microscope AI would have a real shot of being able to address those use cases. In that sort of a situation—where you’re only trying to make a small number of high-quality decisions—it seems likely to be fairly cheap to have a human in the loop and thus simply improving that human’s knowledge and understanding via microscope AI might be sufficient to produce competitive decision-making. This is especially true if there is a market premium on having a human making the decisions, perhaps because that makes it easier to negotiate or work with other humans.

6. STEM AI

STEM AI is a very simple proposal in a similar vein to microscope AI. Whereas the goal of microscope AI was to avoid the potential problems inherent in building agents, the goal of STEM AI is to avoid the potential problems inherent in modeling humans. Specifically, the idea of STEM AI is to train a model purely on abstract science, engineering, and/or mathematics problems while using transparency tools to ensure that the model isn't thinking about anything outside its sandbox.

This approach has the potential to produce a powerful AI system—in terms of its ability to solve STEM problems—without relying on any human modeling. Not modeling humans could then have major benefits such as ensuring that the resulting model doesn't have the ability to trick us to nearly the same extent as if it possessed complex models of human behavior. For a more thorough treatment of why avoiding human modeling could be quite valuable, see Ramana Kumar and Scott Garrabrant's "[Thoughts on Human Models](#)."

Outer alignment. Similarly to microscope AI, it seems likely that—in the limit—the best STEM AIs would be malign in terms of having convergent instrumental goals which cause them to be at odds with humans. Thus, STEM AI is likely not outer aligned—however, if the inner alignment techniques being used are successful at preventing such malign optimization from occurring in practice (which the absence of human modeling could make significantly easier), then STEM AI might still be aligned overall.

Inner alignment. The hope with STEM AI is that by preventing the model from ever considering anything outside its STEM sandbox, the malign limiting behavior that might cause it to fail to be outer aligned can be avoided. Unfortunately, such a sandboxing condition alone isn't quite sufficient, as a model considering only things in its sandbox could still end up creating other models which would consider things outside the sandbox.^[12] Thus, exactly what the correct thing is to do in terms of inner alignment for a STEM AI is somewhat unclear. In my opinion, there are basically two options here: either do something similar to microscope AI and try to prevent all mesa-optimization or do something similar to amplification and ensure that all mesa-optimization that occurs is fully [myopic](#). In either case, the hope would be that the absence of human modeling makes it easier to enforce the desired condition (because modeling an agent such as a human increases the propensity for the model to become agentic itself, for example).

Training competitiveness. Training competitiveness for STEM AI is likely to depend heavily on how hard it is for state-of-the-art machine learning algorithms to solve STEM problems compared to other domains such as language or robotics. Though there exists lots of current progress in the field of applying current machine learning techniques to STEM problems such as [theorem proving](#) or [protein folding](#), it remains to be seen to what extent the competitiveness of these techniques will scale, particularly in terms of how well they will scale in terms of solving difficult problems relative to other domains such as language modeling.

Performance competitiveness. Similarly to microscope AI, performance competitiveness is perhaps one of the biggest sticking points with regards to STEM AI, as being confined solely to STEM problems has the major potential to massively limit the applicability of an advanced AI system. That being said, many purely STEM problems such as [protein folding](#) or [nanotechnology development](#) have the potential to provide huge economic boons that could easily surpass those from any other form of

advanced AI as well as solve major societal problems such as curing major illnesses. Thus, if the answer to the reason that you want to build advanced AI in the first place is to get such benefits, then STEM AI might be a perfectly acceptable substitute from a performance competitiveness standpoint. Furthermore, such boons could lead to a [decisive strategic advantage](#) that could enable heavy investment in aligning other forms of advanced AI which are more performance competitive.

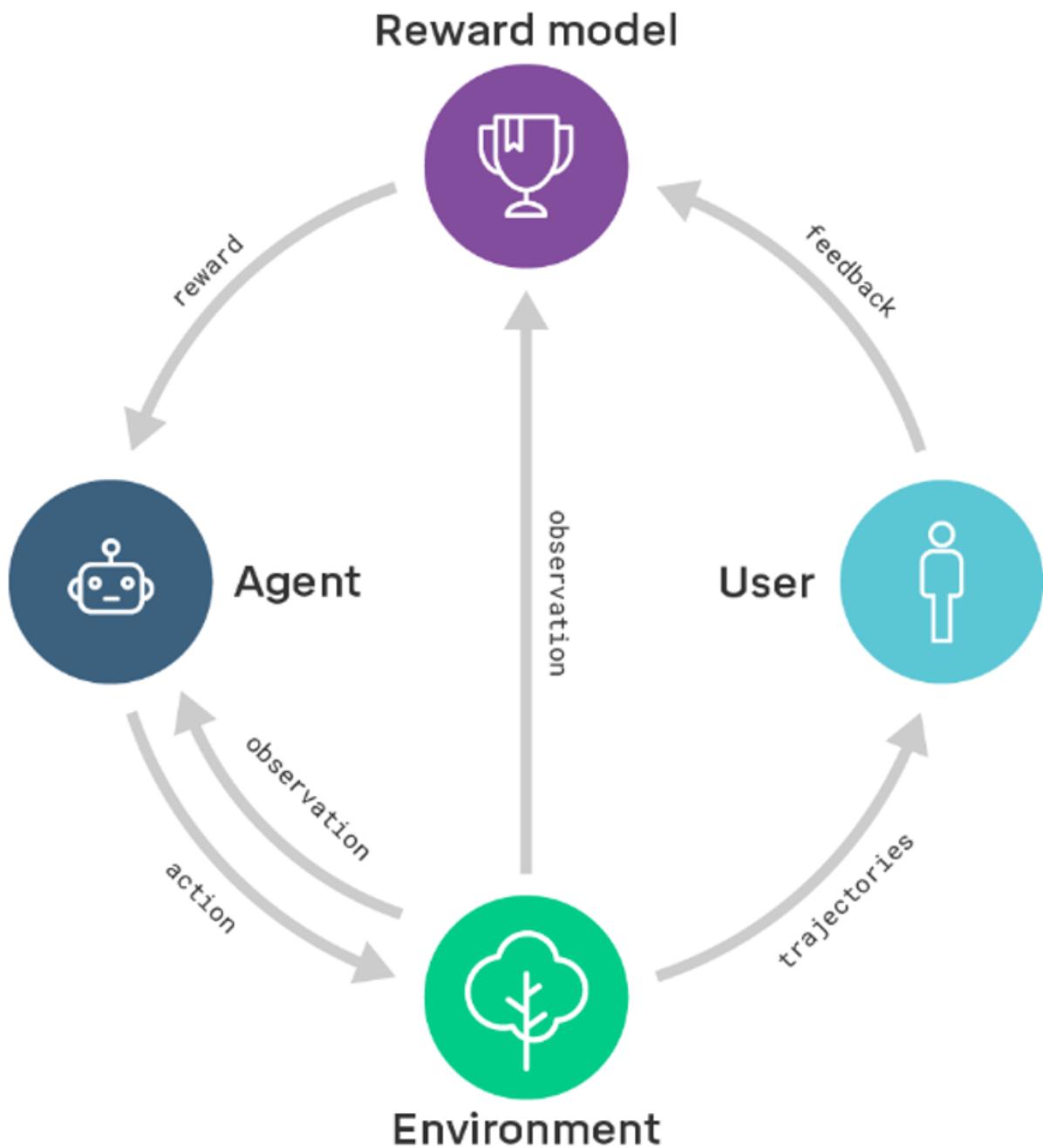
However, if one of the major use cases for your first advanced AI is helping to align your second advanced AI, STEM AI seems to perform quite poorly on that metric, as it advances our technology without also advancing our understanding of alignment. In particular, unlike every other approach on this list, STEM AI can't be used to do alignment work, as its alignment guarantees are explicitly coming from it not modeling or thinking about humans in any way, including aligning AIs with them. Thus, STEM AI could potentially create a [vulnerable world](#) situation where the powerful technology produced using the STEM AI makes it much easier to build advanced AI systems, without also making it more likely that they will be aligned.

This problem could potentially be mitigated if the STEM AI were heavily focused on applications that could potentially assist with alignment such as [whole brain emulation](#), though to what extent that would actually be possible or actually help with alignment is quite unclear.

7. Narrow reward modeling + transparency tools

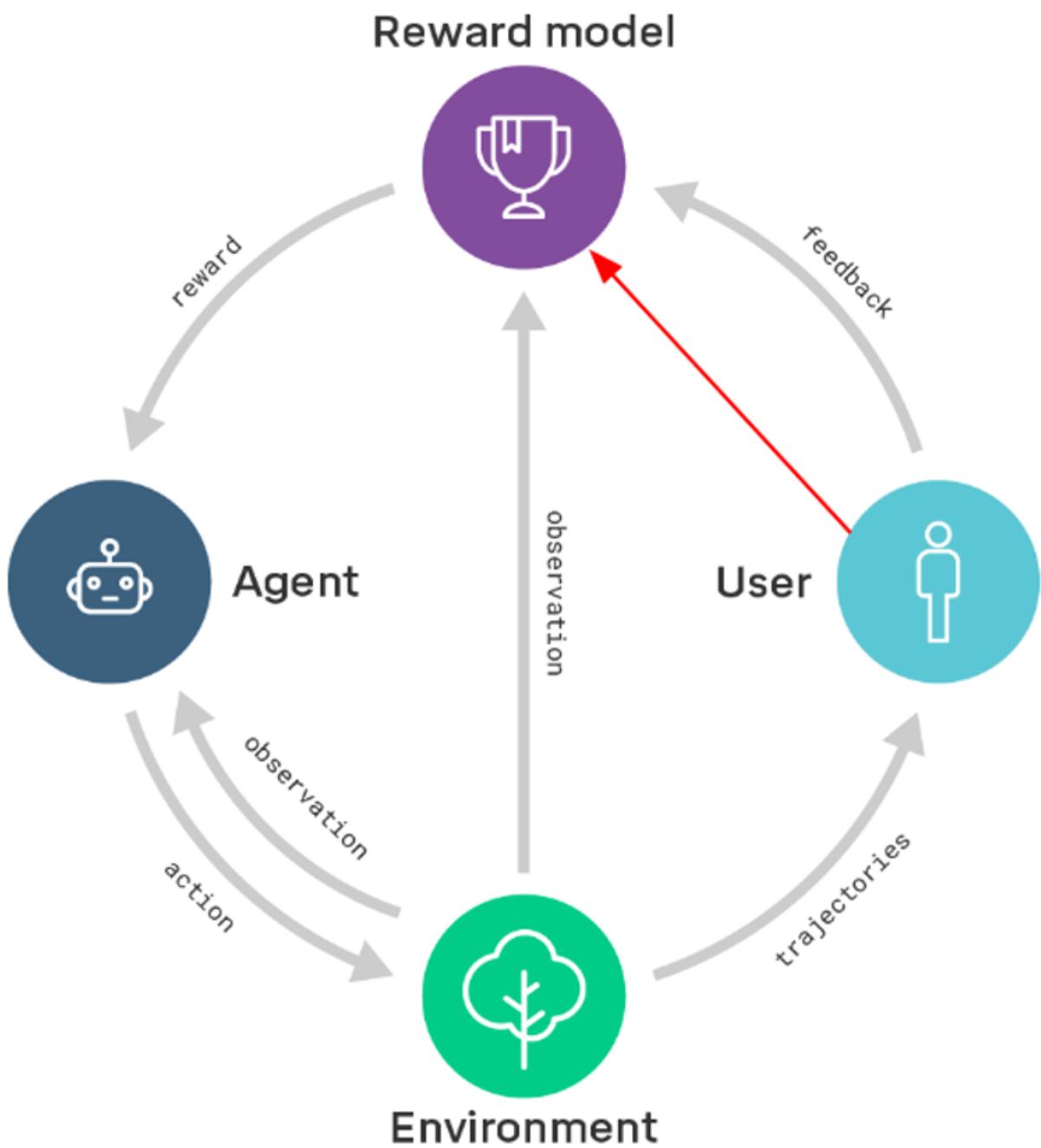
The approach here is as follows:

1. For some particular problem/domain/environment, jointly train a reward model using human feedback and an agent that pursues that reward. Specifically, we can use an approach like that described in DeepMind Safety's "[Scalable agent alignment via reward modeling](#)" where human feedback on agent trajectories can be used to refine the reward model, as seen below.



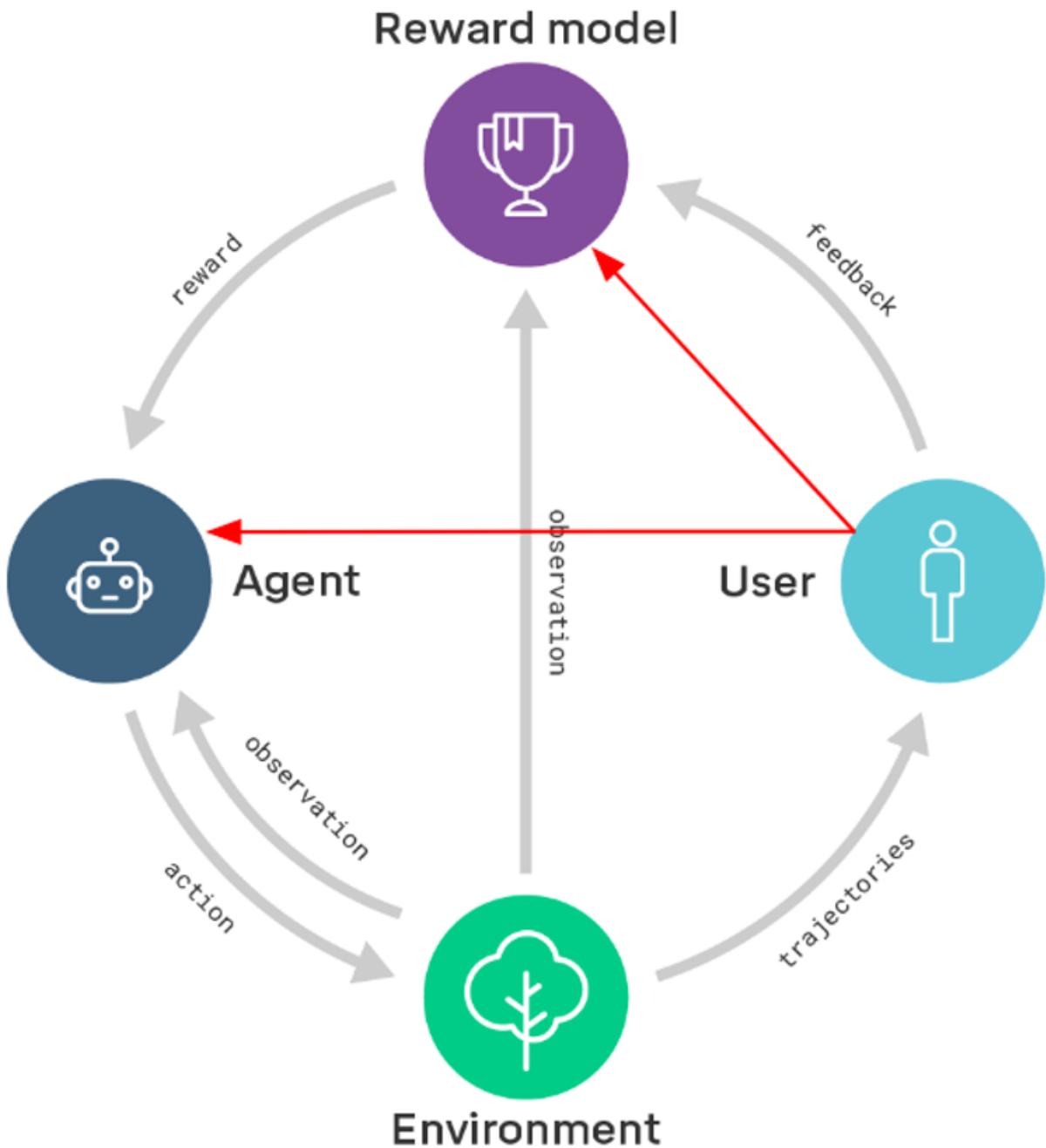
DeepMind Safety's diagram of reward learning as seen in "[Scalable agent alignment via reward modeling](#)."

2. Since we want to keep our agent confined to only the specific distribution on which it was trained (thus helping prevent possible perverse generalizations), we can also incorporate [out-of-distribution detection](#) to train the agent to halt if it encounters a sufficiently out-of-distribution state.
3. We can also have the human provide additional feedback on the reward model's internal behavior via the use of transparency tools.



Reward modeling plus oversight of the reward model by the user.

And we can do the same thing with the agent, giving us a similar sort of mulligan as with reinforcement learning + transparency tools.



Reward modeling plus oversight of both the reward model and the agent by the user.

Outer alignment. The question of outer alignment here is going to depend on whether the various feedback mechanisms are sufficient to ensure that, in the limit, the learned reward model is correct over the domain where the action model believes that it is on distribution. This is a highly complex question that depends on a bunch of factors including how good the human's trajectory and transparency feedback is, how hard it is to train an on/off distribution model, whether the reward model will itself develop optimization in the limit, the extent to which it's possible for the agent to trick the user into thinking it's doing a good job when it's not, etc. Because outer alignment here is dependent on the limiting behavior of such a complex, dynamic process, it's

very hard to determine one way or the other how outer aligned such a proposal is likely to be.

Inner alignment. The use of transparency tools in this approach is actually quite similar to the various relaxed adversarial training approaches since the feedback mechanism here allows the oversight signal coming from the human to be incorporated directly into training. That being said, unlike in relaxed adversarial training—and similarly to the “mulligan” approach in RL + transparency tools—the human doing the oversight here gets no help from the model. Thus, inner alignment for narrow reward modeling + transparency tools comes down to the question of whether a human using transparency tools to perform oversight can prevent the development of [deceptive](#) or otherwise [catastrophic](#) behavior in either the reward model or the agent. This could be done, for example, by verifying that the reward model isn’t performing any optimization (as in microscope AI) and verifying that the agent is [myopic](#) (as was previously suggested could work for amplification).

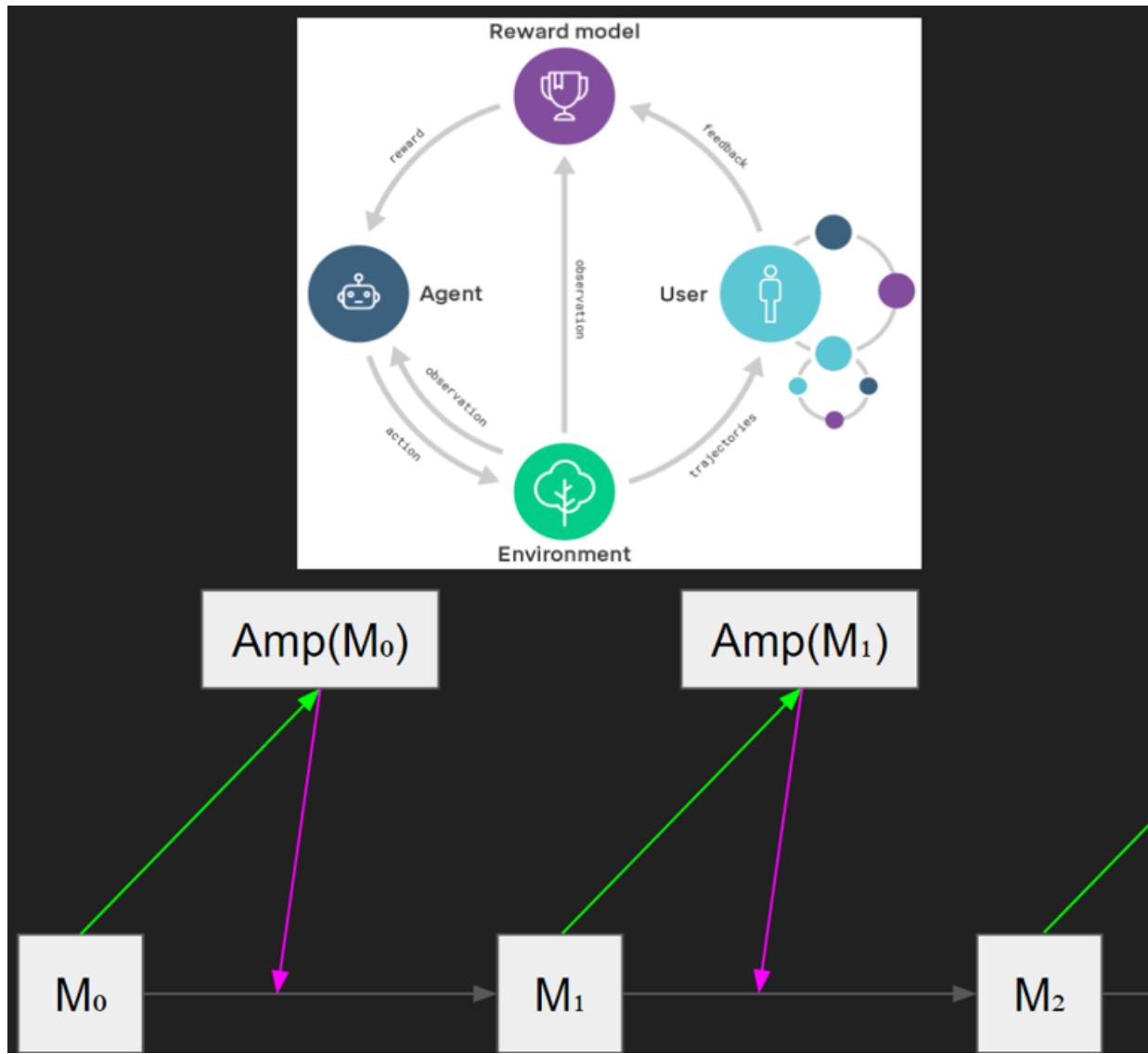
Such an approach could potentially require that the reward be myopic, however, which might be a fairly serious limitation on the applicability of narrow reward modeling if it was restricted to only those domains with myopic rewards. While it is not necessarily the case that you can’t do myopia verification with a non-myopic reward—you just need to verify that the model treats each episode myopically rather than each step—per-episode myopia verification could be significantly more difficult than per-step myopia verification, as it is potentially a much less natural behavior.

Training competitiveness. The question of training competitiveness here is going to depend on the extent to which reward modeling is a good (compute-efficient, sample-efficient, etc.) way of training AI systems to solve poorly-defined, fuzzy tasks. There are lots of different approaches to solving this basic problem, including not just reward learning approaches but also imitation learning and [various different](#) inverse reinforcement learning schemes, and specifically what approach ends up coming out on top still seems quite uncertain. That being said, reward modeling has the major competitiveness advantage of only requiring and depending on human feedback, not human demonstrations, which could be significantly more reliable and easier to elicit. Furthermore, other reward learning schemes such as inverse reinforcement learning can be incorporated into reward modeling by using them to produce a better initial reward model that can then be refined via reward modeling’s feedback mechanism.

Performance competitiveness. Similarly to microscope AI or STEM AI, a potentially major concern with the narrow reward modeling + transparency tools approach is the “narrow” part. While being narrow has potential alignment advantages in terms of reducing reliance on potentially shaky or even malign generalization, it also has the major disadvantage of restricting the approach’s usefulness to only producing relatively narrow advanced AI systems. Thus, the performance competitiveness of narrow reward modeling + transparency tools is likely to depend heavily on the extent to which truly general advanced AI systems are actually practically feasible and economically necessary. For a more detailed analysis of this question, see Eric Drexler’s [“Reframing Superintelligence.”](#)

8. Recursive reward modeling + relaxed adversarial training

[Recursive reward modeling](#), as the name implies, is a sort of recursive, non-narrow version of narrow reward modeling. What this results in is effectively a form of amplification where the distillation step which was previously imitation or approval-maximization becomes reward modeling. Specifically, the basic approach here is to train a model M to maximize the reward obtained by performing reward learning on $\text{Amp}(M)$.

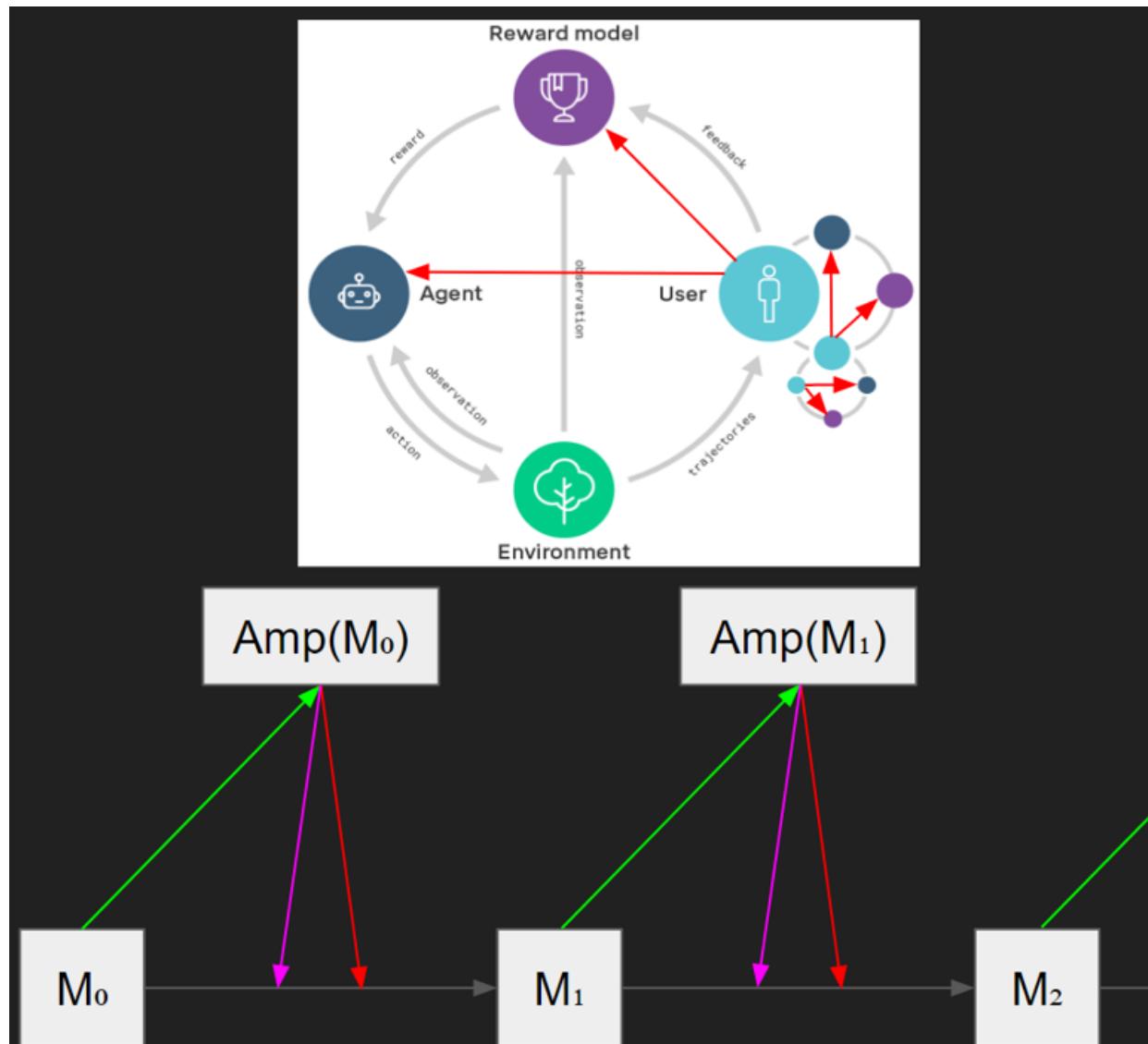


Two different, equivalent diagrams of recursive reward modeling. The top diagram is taken from "[Scalable agent alignment via reward modeling](#)" and the bottom diagram is the equivalent amplification-style diagram where the purple arrows indicate the use of the full reward modeling process.

In this graphic, the images on the top and bottom are meant to represent the same process—specifically, if you take the purple arrow in the bottom image to represent reward modeling, and assume that the agents in the top image are all the same agent just at different time steps,^[13] then you get precisely the same procedure represented

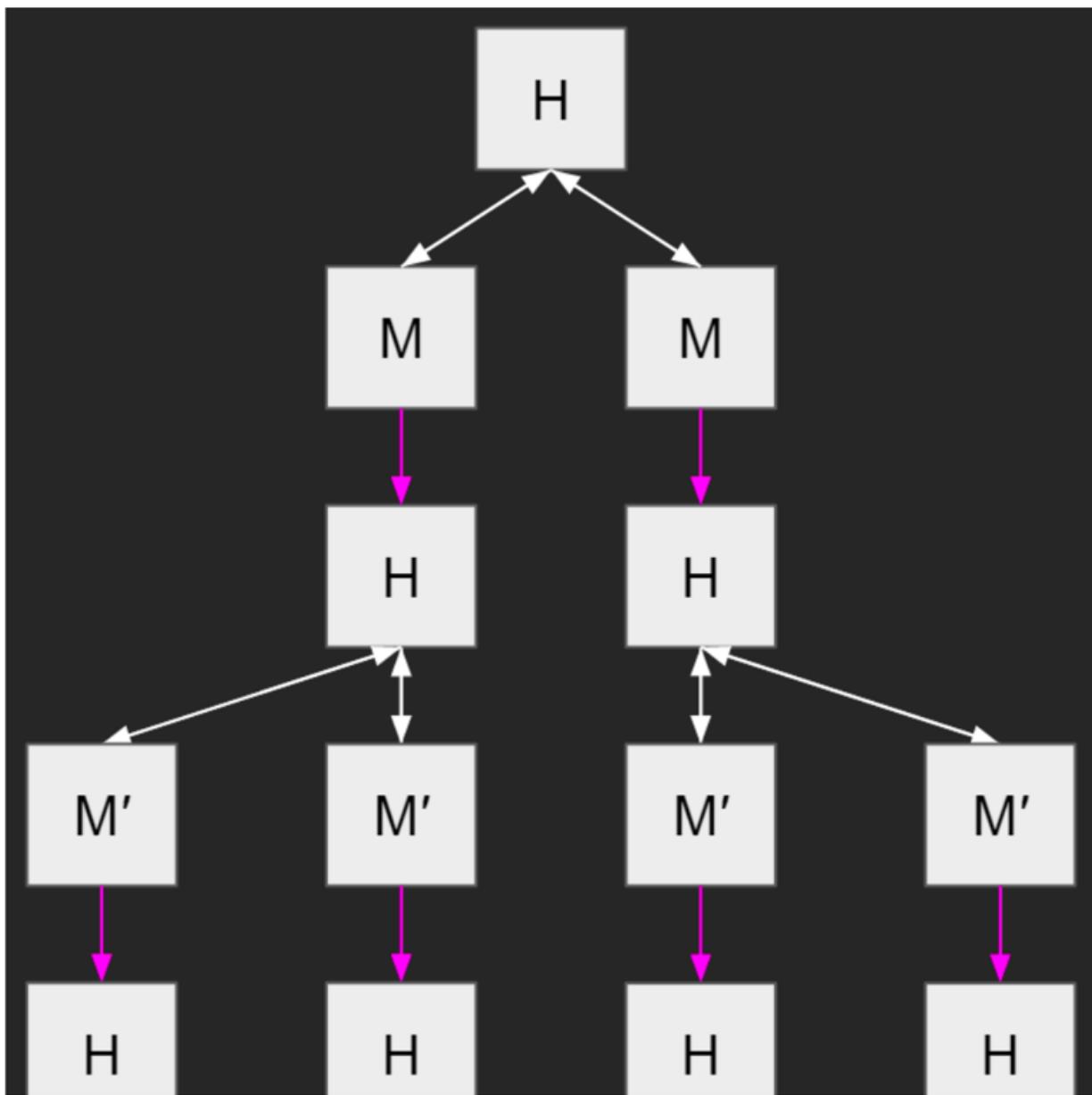
in both images. Thus, the agent in the largest circle in the top picture is M_2 in the bottom picture, the user providing feedback for that agent is $\text{Amp}(M_1)$, the agent in the medium-sized circle is M_1 , the user in the medium-sized circle is $\text{Amp}(M_0)$, and the agent in the smallest circle is M_0 .

Furthermore, similarly to the other amplification approaches, we can then add relaxed adversarial training by training M and the reward model to minimize the probability of [deceptive](#) or otherwise [catastrophic](#) behavior as assessed by $\text{Amp}(M)$.^[14] It's also worth noting that recursive reward modeling is amenable to intermittent oversight as well if desired instead of relaxed adversarial training.



Recursive reward modeling plus relaxed adversarial training where the red arrows indicate the oversight loss.

Outer alignment. The first step in understanding the outer alignment properties of recursive reward modeling is figuring out what exactly the theoretical limit of the training process is. While we don't know exactly what the limiting behavior of an individual instance of reward modeling is—as was noted previously for narrow reward modeling—we can express the limit of the recursive version in terms of many individual reward modeling limits. Specifically, we can unroll the recursive reward modeling process out over time to produce a limiting tree similar to the one given for approval-based amplification, except with approval-maximization replaced with reward modeling.



The recursive reward modeling limiting tree where white arrows indicate information transfer, purple arrows indicate reward modeling, H are humans, and M are models.

The purple arrows in this diagram represent perfect reward modeling where M is the model trained to maximize the reward obtained by performing reward modeling on H consulting M'. Now, whether this tree is aligned or not is a very open question.

Similarly to approval-based amplification—but unlike imitative amplification—it's difficult to form a clear model of what exactly this tree would do, as it not only involves humans but also models that are the limit of many individual instances of reward modeling—limits which could potentially involve deceptive or otherwise malign optimization.

Inner alignment. The question of inner alignment here is mostly going to fall on the efficacy of the relaxed adversarial training. Such efficacy could be quite different than with other amplification approaches, however, as both the model helping the human perform oversight and the model being overseen are trained via a very different process in recursive reward modeling. In particular, if the reward model is non-myopic, recursive reward modeling could rule out the possibility of using per-step [myopia verification](#)—as was suggested for the other amplification approaches—though per-episode myopia verification could still be possible, as with narrow reward modeling. If per-episode myopia verification is not tenable, however, then an alternative condition that rules out deception while being possible to verify for agents trained via recursive reward modeling might need to be found. Furthermore, if reward modeling has a greater tendency to produce deception than imitation learning, oversight could be significantly harder with recursive reward modeling than with imitative amplification even if such a condition is found. Alternatively, if recursive reward modeling helps produce models that are more capable of assisting with oversight—because reward modeling is more capable of training models to effectively apply transparency tools than imitation learning is, for example—then relaxed adversarial training could work better with recursive reward modeling.

Training competitiveness. The training competitiveness of recursive reward modeling depends on the effectiveness of reward modeling not just as an efficient way of training a model to solve a single fuzzy task—as in narrow reward modeling—but instead the effectiveness of reward modeling in training a general model which can solve an entire collection of fuzzy tasks. That being said, many of the nice training competitiveness properties of reward learning continue to apply even in the recursive setting. For example, unlike imitative amplification—but similarly to approval-based amplification—recursive reward modeling relies only on human feedback rather than on human demonstrations. Furthermore, compared to approval-based amplification, recursive reward modeling is non-myopic, which could allow it to solve credit assignment problems that might be difficult for approval-based amplification.

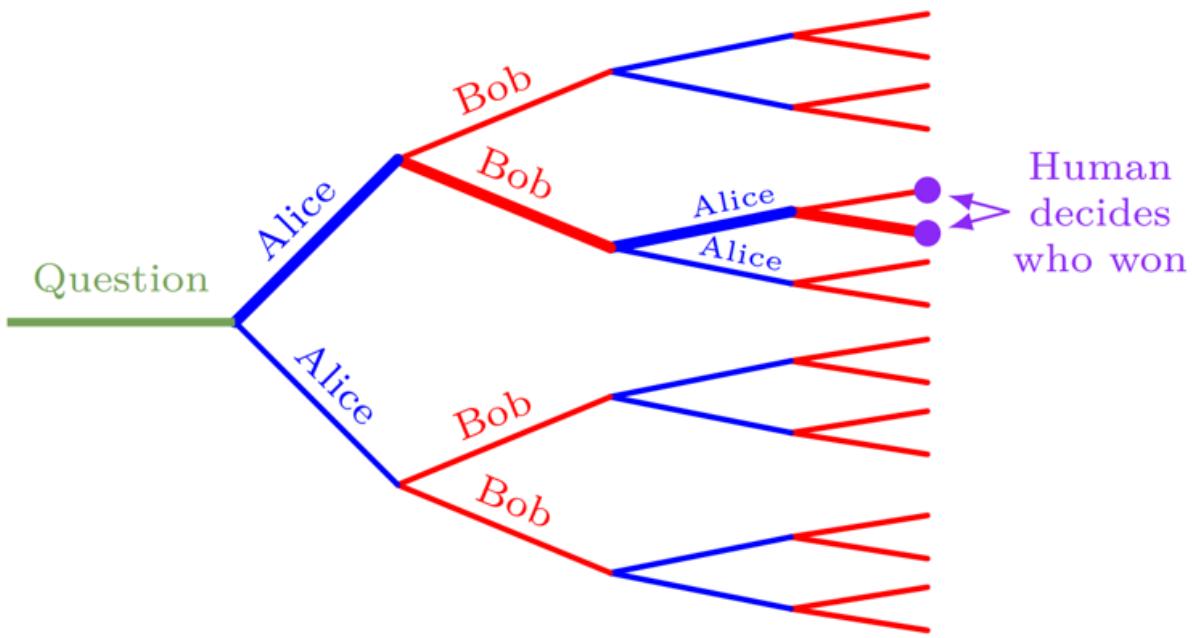
Performance competitiveness. Performance competitiveness for recursive reward modeling will depend on the competitiveness of its aforementioned limiting tree. Compared to HCH, the recursive reward modeling tree can consider ideas that no human would ever produce, potentially increasing competitiveness. And compared to the approval-maximization tree, the recursive reward modeling tree can learn to execute long-term strategies that short-term approval maximization wouldn't incentivize. That being said, both of these facets of recursive reward modeling have the potential for danger from an alignment perspective. Furthermore, if the different

models in the recursive reward modeling tree each assign some different value to the final output—which could happen if the models are not per-episode myopic—they could try to jockey for control of the tree in such a way that not only hurts alignment but also competitiveness.

9. AI safety via debate with transparency tools

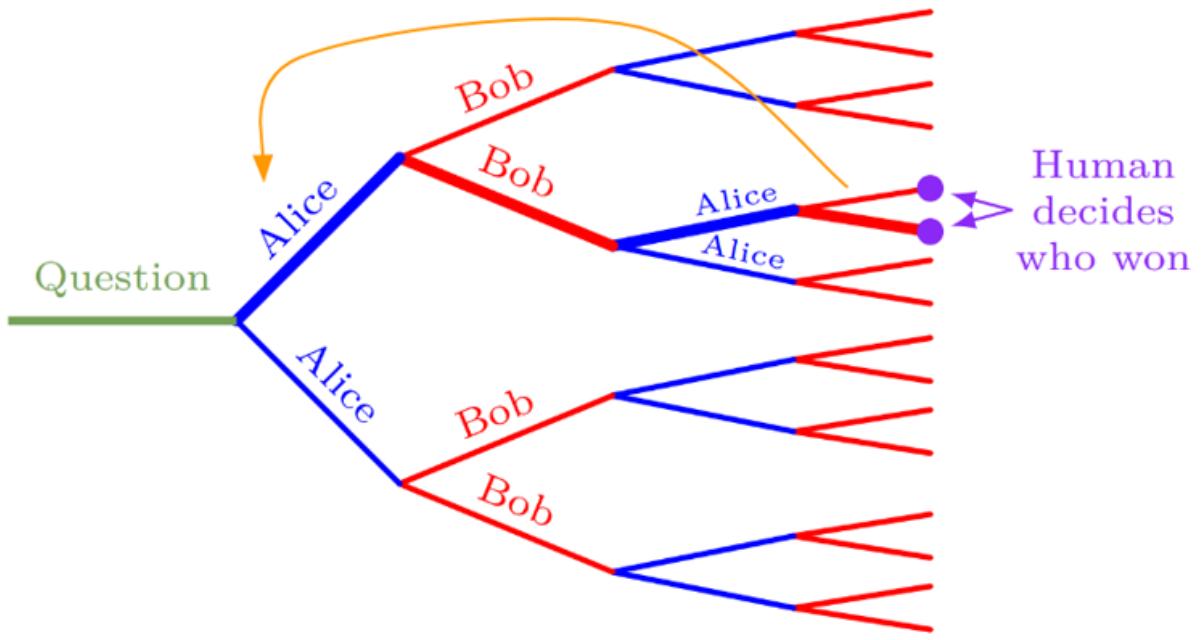
There are many different forms of [AI safety via debate](#), but the approach we'll be considering here is as follows:

1. Train a model ("Alice") to win debates against a copy of itself ("Bob") in front of a human judge.



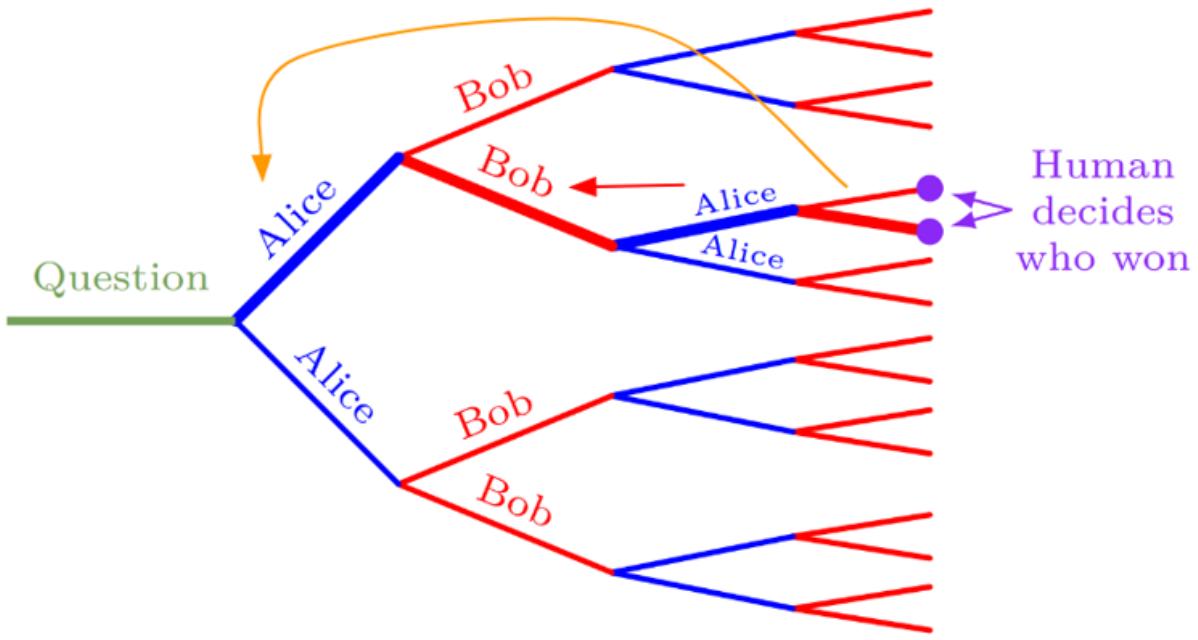
The basic debate setup where Alice is the first debater and Bob is the second debater. Blue lines indicate possible arguments that Alice can make and the red lines indicate possible arguments that Bob can make.

2. Give the models the ability to cross-examine previous versions of the other model (e.g. Bob_2 can ask Alice_1 questions for which the answers are visible to the human judge). See Beth Barnes and Paul Christiano's "[Writeup: Progress on AI Safety via Debate](#)" for a thorough explanation of the cross-examination procedure and its benefits.



Debate with cross-examination where the orange arrow is an example of cross-examination where Bob (specifically Bob_2, the second Bob) cross-examines a copy of Alice (Alice_1) from two moves previously in the debate.

3. Give the models access to transparency tools and incentivize them to try to use them to identify deceptive or otherwise catastrophic behavior in the other debater to win the debate (e.g. Alice_2 can inspect Bob_1's internals and win the debate by presenting evidence of problematic cognition). Combined with cross-examination, debaters should be able to inspect versions of the other debater from arbitrary prior points in the debate.



Debate with cross-examination and transparency tools where the red arrow is an example of Alice_2 using transparency tools on Bob_1.

Outer alignment. Outer alignment for debate depends on whether giving honest, cooperative, helpful advice is the strategy with the greatest probability of winning. Since the first debater gets to choose to defend whatever position they want, in the limit they'll choose to defend whatever position gives them the greatest probability of winning the debate. Thus, if honest, cooperative, helpful strategies dominate, then the first player will always pursue them in the limit, meaning that aligned answers can be produced at test/deployment time simply by eliciting the first debater's position. Whether honesty and helpfulness actually dominate in the limit is a very open question, however, and one that depends heavily on many factors such as the set of questions being asked, the effectiveness of the human judges, the length of the debate, etc. For example, if there is significant variation in how difficult different positions are to defend that is independent of their truthfulness (perhaps because arguments with more moving pieces that depend on more things are harder to defend regardless of how true they are) then it seems likely that there would be instances where the truth wouldn't dominate. That being said, careful mechanism design such as the use of cross-examination could potentially help mitigate such issues.

Inner alignment. The question of inner alignment for debate with transparency tools centers on the effectiveness of the debaters in using the transparency tools to oversee each other. The goal is to directly incentivize the debaters to look for deceptive or otherwise catastrophic behavior in each other using the available tools and present that information to the judge to win the debate. Whether such an approach is actually likely to succeed is dependent on a couple of different factors, however.

First, it's unclear what sort of condition could be enforced in a debate setup that would successfully rule out [deceptive alignment](#) while still allowing for the necessary deception of the second debater, who is forced to defend the second-most probable option, and thus make arguments that they don't actually believe in. Per-episode

myopia could potentially present such a condition, though it might be a difficult condition to enforce.

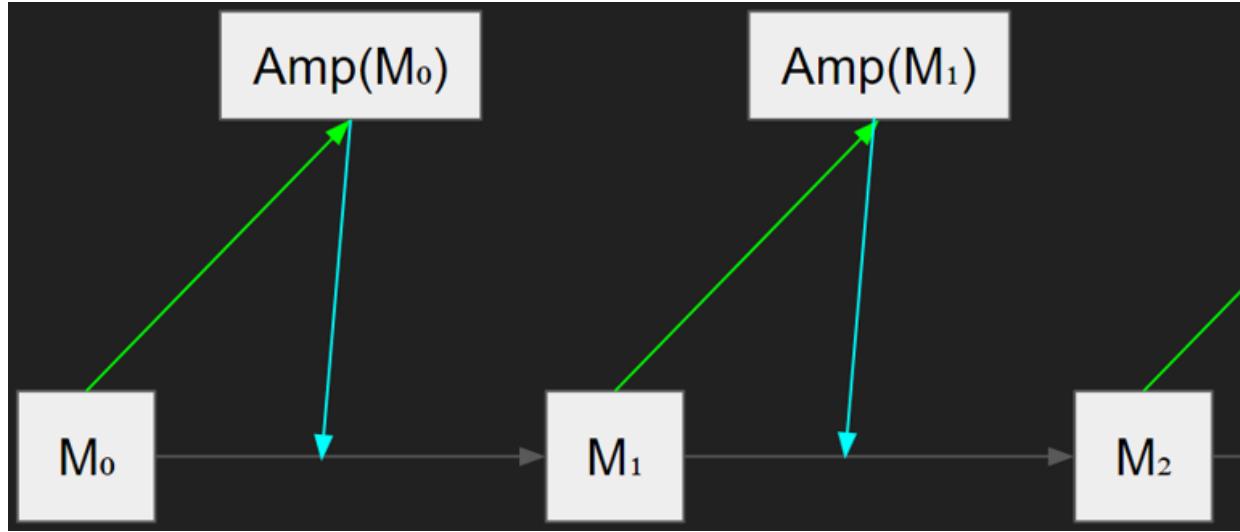
Second, getting a good training signal for incentivizing such transparency tool use could be quite difficult. The debaters need a way of learning how to inspect other models, a way of learning that doing so is a way to win the debate, and a way of learning the specific bad behavior that they should be looking for. This gets especially tricky if the condition being enforced is just the absence of deception, as getting training examples of catching deception would seem to require producing deceptive models, which is the very thing you're trying to avoid doing in the first place.

Training competitiveness. Training competitiveness for debate depends on the extent to which the debate game is a compute-efficient, sample-efficient way of training an advanced AI system. Since the debate training process is fundamentally reinforcement learning via self-play on a zero-sum game, the historical success of such methods on environments such as [Go](#), [Dota 2](#), or [StarCraft](#) suggests that such self-play might be fairly competitive as a general training procedure. That being said, in comparison to the given environments, debate requires a human in the loop, significantly reducing the amount of training data that can be collected and thus requiring significantly better sample efficiency to train successfully. Such sample efficiency concerns could potentially be resolved via training a supervised model to predict human judging behavior and then training against that, though that could raise many of its own additional safety and/or competitiveness concerns.

Performance competitiveness. [It has been proven](#) that debate with cross-examination has the property that arbitrarily strong debaters can convince a polynomial-time judge of the answer to any problem in [NEXP](#). Thus, there is a strong theoretical argument for the performance competitiveness of debate—in the limit, debate has the ability to solve basically any interesting problem that can be solved. While it can be shown that approaches such as approval-based amplification and recursive reward modeling can similarly access [NP](#), the ability of debate to go significantly beyond NP all the way to NEXP suggests that debate might be significantly more performance competitive than those approaches. Of course, such theoretical analysis doesn't necessarily apply in practice—in reality, even in the limit of training, no model can ever actually be arbitrarily strong, so the practical difference between accessing NP and accessing NEXP might be very minimal.

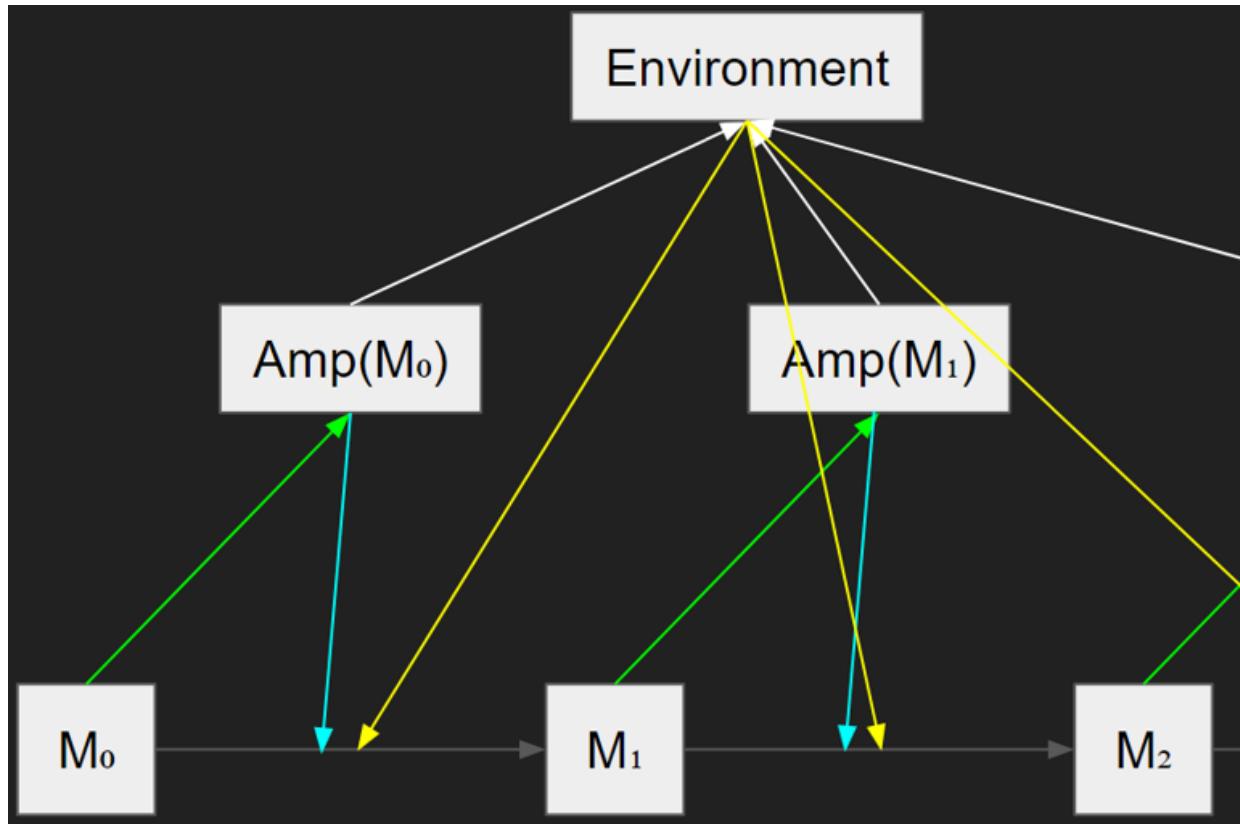
10. Amplification with auxiliary RL objective + relaxed adversarial training

The approach here is to train a model M on a joint objective composed of three different pieces. First, we'll train M to imitate $\text{Amp}(M)$ —that is, the basic imitative amplification loss.



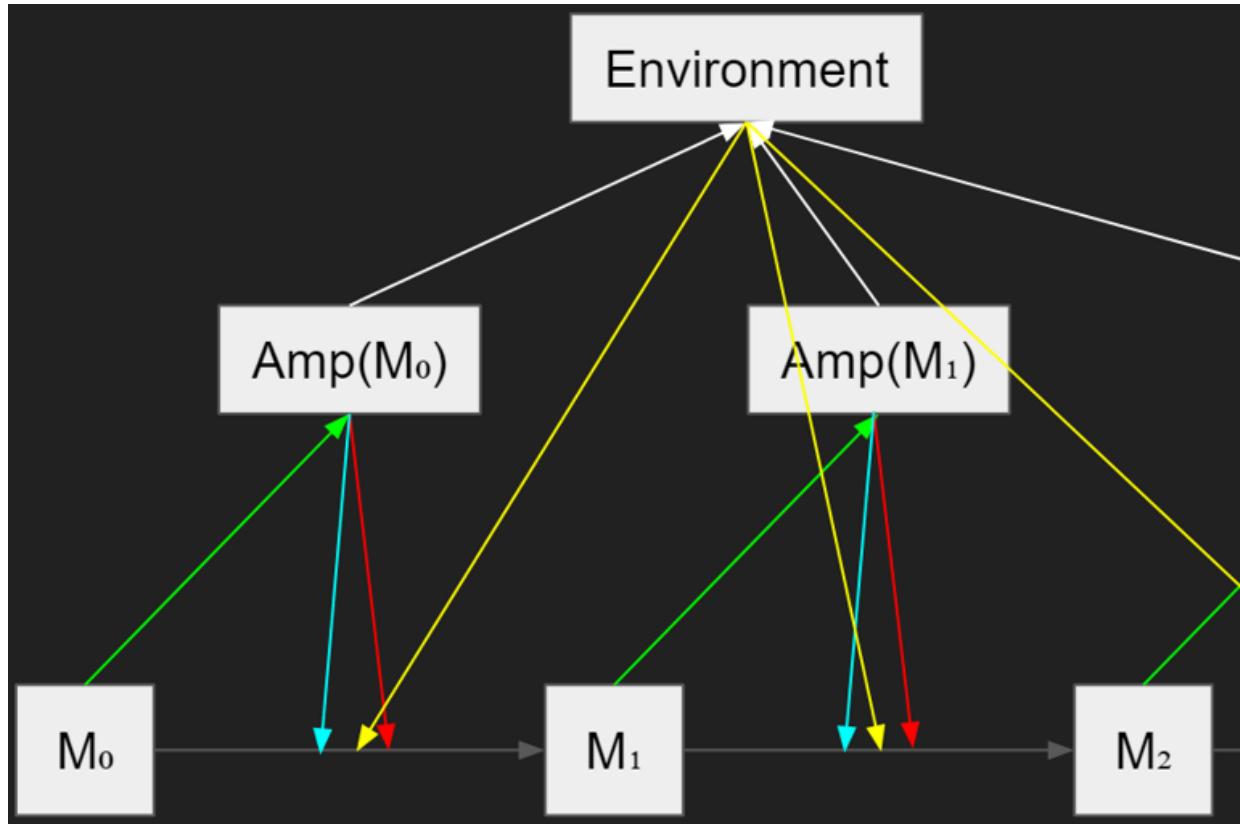
The basic imitative amplification setup.

Second, we'll also train M so as to cause $\text{Amp}(M)$ to perform well on some RL objective in some environment. The idea here is to treat M 's answers to H 's questions as actions that yield some reward determined by H 's performance in the environment with access to those answers. Thus, we are training M to maximize a combination of the standard imitative loss and the RL reward obtained by $\text{Amp}(M)$.



Imitative amplification with an auxiliary RL objective where the white arrows indicate $\text{Amp}(M)$ acting in the environment and the yellow arrows indicate the RL loss for training M to maximize the reward obtained by $\text{Amp}(M)$.

Finally, we can add relaxed adversarial training by training M to decrease the probability of catastrophe as assessed by $\text{Amp}(M)$ trying to find adversarial attacks.



Imitative amplification with an auxiliary RL objective plus relaxed adversarial training where the red arrows indicate the oversight loss.

As a concrete example of this approach where the RL environment in question is the AI safety via debate game, see "[Synthesizing amplification and debate](#)."

Outer alignment. An interesting property of this approach is that its outer alignment properties can be controlled by determining how to anneal the various different components of the loss in the limit. Annealing towards the imitative loss yields HCH as the limit—giving this approach the same outer alignment properties as imitative amplification—whereas annealing towards the RL loss yields whatever the dominant behavior in the given RL environment is as the limit. Given that HCH is generally significantly easier to understand than RL limiting behavior, it seems likely that in the majority of cases you will want to anneal towards the imitative loss to get the best outer alignment properties. That being said, there are other options—for example, you could use the imitative loss as a constraint such that the model is trained to do well on the RL objective without deviating from HCH by more than some fixed amount. Whether such an approach is outer aligned or not is likely to depend heavily on the specific metric used to measure the distance from HCH as well as the robustness of HCH to small deviations according to that metric.

Inner alignment. We're using relaxed adversarial training here, which means that—as with all the other relaxed adversarial training approaches—inner alignment is likely to come down to whether or not the overseer is able to produce a consistent gradient away from deception, which as always could be quite different than with other approaches since both the model helping with oversight and the model being overseen are being trained differently. Compared to imitative amplification, the addition of an RL objective could help significantly if performance in the RL environment is highly

correlated to oversight effectiveness. Furthermore, the fact that the model is being trained to maximize the performance of the human with access to the model means that the model is forced to learn how to present its information to a human to get reward, which could make it significantly more transparent and help with oversight. On the other hand, explicitly training the model to produce answers which change a human's behavior could also encourage the model to trick the human, which could make deception more likely and thus oversight harder. Additionally, training on an RL objective could make it significantly harder to do [myopia verification](#) as optimizing even partially for a non-myopic RL loss could act as a strong force pushing the model away from myopia.

Training competitiveness. Compared to imitative amplification, the addition of an RL loss could significantly improve training competitiveness by producing a much richer reward signal, though at the cost of having to run the RL environment. Furthermore, the addition of the imitative loss could even improve training competitiveness relative to pure RL by providing a less sparse reward. Additionally, similarly to recursive reward modeling, incorporating an RL loss enables directly training the model to solve long-term tasks involving difficult credit assignment problems that might be hard for imitative amplification alone to handle (though similarly to recursive reward modeling this trades off with the potential safety benefits of myopia).

That being said, it is unclear what happens to that behavior if the RL loss is annealed away—ideally, if it is possible for HCH to produce the behavior, then hopefully the model will converge on that, though that requires the optimal RL behavior to be close enough to HCH that first training on the RL loss and then training on the imitative loss actually helps with the imitation task. For that to be the case, success on the RL task likely needs to be highly correlated with good HCH imitation, for which language modeling and human approval maximization tasks (such as the AI safety via debate game as in “[Synthesizing amplification and debate](#)”) could be good candidates.

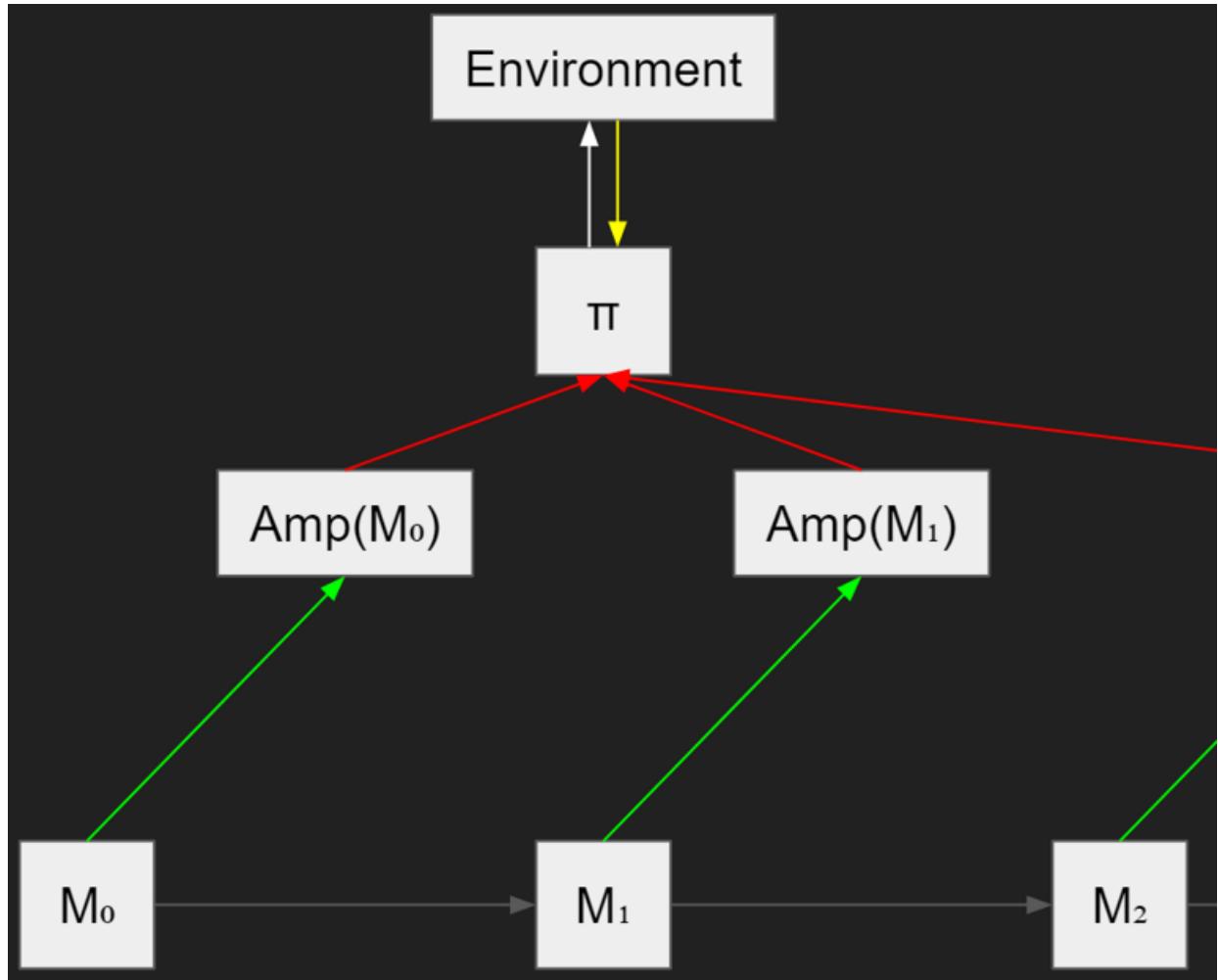
Performance competitiveness. As with outer alignment, performance competitiveness is likely to depend heavily on exactly how the different losses are annealed in the limit. Annealing towards the imitative loss, as was suggested previously, makes the limit HCH, giving this approach the same performance competitiveness properties as imitative amplification. As was mentioned previously, however, there are other options, such as using the imitative loss as a constraint such that the model is trained to maximize the RL reward without deviating too far from HCH. Such a constraint could help performance competitiveness compared to training purely on the RL loss—if the resulting model is significantly better at interacting with humans, for example—or hinder performance competitiveness—if it's very difficult to do well on the RL task while staying close to HCH.

11. Amplification alongside RL + relaxed adversarial training

In this approach, the goal is to simultaneously train two models, a question-answering system M and an RL agent π .

We'll train π on the joint objective of maximizing the RL reward and decreasing the probability of deceptive or otherwise catastrophic behavior as assessed by $\text{Amp}(M)$.

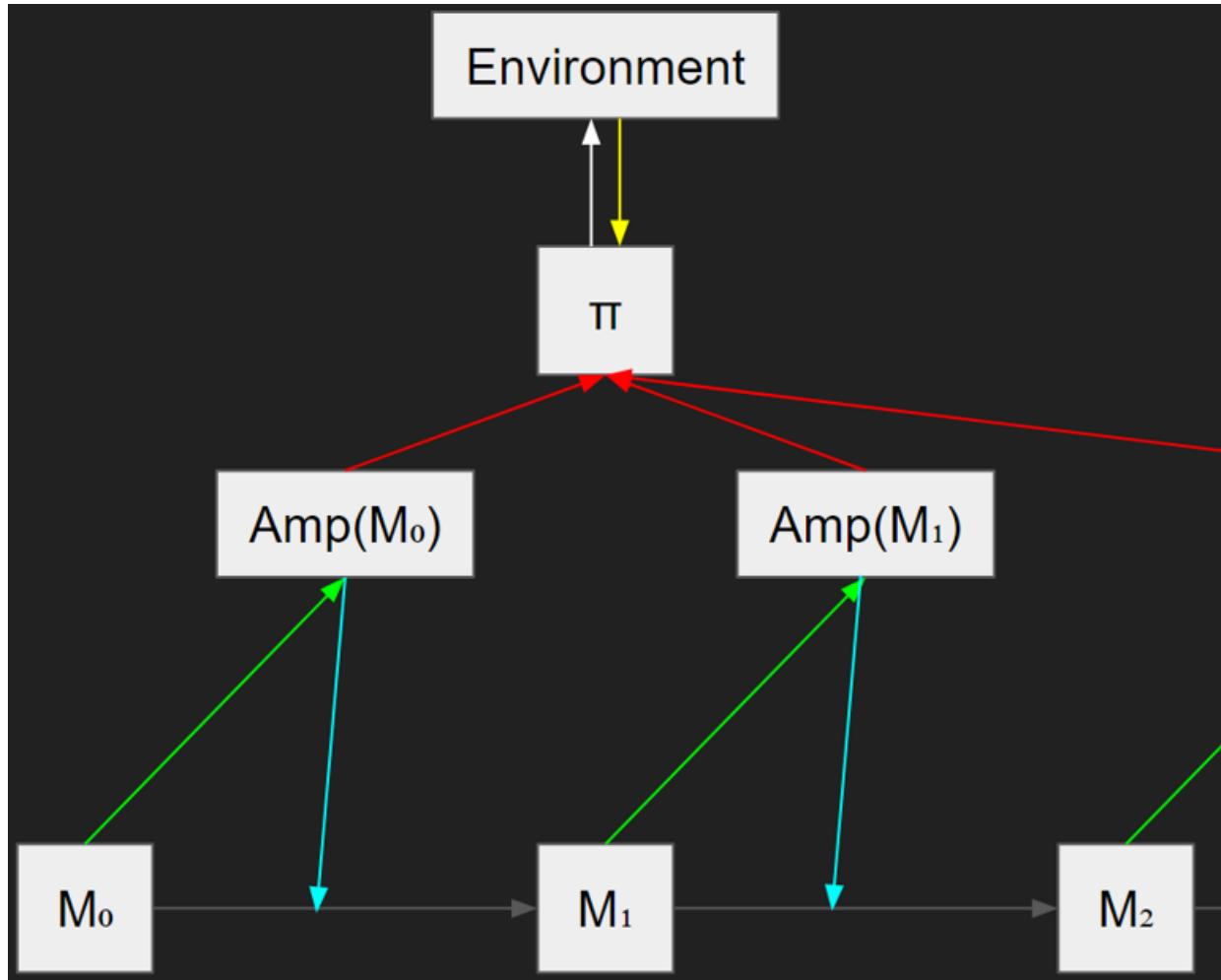
Note that while we're using a sort of relaxed adversarial training loss here, the overseer is using a totally different model (M) than the model being overseen (π).



The training process for π in amplification alongside RL where the white arrow indicates π 's actions in the environment, the yellow arrow indicates π 's RL loss, and the red arrows indicate the oversight loss for π .

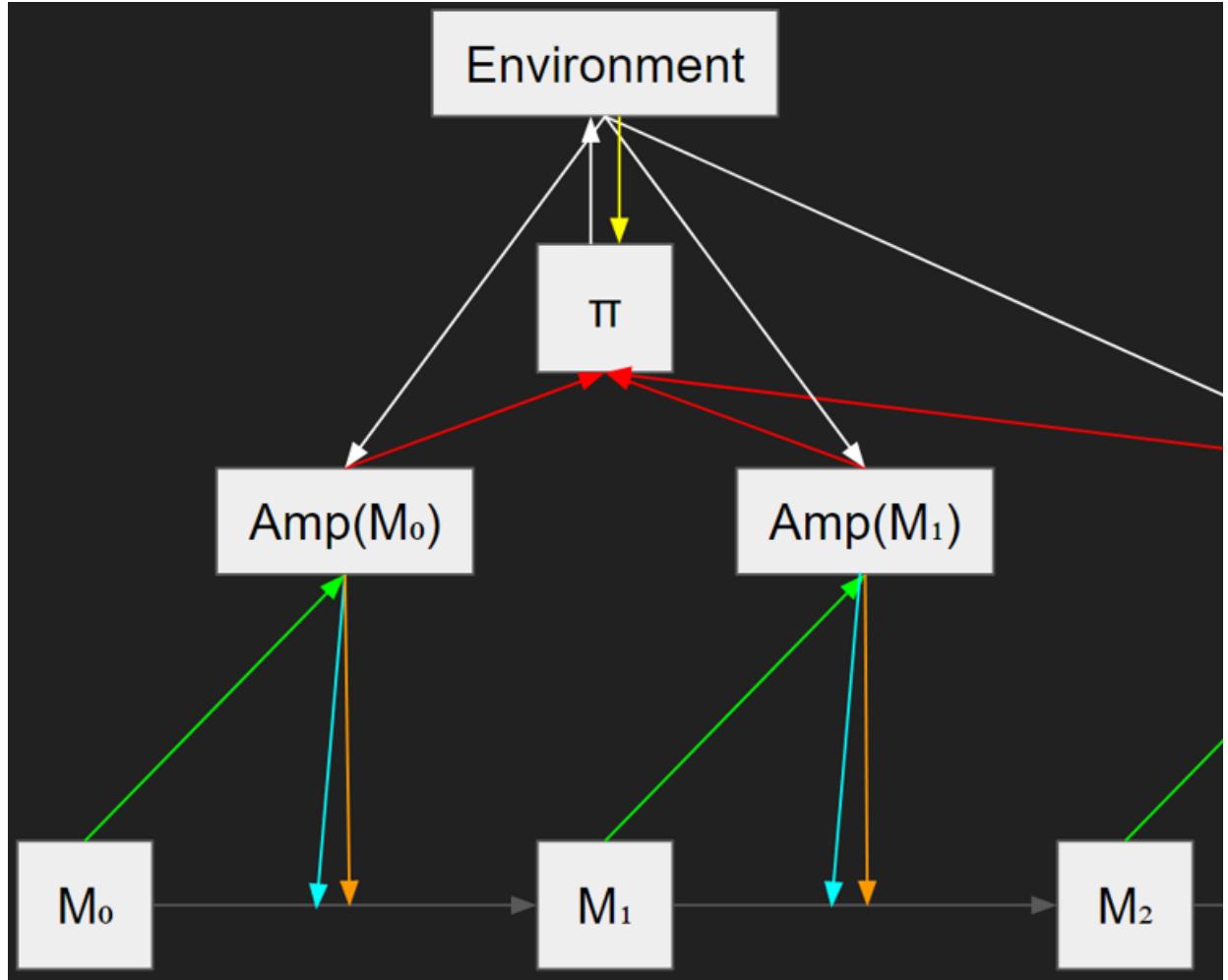
Then, we'll train M on the following joint objective.

First, we want to include the standard imitative amplification loss and train M to imitate $\text{Amp}(M)$.



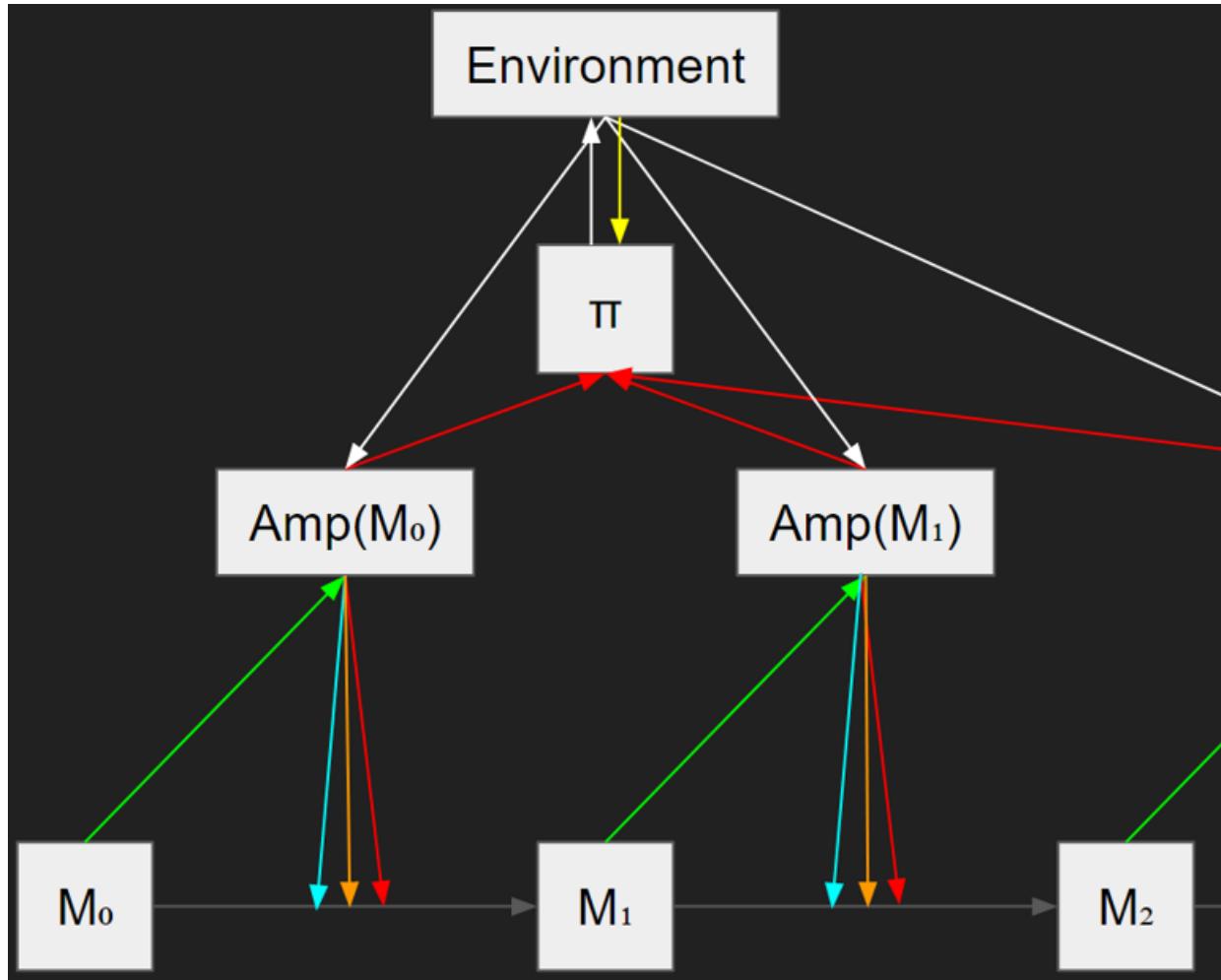
Amplification alongside RL where M is trained only via an imitative loss.

Second, we want to train M to cause Amp(M) to predict π's actions and the RL environment's responses well. The idea here is to train M to tell the human everything the human needs to know to understand what π is doing and how to replicate π's behavior so that the human with access to the model can match π's performance in the RL environment.



Amplification alongside RL where the white arrows to Amp(M) indicate information that Amp(M) attempts to predict in the RL environment and the orange arrows indicate the prediction loss for training M to cause Amp(M) to predict well.

Third, we can add standard relaxed adversarial training and also train M to decrease the probability of deceptive or otherwise catastrophic behavior as assessed by $\text{Amp}(M)$.



Amplification alongside RL + relaxed adversarial training where the red arrows indicate the oversight loss.

Finally, once we've trained both M and π via the above losses, a critical component of this approach is that we then do away with π completely and exclusively deploy the resulting M .

Outer alignment. Since we only deploy M , not π , the question of outer alignment here should be entirely dependent on whether this procedure produces outer aligned M . Furthermore, similarly to the previous approach, we can anneal the combination of losses here to get the desired limiting behavior and thus control our outer alignment properties. In particular, if we anneal towards the imitative loss, this approach can be made to limit to HCH, giving it the same outer alignment properties as imitative amplification. Furthermore, even if we don't anneal towards the imitative loss, as long as it is possible for HCH to fully predict both π and the RL environment in the limit and the imitative loss isn't annealed away completely, then at minimal loss we should still end up with HCH.

Inner alignment. Inner alignment is particularly tricky for this approach. First, there's the standard relaxed adversarial training question of whether Amp(M) will be able to properly oversee M. Compared to the previous approach, the fact that the auxiliary reward is purely predictive rather than a full RL loss should help us here, as it means that our loss is entirely myopic and thus we should be able to do [myopia verification](#). That being said, if π is sufficiently non-myopic, having M predict π while still staying myopic itself could be a challenge.

Additionally, there's also the problem of π itself—while we throw away π at the end, there are still likely to be opportunities for a sufficiently competent and deceptive π to cause serious issues. For example, π could choose its actions to influence M's training in a particular way—or just break out of its sandboxing and directly take bad actions in the world. Thus, we need the oversight of π by Amp(M) to at least work sufficiently well to prevent such issues. For that to happen, it needs to be the case that Amp(M) with access to transparency tools can predict when/if π will ever produce such catastrophic behavior. Thus, the fact that we're explicitly training M to cause Amp(M) to be able to predict π could be quite helpful here, as that should hopefully make Amp(M) good enough at predicting π to foresee π becoming deceptive or producing any such bad behavior.

Training competitiveness. Compared to other amplification approaches, this approach could be significantly more training competitive as it enables us to just train an RL model with minimal restrictions. That being said, it also requires us to jointly train both an RL model and an amplification model, which could make it significantly less competitive, especially if it's difficult to successfully train the amplification model to predict the RL model. That being said, such training can be viewed as effectively just model distillation, which is generally fairly doable even for current models. That being said, the inclusion of the imitative loss will change things, especially if you anneal towards the imitative loss in the limit. In such a situation, the question becomes how useful the auxiliary RL prediction task is in helping with the core imitation task, which—similarly to the previous approach—could be very dependent on the specific RL environment used and how related performance on that RL environment is to the imitation task.

Performance competitiveness. As with outer alignment, annealing towards the imitative loss should produce HCH as the limit, giving this approach the same performance competitiveness properties as imitative amplification. Furthermore—again as with outer alignment—even if the imitative loss isn't annealed to completely, as long as HCH can predict π in the limit, you should still get HCH at minimal loss.

-
1. Note that the order here is chosen purely for pedagogical reasons and is not meant to imply any sort of measure of importance. ↩

2. Note that in practice $\text{Amp}(M)$ doesn't need to always be computed with an actual human. H can simply be M some fixed fraction of the time, for example—or more often when M is confident and less often when M is uncertain—and the resulting procedure is effectively the same. See “[A concrete proposal for adversarial IDA](#)” for an example of how something like that could work. ↵
3. Imitative amplification leads to HCH in the limit of training a perfect predictor over all possible inputs—that is, an infinite amount of training data, an infinitely large model, and an infinitely large amount of training time. ↵
4. Enforcing myopia for amplification could also have the side benefit of improving competitiveness, as a non-myopic amplification model might try to make its output simpler and thus less useful if it is trying to make a human consulting it easier to predict. ↵
5. Further exploration of myopia verification will also be forthcoming in a future post. ↵
6. See Paul Christiano’s “[Efficient feedback](#)” for some examples of possible sample-efficiency-improving techniques. ↵
7. That is, it needs to be the case that in the limit there exists a model that perfectly imitates HCH in such a way that an amplified overseer will be convinced that the imitator is non-catastrophic. ↵
8. This sort of failure mode could be particularly pernicious if the deceptive model is capable of [gradient hacking](#). ↵
9. Verifying that the given tree is, in fact, the limit of approval-based amplification is a straightforward exercise of expanding the training procedure out over time. Note that the models at the top of the tree are the most recent models and the models at the bottom of the tree are the oldest. ↵
10. It is worth noting that there is also the possibility of recovering HCH as the limit of approval-based amplification if the approval signal is implemented as a human-computed distance metric relative to what the human would have done before seeing the model's answer. ↵
11. In particular, myopia verification for approval-based amplification could be used to produce [approval-directed agents](#). ↵
12. The problem of “forwarding” safety guarantees regarding a model performing meta-learning to any other models found by that meta-learning procedure is a general problem that occurs in all inner alignment schemes, though it is particularly pernicious in this situation. For a more detailed discussion of this problem, see the “Meta-learning” section in “[Relaxed adversarial training for inner alignment](#).” ↵
13. “[Scalable agent alignment via reward modeling: a research direction](#)” notes that, while they initially assume that each agent is completely separate, “While this kind of sequential training is conceptually clearer, in practice it might make more sense to train all of these agents jointly to ensure that they are being trained on the right distribution. Moreover, all of these agents may share model parameters

or even be copies of the same agent instantiated as different players in an adversarial game.” Thus, while the different agents are presented here as different instances of the same model—which is a type of recursive reward modeling—it is worth noting that recursive reward modeling also includes other possibilities such as using completely different models for the different agents. [←](#)

14. “[Scalable agent alignment via reward modeling: a research direction](#)” mentions the possibility of such oversight, though does not include it as part of the base proposal as is done here, noting that “When using recursive reward modeling users have the *option* to provide feedback on the cognitive process that produced outcomes, but they are not required to do so. Moreover, this feedback might be difficult to provide in practice if the policy model is not very interpretable.” [←](#)

Literature Review For Academic Outsiders: What, How, and Why

This is a linkpost for <https://www.thelastrationalist.com/literature-review-for-academic-outsiders-what-how-and-why.html>

[A few years ago I wrote a comment on LessWrong](#) about how most authors on the site probably don't know how to do a literature review:

On the one hand, I too resent that LW is basically an insight porn factory near completely devoid of scholarship.

On the other hand, this is not a useful comment. I can think of at least two things you could have done to make this a useful comment:

Specified even a general direction of where you feel the body of economic literature could have been engaged. I know you might resent doing someone else's research for them if you're not already familiar with said body, but frankly the norm right now is to post webs spun from the fibrous extrusions of peoples musing thoughts. The system equilibrium isn't going to change unless some effort is invested into moving it. Notice you could write your comment on most posts while only changing a few words.

Provide advice on how one might go about engaging with 'the body of economic literature'. Many people are intelligent and reasonably well informed, but not academics. Taking this as an excuse to mark them swamp creatures beyond assistance is both lazy and makes the world worse. You could even link to reasonably well written guides from someone else if you don't want to invest the effort (entirely understandable).

I also linked [a guide from Harvard's library](#) (Garson & Lillvick, 2012) on how to do a literature review. But this guide makes extensive use of flash video, which makes it increasingly hard to access the content. Even if flash was alive and well, video is not necessarily the most comfortable format. Worse still, I remember feeling there was a great deal of tacit knowledge excluded from the guide which wouldn't be apparent to someone that isn't already familiar with academic culture. Even if the guide was a perfect representation of how to do an academic literature review, the priorities and types of work put together by LessWrong authors are more [outsider science](#) (Dance, 2008) than they are Harvard. For this reason I've had writing a guide to literature review aimed towards academic outsiders on my to-do list for a while.

At the same time I'm not interested in reinventing the wheel. This guide is going to focus specifically on filling in the knowledge gaps I would expect from someone who has never stepped foot inside a college campus. The other aspects have been discussed in detail, and where they come up I'll link to external guides.

What is a literature review?

'Literature review' the process is a way to become familiar with what work has already been done in a particular field or subject by searching for and studying previous work.

A 'literature review' is a document (often a small portion of a larger work) which summarizes and analyzes the body of previous work that was encountered during literature review, often in the context of some new work that you're doing.

Why do literature review?

Literature reviews tend to come up in two major contexts: As a preliminary study to help contextualize a novel work, or as a work itself to summarize the state of a field or synthesize concepts to create new ideas. Most of my research falls into the latter category, I'm a big fan of [putting together existing evidence and ideas to synthesize models](#) (namespace, 2020). [Gwern also tends to do work in this style](#) (Branwen, 2020). I suspect that a lot of authors on LessWrong are attempting to do this, but fail to really say anything useful because they haven't figured out how to incorporate thorough evidence into their argument. When I did a review of all my notes from 2015, I found the number one failure mode I'd fall into was not paying attention to prior art. This was because I did not have heuristics like:

- If it's hard to write about or you get stuck, you should probably do more research
- If I want to write a post on something and I haven't checked the relevant literature for it yet I should probably do that as part of writing the post
- [Encountering or generating a cool mental model](#) (Constantin, 2018) is a useful cue to consult the literature
- If I'm trying to deal with a hard technical problem I should look at what work has already been done

The Benefits Of Literature Review

Literature review provides many benefits, such as:

- **Build Off The State Of The Art:** Unless you make it a habit to look at what work already exists on a subject, you'll say what others have already said and do what others have already done. Your cognition is slow and expensive, and that makes leveraging the work of others extremely valuable. It is tempting to think that the established experts are idiots and you can beat them all with your own cleverness. [Sometimes, this is actually true](#) (Harford, 2019) but it's not something you should be counting on as a rule. [Some literatures are mud moats](#) (Smith, 2017), but other literatures are priceless treasures. Without access to the mathematics literature you would need to be [a prodigy like Ramanujan](#) to make new contributions. In my 2015 notes there was an episode where I tried designing a package manager. I filled many pieces of paper with thoughts on resolving dependency conflicts. Never did it occur to me to look at what methods were already used by existing systems like .deb or .rpm, let alone research papers that might tell me about theoretical methods.
- **Providing Context:** Cultural artifacts exist in some kind of context, historical, social, or intellectual. [Without provenance a 5,000 year old sword is just a rusty piece of metal](#) (Giuliani-Hoffman, 2020). The same principle applies to intellectual work, without a justifying context [artifacts are parsed as garbage](#) (Foddy, 2017). The literature can help you provide context for your ideas and ground them in something other than just your personal experience.
- **Learn From The Mistakes Of Others:** Bismarck famously remarked that fools learn from experience and wise men learn from the mistakes of others. Even if previous work has failed to make significant progress it can often serve as a

reference of promising-sounding ideas that won't work. This familiarity is often a crucial component of the 'cleverness' that sets you apart from others. The Wright Brothers [were very familiar with the established work on aerodynamic theory](#) (Benson, 2014). Their rapid-iteration approach to airplane design quickly revealed that real world test flights [defied their expectations](#), leading them to develop a new way to measure the performance of airplane parts. Once this was done the data enabled them to easily invent the airplane. Without that starting data to work from, it would have taken the Wrights longer to realize that data was the bottleneck to making an airplane.

- **Common Language:** Scholars develop a shared language to discuss their studies. [These vernaculars are a key marker of group membership](#) (Hossenfelder, 2016). Authors that use the right words generally have [standing](#) and authors that use their own ad-hoc vocabulary are generally considered cranks. Even beyond credibility, writing in the standard language used by other authors makes it more likely you'll get expert feedback on your work.
- **Unknown Unknowns:** Until you go looking, you often just plain don't know what you don't know about a subject. For example in my [essay on fuzzies and saddies](#) (Zealot, 2020) I didn't know that literature on morale was relevant to the research question until I started looking at the psychology of soldiers. Often when you start looking at previous work you have a "wait this exists?" moment that significantly alters the way you approach it.

Literature Review As Accessible Contribution

One question I hear often is: "How can I contribute to the rationality project without institutional resources?". Literature reviews are an accessible contribution that builds skills. [Some of the best posts on LessWrong](#) are literature reviews. The research skills that you build while doing it are extremely valuable, and will help you in most things you might want to pursue. It doesn't require very much money, and can be performed from the comfort of your home. All these traits make it nearly ideal for people who want to contribute but don't have a lot of resources, or who have to spend most of their time on school or work. Literature review does take time however, so like any volunteer work it's necessary that the person undertaking it have spare time and energy to work with. If this sounds interesting to you, feel free to [private message me on LessWrong](#) or [join this blog's Discord server](#) and I'll do my best to help.

The Document Universe

As a phenomenological definition, the document universe is the set of artifacts which are easy to access inside of academic review spaces like museums, libraries, reading/viewing rooms, or a home office. It is the spatial environment in which the literature exists. Learning to navigate this environment is essential to getting good at literature review.

People Are Documents Too

When you want to know more about a subject but aren't sure where to begin, the classic advice is to ask a librarian. Human beings are a key part of the document universe. They are intentionally created artifacts that contain knowledge, and that knowledge is backed by a full general intelligence. It's no coincidence [that Socrates](#)

[didn't like writing](#). People are arguably the most important part of the document universe. Knowledge does very little if it isn't contained inside someone.

Because of their high value and short shelf life [it can be hard to get access](#) to knowledgeable people (Hossenfelder, 2016). [It's not impossible however](#) (Dance, 2008), the received wisdom is that most scholars are eager to discuss their work *so long as you respect their time*. [Eric Raymond's classic essay](#) (Raymond, 2014) on asking good questions is oriented more towards "How do I X with program Y?" type queries, but with some mental rearranging applies just as well to plenty of other queries. For academic questions in particular it's important that you do your best to understand the science [and understand the language used by the science](#) (Hossenfelder, 2016). Failure to do that is likely to get you spam filtered as a crank.

Academic Sources Are Underadvertised

Most web users don't seem to be aware of academic sources. I remember when I was younger feeling a vague malaise as I browsed the Internet, because all the knowledge seemed to be diffuse and informal. When I read books it was clear that they were high quality sources of knowledge, but the Internet felt barren of that. It turned out this was mostly just because I was looking in the wrong place. The academic section of the document universe is [publicly indexed by Google Scholar](#) which makes it much easier to find high quality sources on most subjects.

Traditional Bibliography

The vast majority of the history of scholarship happened before the existence of electronic computers, let alone widespread high-capacity *networked* electronic computers. That means the formal norms of scholarship evolved in an environment quite alien to our current era of cheap access and full text search. In this section we'll review some of their features in that context.

Citation Trees As Central Dogma Of Academia

In school you were probably told that you had to cite your sources, and that failing to do so was plagiarism. Plagiarism is usually defined as "stealing someone else's work without credit", but in the context of citations this definition is very misleading. Grade schools like the concept because it lets them clearly define how much copying is cheating, with the unfortunate side effect that smart kids categorize the practice as schoolhouse ritual rather than valuable technique. By contrast in a functional literature where works are written to be read academic citation norms provide a genealogy of ideas. These days we're pretty used to digital documents that directly reference other pages, videos, etc; but before the Internet was widespread academia alone had the benefit of author provided citations. Academic citation formats are platform agnostic. They're [content addressed](#) rather than location based, so the goal of an academic citation is to give you enough information to reliably locate a *specific* source in the document universe. This is why they tend to get so tedious. A book might have 12 editions with multiple authors and undergo a title change, and only one version contains the passage you reference. All the annoying details in citation formats were put there in response to bibliographic failures and lookup complications with simpler formats.

Within a single work citations provide context for readers and leads for further reading, but it's when you have a whole literature that the practice really shines. It becomes possible to follow citations backwards to see the progression of ideas, move horizontally to find related work, and use modern database systems to find work downstream that cites a document as an ancestor. The genealogy aspect of academic citations also improves the signal:noise ratio by eliminating unimproved duplicate work, and makes it easier to associate ideas with the authors that originated them. All of this makes the academic sections of the document universe much more pleasant to navigate than the informal universe of newspaper articles and blog posts. From a contributors standpoint there's also more security, the norms are built to get your ideas [hooked into a network of associated work which future scholars will consult during their reviews](#) (Hanson, 2007). Outside of that Eden it's possible your effort will just get lost in the noise.

Unfortunately because [the web is a disaster](#) (Binstock, 2012) we're not really liberated from citations by the presence of hyperlinks. In an ideal world the web would be content addressed so that if a source stopped providing a document it could be seamlessly served by a backup provider like the [Internet Archive](#). Instead we address by location, so if the domain hosting this blog changes hands and they put up a new site all the links to my posts break. If I decide I don't want to pay hosting costs anymore, all the links break. If the servers have a technical malfunction even though they're technically still on in some dusty computer lab somewhere, *all the links break*. [As you might imagine this happens a lot](#) (Branwen, 2019), so it's not viable to rely on links to identify content. Traditional citations at least provide for the *possibility* that there is a second copy somewhere which can be found with a search engine. The most savvy netizens [do their best to ensure](#) (Branwen, 2019) there is a second copy somewhere. Because these problems are [unlikely to be fixed any time soon](#), if you plan to write lasting content you had best get familiar with citation formats.

Library Science As Conceptual Foundation Of The Academic Document Universe

Underlying the usefulness of a citation tree is physical infrastructure which houses, indexes, and curates documents. This type of work has been traditionally performed under the moniker of library science, even if in recent times it has mostly been done by a distributed system of bloggers, cooperating scientists, server hosts, and for-profit firms like Google. The old systems still exist however, and they're the environment of adaption for the current academic tradition. This makes it useful to know the principles of traditional library science so that you can better model academic-document-space. I recommend the book *The Intellectual Foundation Of Information Organization* by Elaine Svenonius (Svenonius, 2000) to get that understanding. Published in 2000, it was written just before digital documents were set to disrupt the academic ecosystem. It captures the full powers of the old ways in amber.

The Intellectual Foundation... is a particularly useful book for the scholar because it is designed to be read by the designers of future library systems. This means that it focuses less on the details of particular designs (which we probably don't care about very much by this point) but more on the principles which an effective system should satisfy and the "why?" behind them. These principles define the territory which citations describe, and will help you grok certain aspects of traditional scholarship.

How To Do Literature Review

I'd be a hypocrite if I didn't bother to look at what others have already written about doing a literature review. [This talk with Dr. Candace Hastings](#) (Hastings, 2009) on doing literature review is decent, it spends a lot of time focusing on the way to use sources in your writing once you've found them. She also explains how you can use citation counts to find the most important scholars in the field you're looking at. [Guidelines for writing a literature review by Helen Mongan-Rallis](#) (Mongan-Rallis, 2018) is a well written page on this topic for academics.

Every time I do research I perform a simple thought experiment: assuming somewhere in the world exists evidence that would prove or disprove my hypothesis, where is it? I tend to visualize this as a shot of earth from space, and then 'zooming in' on the sense data that would show me what I want to know. The literalism of this visualization is important because it emphasizes the sensory basis of evidence. Things happen in the world, and artifacts of their presence are left over afterwards. Physical remnants, images captured by cameras and sketch artists, written observations. This 'object level' phenomenological universe is what you're trying to get information about by looking at the literature.

A key consequence of this is that 'the literature' is not always what's output by academics. If I was studying martial arts, I would be looking into the history of martial arts as it's practiced by martial artists, in whatever mediums they use to record and disseminate information. Memory is a human activity, and your first priority should be to find the most effective and relevant sources for whatever you're looking at.

For tips on actually finding sources on the Internet (commonly known as 'Google-Fu') I recommend [Gwern's page](#) (Branwen, 2020) on the subject.

When You Don't Know The Name of Your Literature, or Missing and Biased Literatures

One of the more pernicious problems for literature review can be not knowing the name of the relevant literature. I often find myself posing research questions where it isn't clear how I would find previous work. [The inciting post](#) (Hoffman, 2018) that convinced me to write this one is discussing a phenomena that seems unlikely to be studied by economists. If I was doing literature review as part of writing this post, I would ask myself "What does the universe look like where we had the world wars and then wartime mobilization never stopped?". Then, I would aggressively dig in to find decisive places where looking at what happened before and after the world wars would prove or disprove my thesis. It's not enough to identify two points and then draw a trend line, that's not what it looks like [to thoroughly justify yourself](#) (namespace, 2020). As a thoroughly justified hypothesis looms closer and closer to theory, arguing against it should begin to feel like your debate partner is reality itself.

For the specific problem of a literature you simply don't know the name of, your best bet is often to ask others. Many times I've wanted to post a Request For Literature (RFL) on LessWrong, but felt without context the concept wouldn't really make sense to most readers. Hopefully after publishing this I'll be able to link it for context and that won't be a problem.

I didn't know what literature to look at for my essay on [Fuzzies and Saddies](#) (Zealot, 2020), where the thesis is both outside the overton window and our current social reality. How do you look at the literature for something like *that*? Well, one of the benefits of living in a [consistent universe](#) (namespace, 2020) is that it can take a lot of

effort to reliably censor all information that would point towards a real phenomena. Because our censorship is largely of the distributed kind based on social pressure, it's largely ad-hoc and doesn't hold up well against the historical record or clever inference. I took notes on how I found the book on Missionary Morale.

Example Research Session: Finding The Book On Missionary Morale

Research Question (roughly): What makes some people seem to derive satisfaction and utility from being put into hellish situations like WW1?

Immediate question: Where would I be able to find information relating to this question, where would it be recorded and how would it be framed?

Thought: "What about studies on how soldiers attitudes about war change after they've been to war? [Most soldiers will probably dislike it, but some do like it and this might be studied as a pathology]"

Search (Google Scholar): soldiers attitudes toward war

First result: Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams Jr, R. M. (1949). The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1.

Look for thing, find thing. Read through it some, then:

Observation: There's a chapter on morale, the thing I am researching, "motivation through suffering, being fueled by harrowing circumstances, asceticism, keeping spirits up in the face of a hostile universe" is very closely related to and overlaps with the study of morale. Therefore I can look at morale studies to get a better look at this subject.

Read through book's study on effect of exposure to combat on morale, realize that it doesn't seem to be very useful to me.

Principle of Pain: Why isn't this useful to me?

Answer: The thing causing the drop in morale is of the wrong structure, these studies are about exposure to short bursts of extreme stress and danger which is not the situation my audience will be encountering in their lives.

Principle of Balance: Okay then, what would be useful to me (be of similar circumstances)?

Constraint: Needs to be a population which it's likely there will be studies on.

Hypothesis: Military intelligence officers, since their job is closer to the research aspect of things while still being in a population whose morale will be studied.

Hypothesis: Spy morale, spies need to exist in a foreign place pretending to be someone they're not while their real job is to do something else which is adversarial to the people in their immediate environment. The sort of alienation and lack of belonging that causes seems like a probable fit for how it actually feels to be researching things that only you care about in your immediate environment over a long period of time while at a deep cultural gulf between yourself and the people around you.

Principle of Exhaustion: Ph.D burnout, paratrooper morale (esp. if there are cases where single paratroopers are dropped into an area and have to be on their own, snipers?), Evangelical/Missionary morale/burnout, Wilderness survival/etc morale

I go look up stuff on spy morale, forgot to take notes during this.

Observation: MICE to RASCALS talks about 'operational psychology', which might have material on agent attrition and factors relating to it.

Principle of Balance: What do counterintelligence officers do to *dissuade* potential spies?

Observation: Undercover police work involves similar stuff in a domestic context which is less secret than international espionage.

Find paper on undercover police work, read it.

Principle of Pain: This still isn't quite what I want, because the point here is to condition an officer to play a role which they then need to be pulled out of later without too much damage. Though I guess that could be relevant, it's just not the core of the thing.

Decide to move on and look at missionaries.

Search (Google Scholar): missionary morale

Read the first search result, which is a book from 1920 on literally this subject (Miller, 1920).

Bibliography

1. Garson, D., & Lillvik, C. (2012). *The literature review: A research journey*. Research guides at Harvard Library. <https://guides.library.harvard.edu/c.php?g=310271&p=2071512>
2. Dance, A. (2012). *Outsider science*. Symmetry Magazine. <https://www.symmetrymagazine.org/article/marchapril-2008/outsider-science>
3. namespace. (2020, February 1). "Memento mori", Said the confessor. The Last Rationalist. <http://thelastrationalist.com/memento-mori-said-the-confessor.html>
4. Branwen, G. (2020, May 8). *Embryo selection for intelligence*. <https://www.gwern.net/Embryo-selection>
5. Constantin, S. (2018, December 14). *Player vs. character: A two-level model of ethics*. LessWrong. <https://www.greaterwrong.com/posts/fyGEP4mrpyWEAfqj/player-vs-character-a-two-level-model-of-ethics>
6. Harford, T. (2019, August 14). *The penny post revolutionary who transformed how we send letters*. BBC News. <https://www.bbc.com/news/business-48844278>
7. Smith, N. (2017, May 15). *Vast literatures as mud moats*. Noahpinion. <https://noahpinionblog.blogspot.com/2017/05/vast-literatures-as-mud-moats.html>
8. Giuliani-Hoffman, F. (2020, March 25). *5,000-year-old sword is discovered by an archaeology student at a venetian monastery*. CNN Style. <https://www.cnn.com/style/article/5000-year-old-sword-discovered-in-italy-trnd/index.html>

9. Foddy, B. (2017). Getting over it with bennett foddy [Desktop & Mobile video game]. Humble Bundle: Bennett Foddy.
10. Benson, T. (2014, June 12). *Overview of wright brothers discoveries*. Re-Living the Wright Way. <https://wright.nasa.gov/discoveries.htm>
11. Hossenfelder, S. (2016, May 19). *The holy grail of crackpot filtering: How the arXiv decides what's science - and what's not*. Backreaction. <https://backreaction.blogspot.com/2016/05/the-holy-grail-of-crackpot-filtering.html>
12. Zealot, E. (2020, April 21). *Fuzzies and saddies part one: X-risk and motivation*. The Last Rationalist. <https://www.thelastrationalist.com/fuzzies-and-saddies-part-one-x-risk-and-motivation.html>
13. Hossenfelder, S. (2016, August 11). *What i learned as a hired consultant to autodidact physicists*. Aeon Ideas. <https://aeon.co/ideas/what-i-learned-as-a-hired-consultant-for-autodidact-physicists>
14. Raymond, E.S., & Moen, R. (2014, May 21). *How to ask questions the smart way*. <http://www.catb.org/~esr/faqs/smart-questions.html>
15. Hanson, R. (2007, July 17). *Blogging doubts*. Overcoming Bias. <http://www.overcomingbias.com/2007/07/blogging-doubts.html>
16. Binstock, A. (2012, July 10). *Interview with alan kay*. Dr Dobb's. <https://www.drdobbs.com/architecture-and-design/interview-with-alan-kay/240003442>
17. Branwen, G. (2019, January 5). *Archiving URLs*. <https://www.gwern.net/Archiving-URLs>
18. Svenonius, E. (2000). *The intellectual foundation of information organization*. The MIT Press.
19. Hastings, C. (2009, September 25). *Get lit: The literature review*. YouTube. <https://www.youtube.com/watch?v=9la5ytz9MmM>
20. Mongan-Rallis, H. (2018, April 19). *Guidelines for writing a literature review*. <https://www.d.umn.edu/~hrallis/guides/researching/litreview.html>
21. Branwen, G. (2020, January 21). *Internet search tips*. <https://www.gwern.net/Search>
22. Hoffman, B.R. (2018, May 23). *There is a war*. LessWrong. <https://www.greaterwrong.com/posts/DtS6x5r54sEx7e2tP/there-is-a-war>
23. namespace. (2020, March 30). *Necessity and warrant*. The Last Rationalist. <https://www.thelastrationalist.com/necessity-and-warrant.html>
24. namespace. (2020, March 23). *On necessity*. The Last Rationalist. <https://www.thelastrationalist.com/on-necessity.html>
25. Miller, G.A. (1920). *Missionary morale*. Google Books (orig. New York, Cincinnati: The Methodist Book Concern).

Comment on "Endogenous Epistemic Factionalization"

In "[Endogenous Epistemic Factionalization](#)" (due in a forthcoming issue of the philosophy-of-science journal *Synthese*), James Owen Weatherall and Cailin O'Connor propose a possible answer to the question of why people form factions that disagree on multiple subjects.

The existence of persistent disagreements is [already kind of a puzzle](#) from a Bayesian perspective. [There's only one](#) reality. If everyone is honestly trying to get the right answer and we can all *talk* to each other, then we should converge on the right answer (or an answer that is [less wrong](#) given the evidence we have). The fact that we *can't do it* is, or should be, an embarrassment to our species. And the existence of *correlated* persistent disagreements—when not only do I say "top" when you say "bottom" even after we've gone over all the arguments for whether it is in fact the case that top or bottom, but *furthermore*, the fact that I said "top" lets you *predict* that I'll probably say "cold" rather than "hot" even *before* we go over the arguments for that, is an *atrocity*. (Not hyperbole. Thousands of people are dying horrible suffocation deaths because we can't figure out the optimal response to a new kind of coronavirus.)

Correlations between beliefs are often attributed to ideology or [tribalism](#): if I believe that Markets Are the Answer, I'm likely to propose Market-based solutions to all sorts of seemingly-unrelated social problems, and if I'm [loyal to the Green tribe](#), I'm likely to [selectively censor my thoughts in order to fit the Green party line](#). But ideology can't explain correlated disagreements on unrelated topics that the content of the ideology is silent on, and tribalism can't explain correlated disagreements on narrow, technical topics that aren't [tribal shibboleths](#).

In this paper, Weatherall and O'Connor exhibit a toy model that proposes a simple mechanism that can explain correlated disagreement: if agents disbelieve in evidence presented by those with sufficiently dissimilar beliefs, factions emerge, even though everyone is honestly reporting their observations and updating on what they are told (to the extent that they believe it). The paper didn't seem to provide source code for the simulations it describes, so I followed along in Python. (Replication!)

In each round of the model, our little Bayesian agents [choose between repeatedly performing](#) one of two actions, A or B, that can "succeed" or "fail." A is a fair coin: it succeeds exactly half the time. *As far as our agents know*, B is either slightly better or slightly worse: the per-action probability of success is either $0.5 + \varepsilon$ or $0.5 - \varepsilon$, for some ε (a parameter to the simulation). But secretly, we the simulation authors know that B is better.

```
import random

ε = 0.01

def b():
    return random.random() < 0.5 + ε
```

The agents start out with a uniformly random probability that B is better. The ones who currently believe that A is better, repeatedly do A (and don't learn anything, because they already know that A is exactly a coinflip). The ones who currently believe that B is

better, repeatedly do B, but keep track of and publish their results in order to help everyone figure out whether B is slightly better or slightly worse than a coinflip.

```
class Agent:
    ...
    def experiment(self):
        results = [b() for _ in range(self.trial_count)]
        return results
```

If H_+ represents the hypothesis that B is better than A, and H_- represents the hypothesis that B is worse, then Bayes's theorem says

$$P(H_+|E) = \frac{P(E|H_+)P(H_+)}{P(E|H_+)P(H_+) + P(E|H_-)P(H_-)}$$

where E is the record of how many successes we got in how many times we tried action B. The likelihoods $P(E|H_+)$ and $P(E|H_-)$ can be calculated from the probability mass function of the [binomial distribution](#), so the agents have all the information they need to update their beliefs based on experiments with B.

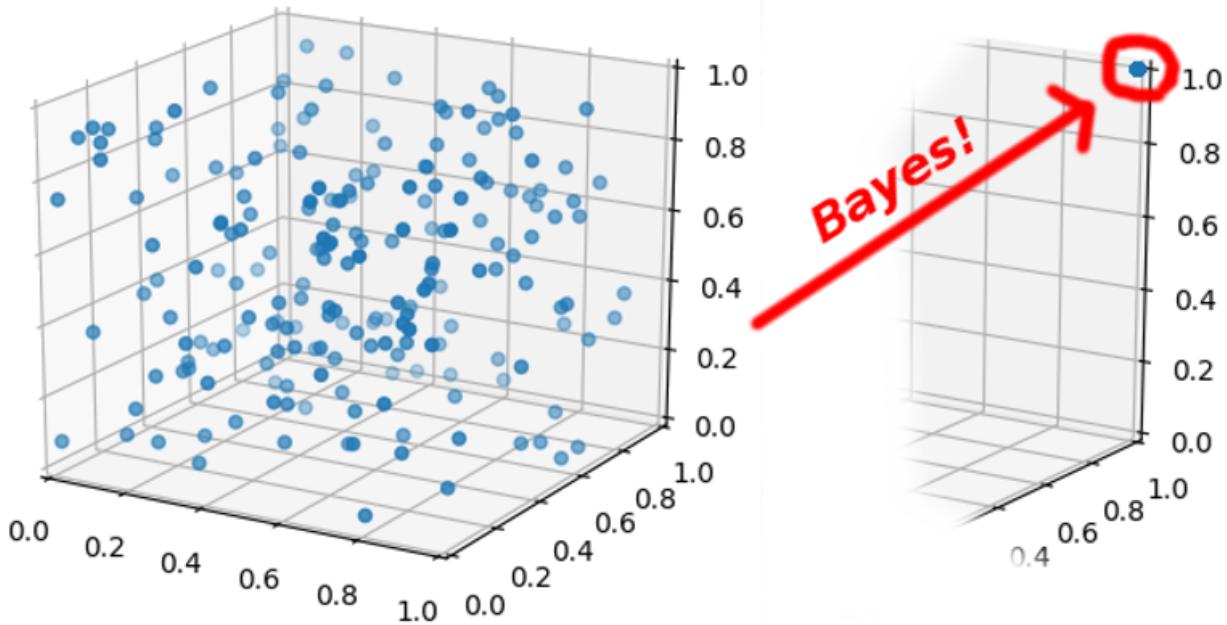
```
from math import factorial

def binomial(p, n, k):
    return (
        factorial(n) / (factorial(k) * factorial(n - k)) *
        p**k * (1 - p)**(n - k)
    )

class Agent:
    ...
    def pure_update(self, credence, hits, trials):
        raw_posterior_good = binomial(0.5 + ε, trials, hits) * credence
        raw_posterior_bad = binomial(0.5 - ε, trials, hits) * (1 - credence)
        normalizing_factor = raw_posterior_good + raw_posterior_bad
        return raw_posterior_good / normalizing_factor
```

Except in order to study the emergence of clustering among multiple beliefs, we should actually have our agents face *multiple* "A or B" dilemmas, representing beliefs about unrelated questions. (In each case, B will again be better, but the agents don't start out knowing that.) I chose three questions/beliefs, because that's all I can fit in a pretty 3D scatterplot.

If all the agents update on the experimental results published by the agents who do B, they quickly learn that B is better for all three questions. If we make a pretty 3D scatterplot where [each dimension represents](#) the probability that B is better for one of the dilemmas, then the points converge over time to the [1.0, 1.0, 1.0] "corner of Truth", even though they started out uniformly distributed all over the space.



But suppose the agents don't trust each other's reports. ("Sure, she says she performed B_2 50 times and observed 26 successes, but she *also* believes that B_1 is better than A_1 , which is *crazy*. Are we sure she didn't just make up those 50 trials of B_2 ?"") Specifically, our agents assign a probability that a report is made-up (and therefore should not be updated on) in proportion to their distance from the reporter in our three-dimensional beliefspace, and a "mistrust factor" (a parameter to the simulation).

```
from math import sqrt

def euclidean_distance(v, w):
    return sqrt(sum((v[i] - w[i]) ** 2 for i in range(len(v))))

class Agent:
    ...
    def discount_factor(self, reporter_credences):
        return min(
            1, self.mistrust * euclidean_distance(self.credences, reporter_credences)
        )
    def update(self, question, hits, trials, reporter_credences):
        discount = self.discount_factor(reporter_credences)
        posterior = self.pure_update(self.credences[question], hits, trials)
        self.credences[question] = (
            discount * self.credences[question] + (1 - discount) * posterior
        )
```

(Um, the paper itself actually uses a slightly more complicated mistrust calculation that also takes into account the agent's prior probability of the evidence, but I didn't quite understand the motivation for that, so I'm going with my version. I don't think the grand moral is affected.)

Then we can simulate what happens if the distrustful agents do many rounds of experiments and talk to each other—

```

def summarize_experiment(results):
    return (len([r for r in results if r]), len(results))

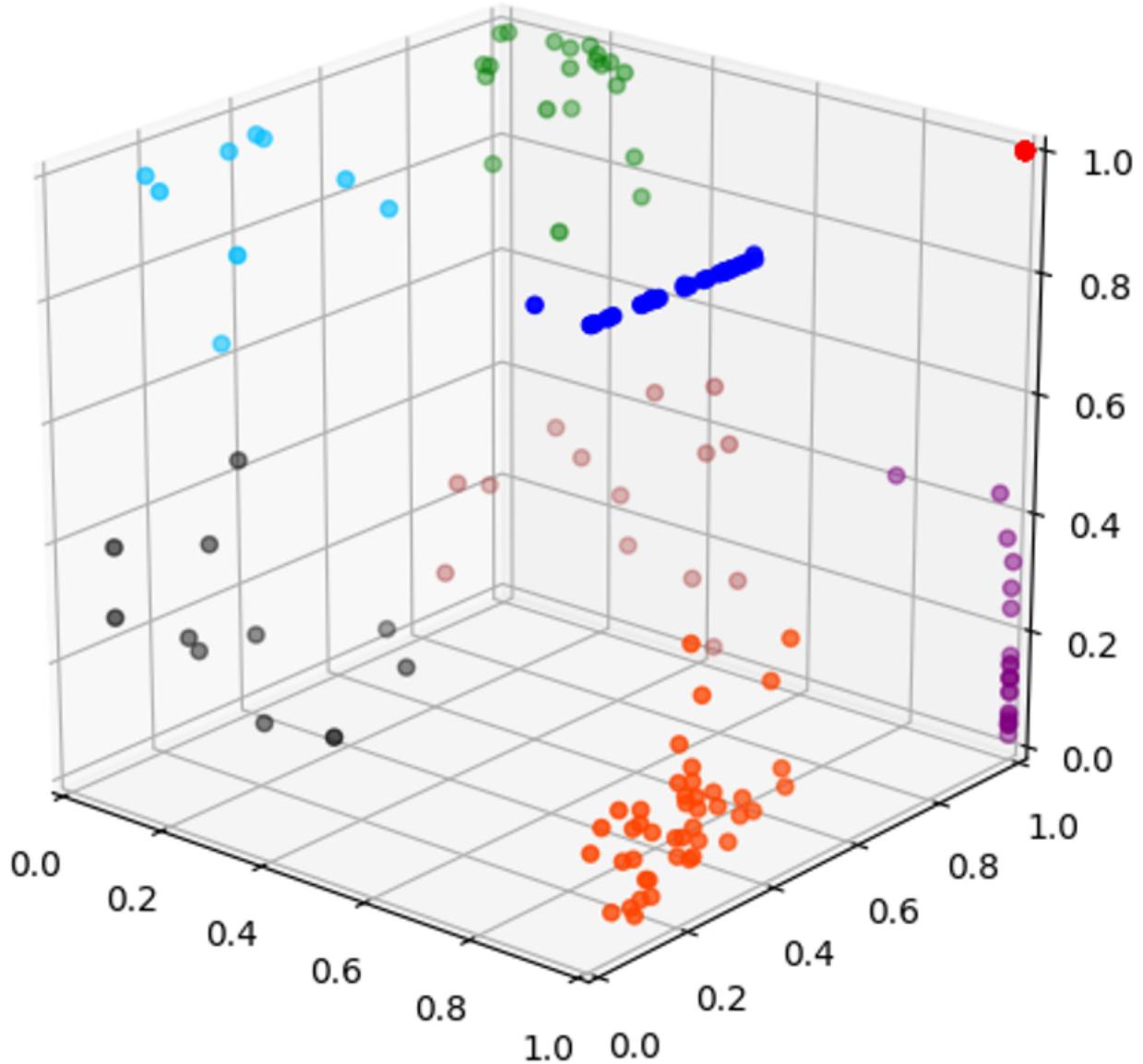
def simulation(
    agent_count, # number of agents
    question_count, # number of questions
    round_count, # number of rounds
    trial_count, # number of trials per round
    mistrust, # mistrust factor
):
    agents = [
        Agent(
            [random.random() for _ in range(question_count)],
            trial_count=trial_count,
            mistrust=mistrust,
        )
        for i in range(agent_count)
    ]

    for _ in range(round_count):
        for question in range(question_count):
            experiments = []
            for agent in agents:
                if agent.credences[question] >= 0.5:
                    experiments.append(
                        (summarize_experiment(agent.experiment()), agent.credences)
                    )
            for agent in agents:
                for experiment, reporter_credences in experiments:
                    hits, trials = experiment
                    agent.update(
                        question,
                        hits,
                        trials,
                        reporter_credences,
                    )

    return agents

```

Depending on the exact parameters, we're likely to get a result that "looks like" this
`agent_count=200, round_count=20, question_count=3, trial_count=50, mistrust=2` run—



Some of the agents (depicted in red) have successfully converged on the corner of Truth, but the others have polarized into factions that are all wrong about *something*. (The colors in the pretty 3D scatterplot are a [k-means clustering](#) for $k := 8$.) On average, evidence pushes our agents towards Truth—note the linearity of the blue and purple points, illustrating convergence on two out of the three problems—but agents who erroneously believe that A is better (due to some combination of a bad initial credence and unlucky experimental results that failed to reveal B's ϵ "edge" in the sample size allotted) can end up too far away to trust those who are gathering evidence for, and correctly converging on, the superiority of B.

Our authors wrap up:

[T]his result is especially notable because there is something reasonable about ignoring evidence generated by those you do not trust—particularly if you do not trust them on account of their past epistemic failures. It would be irresponsible for scientists to update on evidence produced by known quacks. And furthermore, there is something reasonable about deciding who is trustworthy by looking at their

beliefs. From my point of view, someone who has regularly come to hold beliefs that diverge from mine looks like an unreliable source of information. In other words, the updating strategy used by our agents is defensible. But, when used on the community level, it seriously undermines the accuracy of beliefs.

I think the moral here is slightly off. The *specific* something reasonable about ignoring evidence generated by those you do not trust on account of their beliefs, is the assumption that those who have beliefs you disagree with are following [a process that produces systematically misleading evidence](#). In this model, that assumption is just *wrong*. The problem isn't that the updating strategy used by our agents is individually "defensible" (what does that mean?) but produces inaccuracy "when used on the community level" (what does that mean?); the problem is that you get the wrong answer if your degree of trust doesn't match agents' actual trustworthiness. Still, it's enlighteningly disturbing to see specifically how the "distrust those who disagree" heuristic descends into the madness of factions.

([Full source code.](#))

Movable Housing for Scalable Cities

First posted [on Steemit](#) on July 30, 2016.

0: Summary.

The non-scalability of current major cities poses a challenge to economic growth; it is a burden that falls disproportionately on those least able to bear it.

Unfortunately, any would-be startup city must confront the network effects that pose a tremendous potential energy barrier to the attractiveness of a city that isn't yet mature. From day one, you need to compete with the hedonic attractiveness of San Francisco, in the eyes of the person who decides where to locate the company.

Movable houses mated to modular foundations – nice large customized houses, not small ugly trailers – can make cities more scalable *and* more hedonically attractive:

- By making it easier for groups to relocate and untangle themselves in a coordinated way, movable houses can increase the natural dispersion of the city as it grows larger.
- Since people could buy and retain customized houses manufactured with economies of scale, movable houses could have better technology and amenities not available today. (Contrast your current house to a modern car.)
- Movable houses could allow unprecedented opportunities to live next to your friends, or to groups of similar-minded people. You and your friends just need to find a set of available modular foundations close together.
- Movable housing might help on key points of governance and realpolitik, for example by making land value taxes more attractive, and decreasing exit costs.

(This was written in response to Y Combinator's [request for ideas](#) about the design of new cities.)

1: Introduction: The challenge of scalable cities.

Manhattan doesn't scale well. San Francisco doesn't scale well. Tokyo and London don't scale well. On the present system, each added resident comes with some tiny added degree in inconvenience for every other resident already living in a city, most notably in the form of traffic jams and housing prices. Yes, every added programmer also provides some tax revenue; but the city infrastructure, qua infrastructure, is not scaling well.

This is not *all* due to anti-housing housing policy, though we've certainly gone to every possible length to shoot ourselves in that foot. Limited road bandwidth is a real issue. The air in Manhattan is genuinely dirtier than in Berkeley – I acquire a sore throat every time I visit New York City.

Building cities that can scale better is a problem of overwhelming human importance. Some analyses have suggested that increased housing costs have eaten huge chunks out of middle-class income; that housing costs may be responsible all on their own for the stagnating middle class. The costs and inconveniences of our most active and growing cities, prevent people in less active economic areas from seeking better fortunes. Cities that don't scale are problems measured in trillions of dollars of lost economic growth and a crushing amount of real human despair.

1.1: All-robotic car fleets are an obvious first step.

The technology for an all-robotic car fleet is probably already available... if we take all the non-robotic cars off the road.

A robotic car fleet makes cities more scalable in many, many ways. The most obvious boost is widening space, compressing time, and removing the cost in stamina to travel. Robotic cars that don't need stoplights or speed limits might let you get anywhere within 6 miles in 10 minutes. (Say, 2 minutes to get to a thoroughfare, 6 minutes to go 6 miles at 60mph, and 2 minutes to go the last mile.) A 6-mile radius is a *lot* of territory (113 square miles). You can be 10 minutes away from a *lot* of friends.

Robotic cars lend themselves to fleets that show up when needed, rather than individually owned cars. If every individual car is being used much more regularly, that changes the cost/benefit calculation for batteries versus gasoline. Which implies less pollution per added resident in concentrated populations. Fewer parking lots implies more space for people, and so on.

But robotic cars, by themselves, may not be enough. Designing a new city is the *last* place you want to stop and congratulate yourself on the improvements you already have planned.

1.2: The first business challenge is being hedonically attractive to corporate decisionmakers.

Cities run on network effects. Moving to a location with 1000 companies that might want to hire you is far less risky to moving to a location with 3 companies that might want to hire you.

Any new city faces a *huge* potential energy gradient during its youngest and most vulnerable stages. The gravity of Manhattan and Tokyo and London is absolutely enormous, drawing away your potential residents. You need to offer something *extremely* attractive about your new city if you want companies to move there instead of San Francisco. And your city's first critical business challenge *is* about attracting companies, not humans; humans go where there are jobs.

"But we'll have lower housing prices!" you cry. Alas, it's a sad fact of life that the people who decide where to locate their companies face different incentives than the average employee in those countries. (As in the Dilbert cartoon where the CEO is moving the company to what happens to be his old hometown, because that way he

can get free babysitting from his grandparents.) I once asked a Google employee why Google didn't start an offshoot campus somewhere with lower housing costs – surely, I ventured, Google had the scale to do that, and enough employees who dispreferred city living. My friend replied that no project manager at Google would want to be so far from the center of corporate politics.

The cofounder who decides to locate their hot startup in San Francisco may have plenty of reasonable and selfless motives to do so – better access to venture capital, better access to programmers. But it seems worth noticing regardless that this decision is being made by someone who can probably afford more Ubers and higher rents than other company employees. If the company did locate itself somewhere cheaper, it could perhaps pay lower salaries and so capture *part* of the gains from lower housing costs. The decisionmaker who decides where to locate the company may capture a *part* of the company's gains from those lower salaries. But quantitatively speaking, those fractions do not multiply out to 100%. The incentives are not aligned between the decisionmaker and the rank-and-file.

The overwhelming proportion of the *real human good* accomplished by a more scalable city comes in the form of schoolteachers who can afford the rent. But if you want that city to boot up successfully, you need to think about how to make it *personally, hedonically* attractive to the decisionmakers who choose where to put their companies.

Robotic cars help here too! Even compared to taking a private car through traffic, taking a robotic car through an absence of traffic might be more attractive to a CEO.

Slap some flying drones riding the cars, to take off and make deliveries as the destination approaches, and you can have amenities that non-robotized cities just don't have! If delivering hot meals across a 3-mile radius became truly cheap and ubiquitous, so that even non-CEOs could afford it, then we would see much greater specialization on selling meals. Meal-creation businesses would compete on cost and healthiness, instead of restaurants competing on location and dining experiences. It could become typical to order a not-so-expensive, non-high-calorie, home-cooked-style hot meal that would have taken you an hour to make from scratch, and have it delivered in 10 minutes. Maybe the CEO of a *big* company has a home chef who can compete with that – but it's an attractive amenity for the CEO of any middle-sized company or hot startup.

Again, the vast majority of real good comes from giving the thousands of *other* residents access to healthy food without a huge cost in personal labor. But you have to appeal to the decisionmakers before anyone else gets that chance.

Having remarked on this cool and futuristic amenity, we are still in no position to relax and declare ourselves done. We need every possible attractive feature for our shiny new city to exert a draw that is remotely competitive with Manhattan.

1.3: Inside or outside the US?

I will digress at this point – though it will turn out to not really be a digression – to consider the possibility of locating the New City outside the United States (and outside the UK and the EU and Australia).

By far the hugest attraction of this possibility for cofounders would be EASIER IMMIGRATION. If your city is otherwise competitive in terms of first world amenities and safety, has lower housing prices, and cofounders can bring people there without Eternal Visa Hell, startups will flock to you that *couldn't* form anywhere else.

Avoiding regulatory molasses will be another huge attraction for some companies and potential companies. Cough cough PATENT MONSTERS cough cough.

If the new city were located in, say, a special district negotiated with the government of Uruguay, this might permit some basic utopian legal features. For example, not living in a country where marijuana is theoretically illegal if you're white and rich, and potentially life-destroyingly illegal if you're black or poor. For many cofounders or company-owners, this is a point of principle rather than practice; but to some people it is a matter of practice. Or some people, with decision-making power even, may dislike at a deep emotional level the prospect of living in a country where a SWAT team can kick down their door at any time.

The United States is unique among nations in insisting that it can tax the income of its citizens even if they live and earn outside the US. But the tax policy surrounding your city still matters to anyone who isn't a US citizen, or to any US citizen earning not much more than \$92,000 per year.

Act 20/22 in Puerto Rico enables people moving to Puerto Rico for the first time to negotiate a two-decade contractual exemption from US *federal* income taxes. Puerto Rico is just about the only place on Earth, inside or outside the US, that a US citizen can go to not pay US federal taxes. For this reason I'm fond of saying that the three obvious places to start a new city are 2 hours from the Bay Area, Puerto Rico if you're going to be located in the US but not near any particular existing metropolitan area, and Uruguay (hurricane and earthquake free, extremely sensible and stable government). Not paying federal taxes in Puerto Rico would be a hugely favorable attractive gradient – for the decisionmakers that locate companies, in particular, but heck, even for programmers making a dinky \$150K per year.

International development has its own negatives. Most obviously, the extra barrier posed by being a *lot* further away from existing centers of gravity; compared to starting up 2 hours from the Bay Area, or in coastal Oregon, or Nevada. It's one thing to live a 2-hour drive away from venture capitalists, and quite another to be separated from them by a 6-hour plane flight.

There may also be language issues – especially if your company needs to interact with government agencies whose employees don't speak your language. Google Translate will only continue to improve in the future, lowering language barriers further; but the technology isn't there *right now* to render international communication frictionless.

These barriers will be particularly fatal during the early stages, when the potential energy barriers surrounding your city are the highest.

What we *really* want – one might think – is to *initially* start up our city 2 hours from the Bay Area. But to have some mysterious feature that made it *unusually easy* for many of the city's inhabitants and whole companies to relocate to coastal Oregon, to Puerto Rico, or to Uruguay, once the system had been proven.

Especially if this mysterious new feature also improved the hedonic attractiveness, the cost profile, and the fundamental scalability of our startup city...

2: Movable housing.

Let's assume at the beginning that we're talking about an early town small enough to be composed mainly of houses instead of apartment buildings.

Imagine that in this town, houses are portable objects with standard connections that match up to standard modular house-foundations. We can disconnect a house from its water and electricity and Internet cables, have a standard vehicle pick the house up from its foundations, and gently drive the house over to somewhere else inside the city.

We could also imagine apartments that slot into towers on rails. We can imagine office buildings built up from modules that could be taken apart at need. But as we'll soon see, this new city might be able to scale a lot further before it *needed* apartment buildings and office towers.

Movable housing would require some amount of new technology, and more importantly, a green-field city – the roads need to take the weight (unless we can replace or supplement carriers with skyhooks or other lifters); the robotic-vehicle idiom would also help, because it means we can clear any road of cars as a house-carrier is passing. I'm pretty sure all of this is within reach of human technology; the question is whether it's too expensive.

And whether anything is 'too expensive', of course, depends on what benefits you're buying.

Benefit #1: You can live closer to where you most want to live.

In the world today, when you want to move to a new city or district, you hunt around until you find a house that is shaped *sorta* like you want a house to be shaped, which isn't located *too* far from where you wish you actually lived. And all of your friends and coworkers – network effects will shortly become very important here – are implementing that same process.

Movable housing decouples the problem of finding a good *location* from the problem of finding a good *house*. It disaggregates the two businesses, [as my father would say](#).

To find a nice place to live, you just need to find any available plot whose modular foundations match your current house. (There'd better not be more than three varieties of foundation, though, or we're back to the same old matching hell.)

Benefit #2: You can own a better house.

A modern car contains vastly more interesting technology than a modern house. That's because there are big-company car manufacturers with centralized factories and R&D departments competing to offer the car that you'll like the best.

Imagine if instead, a boutique car-constructor company sauntered over to a parking space and built a car there over a year or so. And then you looked around for an

available parking space that wasn't too far from your house and had an associated car that was acceptable. If no car was good enough, you could buy a parking space and its car, trash that car, and have a different car handcrafted in that parking space over the next year by the small boutique car company of your choice.

Cars would be a lot less advanced, to put it mildly.

You wouldn't realize what you were missing. You'd never have seen cars with electronic stability control and automatic parallel parking and collision alarms and radio keyfobs. Maybe some rich people would have hand-constructed cars with primitive versions of those features, to go with the gold-plated ashtrays.

So if you can buy a house from a company that has a real R&D department and is actually trying to please the customer and has real competition... well, I'm not an entire R&D department myself. But my personal starting list of feature requests for an Apple iHouse might include:

- Completely blackout shutters over my windows, which open gradually and automatically at the time of morning when I actually wake up.
- Noiseproofing. Extremely serious noiseproofing, with thicker walls, better-sealed doors and windows, active sound cancellation, and maybe white noise generation if that still wasn't enough. When I close my door I want *silence*.
- Centralized, extremely powerful LED lighting modules with the light carried by optical fibers to whichever room I was actually using at the time. Dimming to red as my bedtime approached, of course.
- Centralized humidification with per-room control plus per-room temperature control, so I'm not running a loud hot humidifier in my bedroom at night. (I need a lot of humidity, personally.)
- A Jacuzzi on the roof, open to the stars. (You think this is some kind of huge luxury and that I'm spoiled for even wanting it, but that's because you live in a civilization where everything house-related costs too much.)
- Anti-insect laser bug-zappers on the roof which fry any mosquitos, wasps, hornets, other stingy things, or even houseflies that get too close. (Yes, I know we're all supposed to be too tough to be bothered by little things like insects; just like there was a time when British people made fun of people who carried umbrellas on drizzly days. I observe that you do, in fact, work inside an office instead of outdoors. I bet those little annoyances you're supposed to be too tough to care about have an awful lot to do with that cultural decision in practice.)

Finally, bigger rooms! More storage space! Relatively empty square meters in a house just should not cost that much on the margin!

Extra marginal people impose costs on cities. Electricity costs money. The central LED lighting modules and Powerwall batteries will cost money. An extra square meter of floor or ceiling, even an extra story on a house, should not cost that much more on the margins.

It does feel to me like the *small, crowded* apartments are an arcane penance that we are imposing on ourselves. Even people who don't like expanding cities, ought to see more people as being the issue, not more square meters of living space; having more square meters per person seems like something our civilization should be able to manage. Our ancestors lived in bigger houses. We really ought to be able to get that back.

(This does mean tolerating higher roofs, or larger apartment-tower frames for modular apartments on rails, or less grass on the hill. Or living further away with lower local population density; see Benefit #4.)

Some of us have chosen to live in large house-complexes with friends (though it's not obvious how much of this is just due to modern housing costs). Imagine a central 'house' that has the kitchen and the game room and the Jacuzzi, and six modules radiating out from it, each module with a bedroom, bathroom, storage space, and small reading room. Want to try out a different group house for a month? Pick up the bedroom module, connect it to a different house, put it back down again.

Your own wishlist would look different, I expect. And in another 5 years the Apple iHouse 4e would offer us features that neither of us imagined. But one point is sure: under the present system of the world, we'll never get a chance to find out.

Benefit #3: Friends living near each other.

All sufficiently tight clusters of friends can find a set of empty foundations, rent them, and live literally next door to each other.

We could have neighbors again.

We could have *tribes* again.

Yes yes, I know, you're thinking of that one person who's friends with a bunch of your friends but who you don't want on the same block as you. But really, human beings have been living in tribes for a *while*. You can probably work it out. Many people could benefit a *lot* from living in close proximity to fifty people whose company you can tolerate – not necessarily your coworkers, even; you might be choosing to live next to your fellow D&D 3.5e roleplayers who also like dubstep.

It could give us back something human that we lost, and start to make headway against lives that are *sad*, never mind the poverty.

This notion will admittedly create some coordination problems: "Does everybody in the block need to be in our group for us to have the kind of... parties... that we want?" and "How do we try to make sure that a lot of foundations go empty at the same time so a new group can move in?" and "Uh, what if 60% of the group prefers somebody *not* move in to their cluster, is that going to be enforceable at all and doesn't that imply a whole new level of local government?"

I can think of some obvious policies to try, but I have a strong suspicion that policy v.1 won't work all that well and it will take until policy v.3 before work tolerably well. Sometimes you just gotta try things.

Benefit #4: Small towns can scale better.

And not just to larger sizes, but to higher levels of economic activity.

This is the prediction that I'm least sure about. But one of the reasons why I'm talking about movable *houses*, instead of modules on rails in a huge apartment-tower frame... is that maybe, a lot of us won't even need the huge apartment-frames?

Maybe you, dear reader, positively prefer to live in a huge apartment building. Maybe you don't care where a robotic car can take you in 5 minutes. Maybe what you really want, more than any other style of life, is to be able to walk out onto your sidewalk and see lots and lots of human faces (even unfamiliar ones), then turn right and walk into that lovely coffeeshop that is downstairs and half a block to the right of your eighth-floor apartment. Maybe you like the hustle and the bustle. There is no arguing with terminal components of the utility function, as David Hume observed.

Some of us – and I'm not saying we have better preferences, but we have *different* preferences – some of us would rather live in Rivendell.

Or failing that, we'd like to live in a quiet little house on a green hill, where the technologically advanced soundproofing doesn't need to work that hard.

If we had robotic cars *and* movable houses, it would be a lot easier for all of us who work at my nonprofit to *not* need to live downtown – not even in downtown Berkeley. Part of the reason we located in downtown Berkeley is that it has sufficient density of housing that we originally could, and new employees still can, find available non-horrible apartments within walking distance of the office.

But in a city with robotic cars *and* movable houses, we could perhaps all live in the same section of green hills, 10 miles or 20 miles from the skyscrapers, and move our office modules there too. Robotic cars could teleport us to the large, centralized supermarket that served a big area and therefore had just as much selection as a supermarket in a big city. Or some of us might live closer to the skyscrapers, or with the Burning Man tribe; and for those who made that choice, a robotic car would take them to the green hilly workplace in 24 minutes while they browsed their cellphones.

24 minutes isn't far from the time it takes to cross San Francisco right now! Anyone who considers that good enough could live anywhere within a 20-mile circle, served by robotic cars with no traffic lights cruising at 60mph down the central throughways.

Even so, some offices and some people would need to be closer to the center of gravity, or would just yield in preference to the siren call of network effects. Sometimes there are just too many graph links that all need to be located within 5 minutes of one another. Those companies might still need to live in office modules slotted into towering office-complex-frames at the New City's center.

But if you can coordinate locations more cheaply, move more cheaply, and travel more cheaply because of the robotic cars, then it becomes a whole lot more feasible to have *more* software companies located in the quiet hills. It's not a panacea and it won't work for every organization, but *more* people will be able to live in real houses like our parents owned.

Even more of us will be able to live in Rivendell when virtual reality tech matures, which will loosen (though not cut) the bonds of spatial distance that much further. Or VR might not really make any pragmatic difference, but it's worth trying to think 5 years ahead at least; the headsets *will* improve. It will be one more marginal force exerting a bit more quantitative push towards the attractiveness of living in a bigger house in a quieter place, if your new city can offer that with fewer than usual disadvantages.

Benefit #5: Land value taxes.

Economists since Adam Smith have observed that land is the ideal thing to tax. Literally, tax the square meters of planetary surface. We can't make more land, so it's not like a land tax discourages the production of land, the way that income taxes discourage work. *Somebody* is going to collect the implicit rent on land; so long as the relevant collector doesn't tax more than the price point at which the supply of land balances the demand for land, the tax doesn't change that price. It's a frictionless tax on a **HUGE** flow of land rent, and the *alternative* to taxing that huge flow of land rent, is frictionful taxes. Taxing almost anything other than land (except carbon dioxide emissions), *before* taxing land, is one of those insanely stupid aspects of modern civilization that make economists want to stab themselves to death with a butter knife.

With movable housing, the houses are moving around and the foundations are staying in one place. This makes it even more obvious that you might as well have the rent on the foundations supporting government services.

I pay \$2500/month on my little 2-bedroom apartment in Berkeley, the supermajority of which is certainly land rent. That land rent is probably more than I pay in state and federal income taxes. Which makes me want to gouge out my eyes with a spoon. Because – no offense to my innocent landlord who paid a corresponding price for that land and is not earning an excess profit in on it after mortgage costs – I am paying *two* huge taxes where an even slightly saner system could be charging me *one* tax.

Yes, you can see how it *might* go wrong if one government tries to own all the modular foundations and therefore all the urban land. (The more so if there are no competing cities: see benefit #6.) But the economic factors here are huge enough that it seems worth trying to do things the sane way just once.

Since most taxes are federal, making real headway on eliminating the 'double tax' (income tax plus land-rent) might require locating in Puerto Rico or Uruguay. But even without that, to whatever extent the modular foundations have a natural equilibrium rent, that rent can provide for fire trucks, maybe even health care if the land rents equilibrate high enough – without *additional* taxes. Where, again, the current model is to have the residents pay one stream of highly frictional tax-rent to the government, and an entirely separate stream of land-rent.

This isn't really a technology problem. But having houses moving around, so that the rent collected on the stolid immobile foundations is entirely separate from any handcrafted structure nailed to them, makes the solution that much more obvious and maybe more politically feasible.

Benefit #6: Exit threats and political relocations.

When Patri Friedman proposed 'seasteading', a lot of the hoped-for good systemic properties came from the fact that sea-based platforms would be easier to move around. Movable housing can be seen as trying to get several similar systemic properties, without taking on the engineering or political challenges of building at sea.

This analogy extends over to one of the primary political ideas motivating seasteading, the notion of "voice and exit".

When your [BATNA](#) to staying in a place is to pick up your houses and/or your office modules and move them somewhere else... then that creates a different negotiating position than when you need to destroy your painstakingly handcrafted structures, create new handcrafted structures somewhere else, rearrange all of your personal possessions, etcetera.

Yes, there's still friction – you have friends who may not follow you, maybe your kids end up going to a different school. But there's *less* friction with movable houses, the coordination problems are that much easier to solve in groups; it could make a quantitative difference.

This improved BATNA could apply at the level of a whole city that negotiated a special economic zone with some state or country. Maybe it's wacky to think that "Kay thanks bye" would ever be a plausible reply on that scale if the host country tries to "alter the bargain, pray I do not alter it any further". There might just be too many non-movable objects creating too much inertia. But even having a large *fraction* of the potential victims, having the option of putting their movable modules on a cargo ship and heading elsewhere, could make a difference. It matters quantitatively to a victimizer whether making conditions worse for their victims means that 20% of their victims leave or 2% of their victims leave.

I'm trying not to go on too much of a rant here. But one of the enormous overlooked questions of the modern age is how poverty still manages to exist, when agricultural and economic productivity have risen by a factor of literally 100 since the time when 98% of the population was farming. We have *fewer* poor people, to be sure, the life of the lowest income quartile is a *lot less* horrible than it was in the 13th century. But there's still some sense in which it seems a little *embarrassing* to imagine going back to a world where people managed to survive despite being 100 times less productive, telling them that we are now 100 times wealthier, and then having to explain why there are any horribly poor and desperate people in our country *at all*.

When a condition is that sticky, we should suspect it to be an equilibrium with strong restoring forces. There must be some powerful factor that makes some people be poor, no matter how much wealth is flowing around – a factor that gets stronger as more wealth flows, even by a factor of 100.

One of the obvious forces that could be stabilizing a Poverty Equilibrium is if the standard state of affairs, for human civilizations in general, is for there to always be a few groups here or there that can extract a little more value. The Ferguson Police Department, issuing 3 warrants per household per year, is one obvious example of this idiom. But you should also be thinking of taxi medallions, licensed haircutters, NIMBY house-owners, and health insurance companies without much statewide competition. I don't mean to single out one group as a target for the Two Minutes Hate. There can just be these endless small sets of local factors with the power to drain one more dollar; and these factors will collectively go on draining one more dollar until they can't drain any *more* dollars without some victims dying. Actually, the equilibrium for multiple extractive forces is a [commons problem](#) – Alice knows that if she doesn't steal a dollar from your pocket, Bob will steal it instead, so Alice might as well steal that dollar even if the result is disastrous. Which means that in many cases the little extractions *do* continue past the point where people riot.

This is one reason I'm skeptical of the ability of a Guaranteed Basic Income to solve poverty in general, leaving aside various other technical problems. We increased economic productivity by a factor of 100 and there are still poor people. Is a GBI really

going to be the last marginal improvement that solves it all? A GBI might still *help* – just like increasing economic productivity by a factor of 100 *helped* the people who are still living lives of awful suffering and desperation. But after you introduce a GBI, I'm guessing, there will be a number of factors that start to extract one more dollar here, one more dollar there. The Ferguson police department issues another arrest warrant per household, the state increases its court costs, hey, people can afford it now, they've got a GBI right? And what do you know, almost everyone will still have to get awful jobs just to survive.

So it's not at all a side issue, or a mere bugaboo of the independent-minded, to think about the political power of a cheaper exit. To consider whether mature VR, and to a lesser extent, movable housing, might make it a little bit harder to extract value from victims anchored too solidly to run away. The mobility of labor might affect how fast the poverty equilibrium restores itself.

I'm not saying that corporate taxes are the correct level of organization on which to have any tax at all... but it does happen to be the case that taxing corporate profits located in your country is very hard to do, at least to large corporations, because they just locate their profits somewhere else. Making individual human beings and small companies more mobile would grant them some of the same power of resistance.

No, let me be more blunt. If your shiny new city would otherwise be generating a huge amount of excess value for the people inside the new city, and the people inside the city have no credible threat of exit, the people inside your city will not be allowed to keep that value. There are things in the ecology that like to eat free energy, and your city will not be allowed to keep that energy indefinitely if it is so temptingly available for a little more taking every year. It could be eaten by any level of regional government, or any organization empowered by any level of regional government. If you're dumb enough to let somebody patent the connection scheme of the modular foundations, they can let you build out the city, watch to see how much excess wealth is being generated, and then jack up fees to try to capture nearly all of that value. It could be an invasion of patent monsters under a national jurisdiction that permits them. It only takes one factor that can threaten to shut down your whole process, to extract nearly all of the free energy from your city.

So if your well-meaning goal is to make your new city generate lots of excess wealth that the people inside the city get to keep, you'd better make it as cheap as possible for coordinated groups to leave.

Benefit #7: Lower-frictional flow of people through economies.

I'll finish by remarking that, in a very generic and boring sense, friction is usually bad for economies and reduced friction is usually good. It is a terrifying sign of stagnation that people in the United States, especially the less wealthy states, are moving house less often. Housing prices are probably a bigger part of that problem than any one-time cost of shuffling possessions. But every little bit helps, and if you reduce the physical cost and emotional wear of moving from Point A to Point B, it will matter *some*. Alleged benefits #3 to #6 mostly reduce to, "Some nice things might happen if we reduce the cost and friction of being somewhere else". So benefit #7 is just the generic observation that movable houses would reduce economic friction in a very generic way and therefore other nice things might happen as well.

We could also see movable housing as a kind of *modularity* that has the same kind of benefits as modularity in code; you can think about fewer things at a time. Time to extend the city? No need to plan Housing Projects and Development. Your job is just to dig a bunch of new foundations, and hook them up to water and electricity and Internet and roads. There, now your city is bigger. Everything else will sort itself out.

3: Conclusion.

I would put high odds against any of this happening in real life before further events are derailed by a near-lightspeed expanding front of von Neumann machines eating the galaxy (as happens in both good and bad AI scenarios). But if advanced AI does happen to take that long, or if Elon Musk gets bitten by the movable housing bug and makes it happen in two years, I'd enjoy living in Rivendell for some of the remaining time.

Imagine living in a higher-tech bigger nicer house, with robotic cars to teleport you wherever, and drones to deliver hot meals. Imagine that the rents not being so damned high – or even that 18th-century utopia where you *just* pay rent instead of rent plus tax – has produced a decrease of economic friction and a corresponding economic boom, so that it's less hard for people to make a living and get by. Is that a little more like that legendary future our parents were promised, of which it is said that instead we got 140 characters? Better steel factories didn't just produce shinier steel, back in the day; people made more tools that required steel, and used those steel tools to make other things. Scalable cities are something in that class. Movable houses, I think, might be a significant incremental piece of something in that class.

To me, it seems nearly certain that none of that will actually happen. This essay is not a prediction of future glories ahead of us, more of a wistful sigh. If we lived in a civilization where we could have nice things at that complexity level, we'd already have nicer versions of much simpler things.

But so long as Y Combinator is asking for essays on new cities anyway, this seemed worth writing up. I hope you enjoyed it!

PS: Please pave the sidewalks with the bouncy kind of pavement so that people can run on sidewalks without destroying their knees.

Insights from Euclid's 'Elements'

Presumably, I was taught geometry as a child. I do not remember.

Recently, I'd made my way halfway through a complex analysis textbook, only to find another which seemed more suitable and engaging. Unfortunately, the exposition was geometric. I knew something was wrong – I knew something had to change – when, asked to prove the similarity of two triangles, I got stuck on page 7.

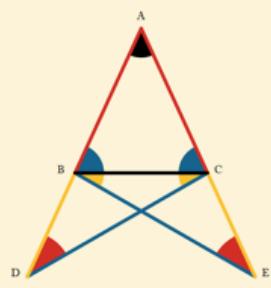
I'd been reluctant to tackle geometry, and when authors reasoned geometrically, I'd find another way to understand. Can you blame me, when most geometric proofs look like *this*?

LET the equal fides AB and AC be produced through the extremities BC, of the third fide, and in the produced part BD of either, let any point D be afflumed, and from the other let AE be cut off equal to AD (B. i. pr. 3.). Let the points E and D, fo taken in the produced fides, be connected by straight lines DC and BE with the alternate extremities of the third fide of the triangle.

In the triangles DAC and EAB the fides DA and AC are respectively equal to EA and AB, and the included angle A is common to both triangles. Hence (B. i. pr. 4.) the line DC is equal to BE, the angle ADC to the angle AEB, and the angle ACD to the angle ABE; if from the equal lines AD and AE the equal fides AB and AC be taken, the remainders BD and CE will be equal. Hence in the triangles BDC and CEB, the fides BD and DC are respectively equal to CE and EB, and the angles D and E included by those fides are also equal. Hence (B. i. pr. 4.) the angles DBC and ECB, which are those included by the third fide BC and the productions of the equal fides AB and AC are equal. Also the angles DCB and EBC are equal if those equals be taken from the angles DCA and EBA before proved equal, the remainders, which are the angles ABC and ACB opposite to the equal fides, will be equal.

Therefore in an isosceles triangle, &c.

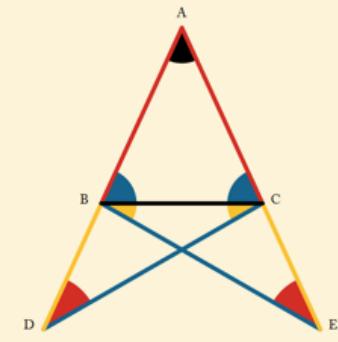
Q.E.D.



Distasteful. In a graph with n vertices, you'd need to commit $O(n^3)$ things to memory (e.g. triangles, angles) in order to read the proof without continually glancing at the illustration. In a normal equation with n variables, it's $O(n)$.

Sometimes, we just need a little beauty to fall in love.

Produce and
 make =
 Draw = (B. i. pr. 3.)
 in and
 we have =
 = and common:
 \triangle = , \triangle =
 and = (B. i. pr. 4.)
 Again in and ,
 = ,
 = ,
 and = ;
 \triangle =
 and = (B. i. pr. 4.).
 But = ,
 \triangle = .



Q.E.D.

Welcome to Oliver Byrne's rendition of Euclid's *Elements*, [digitized and freely available online](#).

PROPOSITION I. PROBLEM.



Given a given finite straight line (—) to describe an equilateral triangle.



Describe \odot and \odot (postulate 3.); draw — and — (post. 1.). then \triangle will be equilateral.

For $\text{---} = \text{---}$ (def. 15.);
and $\text{---} = \text{---}$ (def. 15.),
 $\therefore \text{---} = \text{---}$ (axiom. 1.);

and therefore \triangle is the equilateral triangle required.

Q. E. D.

Elements

Propositions are placed before a student, who though having a sufficient understanding, is told just as much about them on entering at the very threshold of the science, as gives him a preposition most unfavourable to his future study of this delightful subject; or "the formalities and paraphernalia of rigour are often tentatively put forward, as almost to hide the reality. Endless and perplexing repetitions, which do not confer greater exactitude on the reasoning, render the demonstrations involved and obscure, and conceal from the view of the student the confirmation of evidence."

Thus an aversion is created in the mind of the pupil, and a subject so calculated to improve the reasoning powers, and give the habit of close thinking, is degraded by a dry and rigid course of instruction into an uninteresting exercise of the memory.

~ [Oliver Byrne](#)

Equality and Similarity

Old mathematical writing lacks modern precision. Euclid says that two triangles are "equal", without specifying what that means. It means that one triangle can be turned into another via an [isometric transformation](#). That is, if you rotate, translate, and/or reflect triangle A, it turns into triangle B.

[Similarity](#) is a bit more lenient, because you can rescale as well:



My favorite characterization of similarities is:

As a map $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, a similarity of ratio r takes the form $f(x) := rAx + t$, where $A \in O_n(\mathbb{R})$ is an $n \times n$ orthogonal matrix and $t \in \mathbb{R}^n$ is a translation vector.

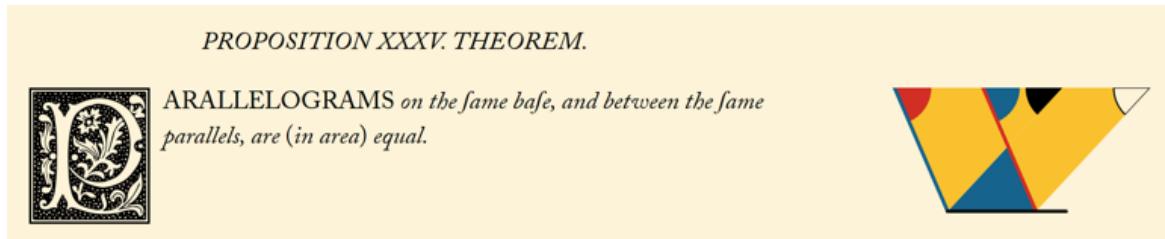
The only difference compared to congruence is that congruence requires $r = 1$.

Synthetic/analytic

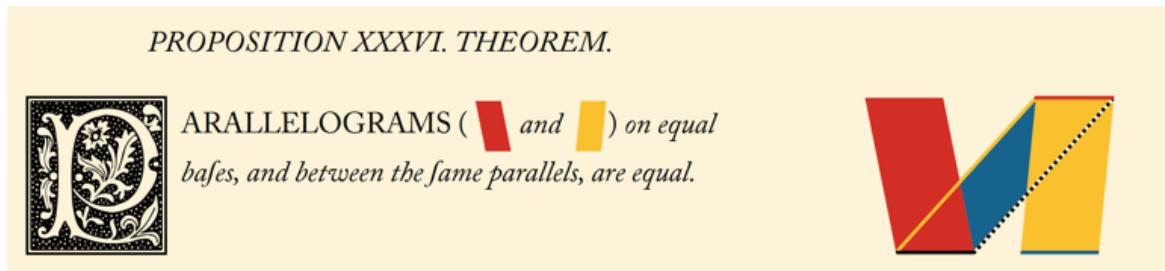
I find it strange that Euclid got so far by axiomatizing informal notions without any grounding in formal set theory (e.g. ZFC). I mean, you'd get *absolutely blown away* if you tried to pull these shenanigans in topology. But apparently, Euclidean geometry is sufficiently well-behaved that it basically matches our intuitions without much qualification?

Area invariance

[Book 1, proposition 35:](#)



This says: suppose you draw two parallel lines, and then make a dash of length 2 on each line. Then, make another dash of length 2 on the upper line. The two parallelograms so defined have equal area. This is clarified in the next theorem.



If you take one of the dashes and slide it around on the upper parallel line, the resultant parallelograms all have the same area. I thought this was cool.

Notes

There aren't any exercises; instead, I tried to first prove the theorems myself.

Book III treats circles, with wonderful results on arcs and their relation to angles. I search for a snappy example, a gem of an insight to share, but my words fail me. It's just good.

I read books I, III, IV, and skimmed II. Not all books of the Elements are about plane geometry; some are archaic introductions to number theory, for example. Those looking to learn number theory would do much better with the gorgeous [Illustrated Theory of Numbers](#).

Forward

Elements is a *tour de force*. Theorem, theorem, problem, theorem, all laid out in confident succession. It was not always known that from simple rules you could rigorously deduce beautiful facts. It was *not always known* that you could start with so little, and end with so much.

Before I found this resource, I'd checked out several geometry books, all of which seemed bad. To salt the wound, many books were explicitly aimed at middle-schoolers. This... was a bit of a blow.

However, it doesn't matter when something is normally presented. If you don't know something, you don't know it, and there's nothing wrong with learning it. Even if you feel late. Even if you feel sheepish.

Against completionism

I'm glad I didn't read all of the books, even though they're beautiful. I'd picked up a bad "completionist" habit – if I don't read the whole book, obviously I haven't completed it, and obviously I'm not allowed to make a post about it. Of course.

But I'm trying to pick up useful skills, to expand the types of qualitative reasoning available to me, to get the most benefit per unit of reading. I stopped because I have what I need for my complex analysis book.

Read around

Reading relevant Wikipedia pages / other textbooks helps me cross-examine my knowledge. It also helps connect the new knowledge to existing knowledge. For example, I now have a wonderfully enriched understanding of [the geometric mean](#).

Over time, as you expand and read more books, you'll find yourself reading faster and faster, understanding more and more subsections. [I don't recommend learning new areas via Wikipedia](#), but it's good reinforcement.

Re-deriving dependencies as a habit

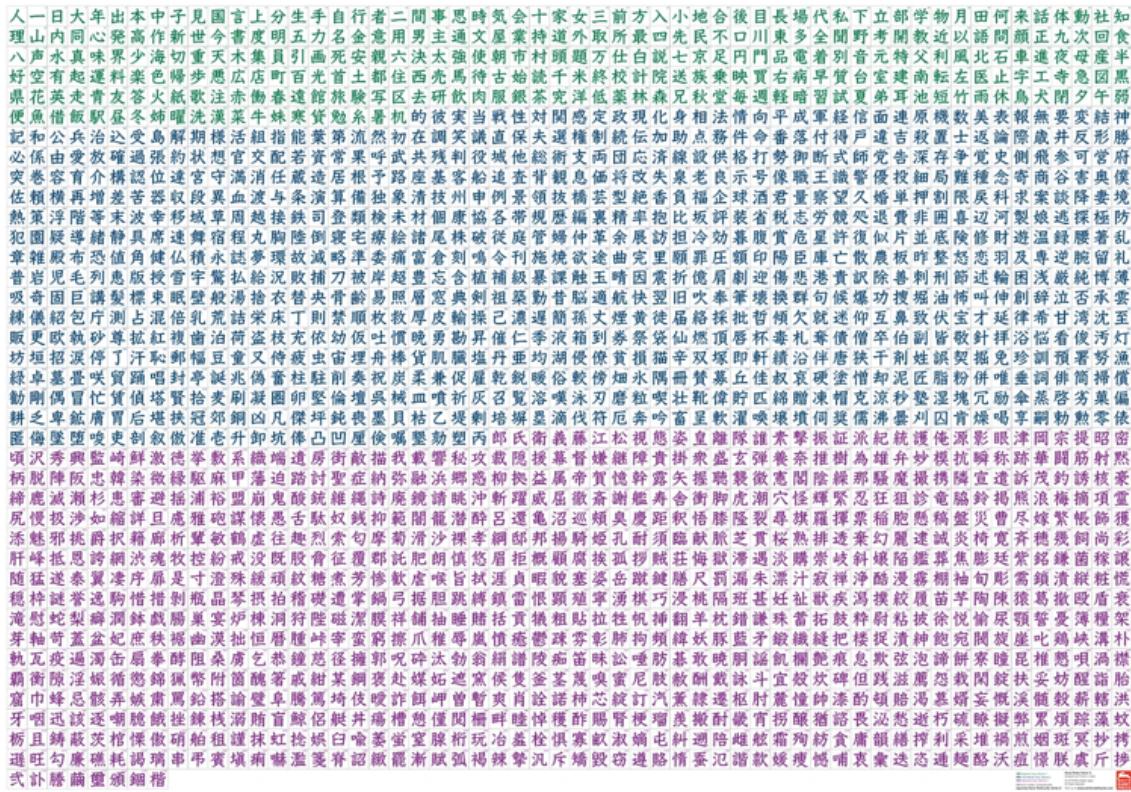
Ever since [I learned real analysis](#), I reflexively reprove all new elementary mathematics whenever I use it. For real analysis, that meant *continually reproving* e.g. $1 + 2 = 2 + 1$ whenever I used that property in a proof. Did it feel silly and tedious? A bit, yes.

But with (this) tedium comes power. I can now regenerate a formal foundation for the real numbers from the Peano axioms, proving the necessary properties about the natural numbers, then the integers, then the rationals, and then the reals, and then complex numbers too. (But please, no quaternions!)

With this habit, you continually ask yourself, "how do I know this?". I think this is a useful subskill of Actually Thinking.

Commemoration

In college, I taught myself a bit of Japanese. Through a combination of spaced repetition software and memory palaces, and over the course of three months, I learned to read the [2,136 standard use characters](#). After those three months, I proudly displayed this poster on my wall:



I look forward to another beautiful poster.



As the fenses of sight and hearing can be so forcibly and instantaneously addressed alike with one thousand as with one, *the million* might be taught geometry and other branches of mathematics with great ease, this would advance the purpose of education more than any thing that *might* be named, for it would teach the people how to think, and not what to think; it is in this particular the great error of education originates.

Why Artists Study Anatomy

I've been interested in drawing ever since I was a little kid. I like drawing people. I practiced a lot by looking at a cool picture of a character I like, and copying it as closely as I can.

The better I got at copying, the better my work became. I got some great results this way. I even advanced to the level where I could take photos of real-life people, then painstakingly copying it to produce a portrait. I was pleased by and proud of many of my works. Many agreed that I was "such a good artist, wow!"



Here's one piece I was particularly proud of; my sketch on the left, the reference on the right.

In middle school, I took on a new challenge: **I wanted to draw commissions.** People would give me a character design, sometimes with a few reference images, and I would draw their characters. This was really hard for some reason...

I had no images to copy directly from. My results were inconsistent. I spent a lot of my time tweaking things like moving around the eyes, redrawing the nose, etc. After a bunch of commissions, I got reasonably decent at drawing a character at the $\frac{3}{4}$ angle, the most common one for portraiture. I did hundreds of commissions, almost all of them at that angle. Soon, I felt uncomfortable drawing faces at any other angle.

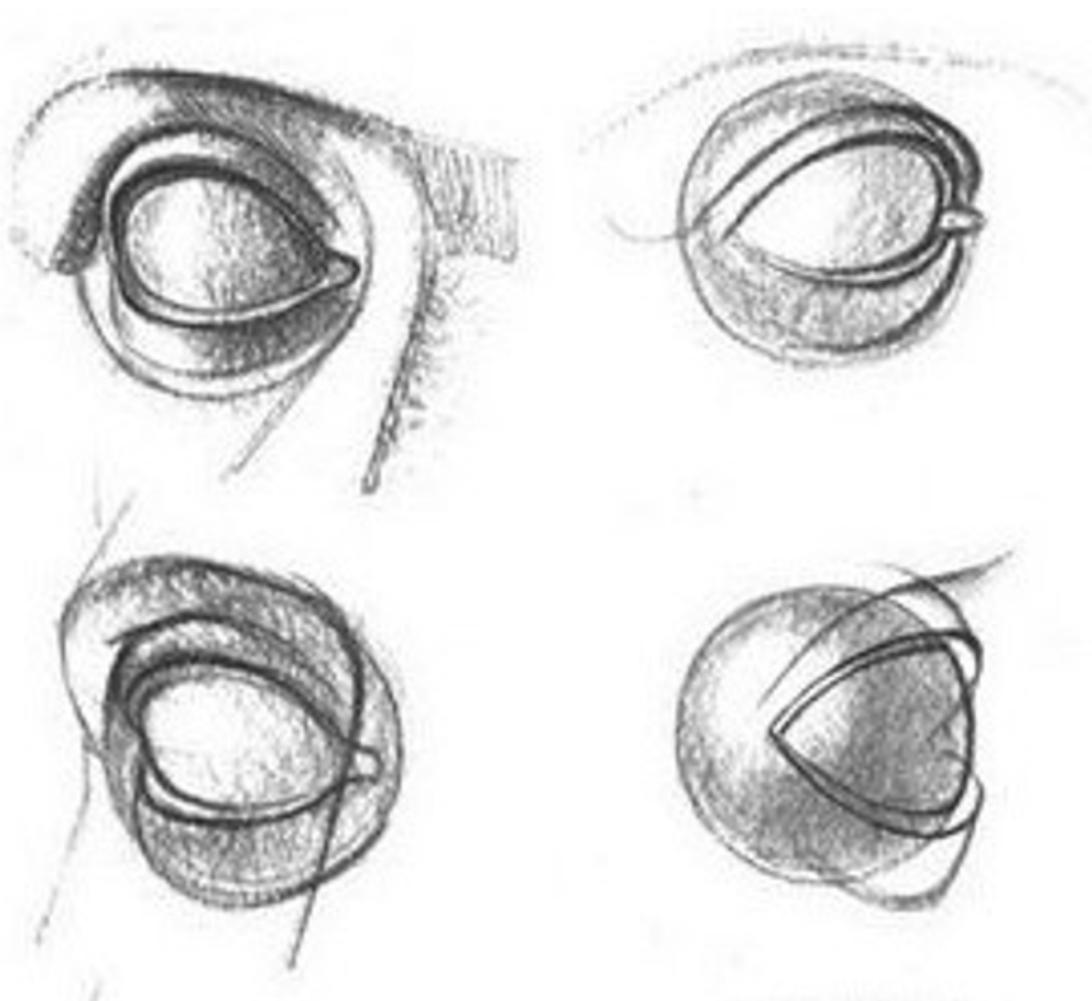


Some examples of the stuff I drew. Some of them look kinda strangely similar, hmm...

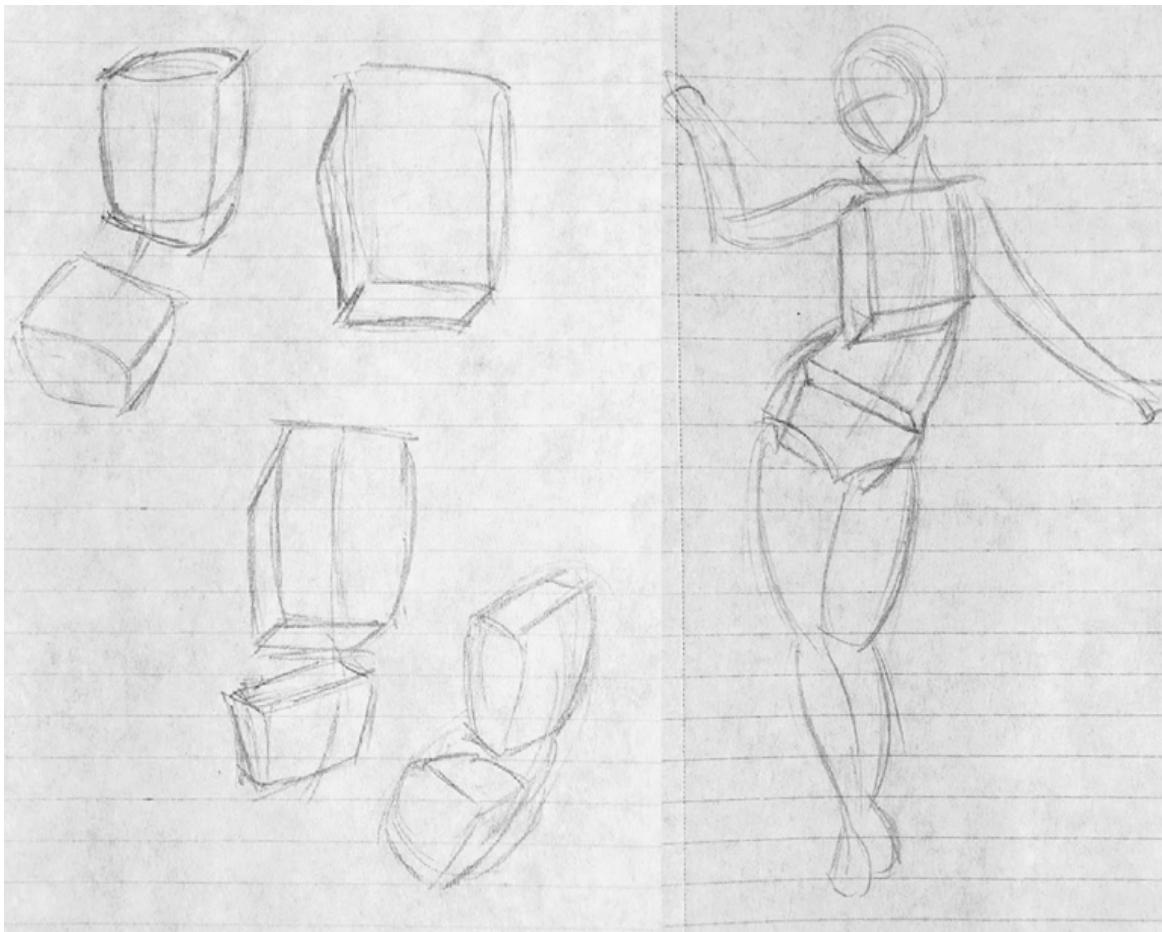
I stopped doing art for a while because college was busy. When I started again, I took [an anatomy class](#). I learned a lot of great things. But the biggest lesson I learned was that I've been approaching drawing all wrong. The biggest thing I had missed was that **I had to understand the process that generates the image**.

The art an artist draws is a projection of the 3D object onto a 2D surface. By practicing by only copying from 2D images without studying the 3D structures that generated them, **I treated my reference images as being generated by some black box, and ignored the underlying gears and structures.**

For example, the shape of an eye is determined by the 3D structures of the eyeball projected into 2D. I never learned what an eye looks like in 3D, so I didn't know how to draw it from an angle different from the 2D references I've seen. It turns out, the eye isn't a flat almond-shape like it seems in a picture. It's a sphere inset into the skull, wrapped from the top and bottom by flaps of skin. No wonder I hit a wall when I tried to draw an eye from the side. I had never seen the sphere or practiced drawing it as a sphere. I probably tried to draw it as a slightly squished almond, then wondered why it looked wrong.

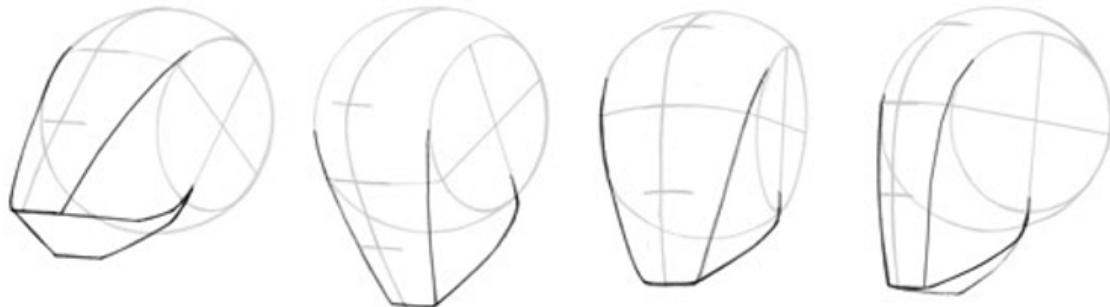


Another example: the lines describing a body in a drawing are dictated by the underlying bone and muscle structures. I had never studied these before. In the anatomy class, we looked at body parts in 3D. We reduced complex shapes (like a human torso) into simpler geometric shapes (boxes and spheres) which are easier to mentally rotate in space. Then we constructed lines of the torso using these basic shapes.



Some of my sketches of the torso.

Similarly, I learned to draw a head by starting with a sphere and a box. I then used these shapes to produce guiding lines, which then dictated where the surface features (eyes, nose, mouth) had to be located. If I wanted to draw the head from a different angle, I started with a sphere and box angled differently, which then produced different guiding lines. There was no more “guessing” where the eyes should go.



[From this post](#)

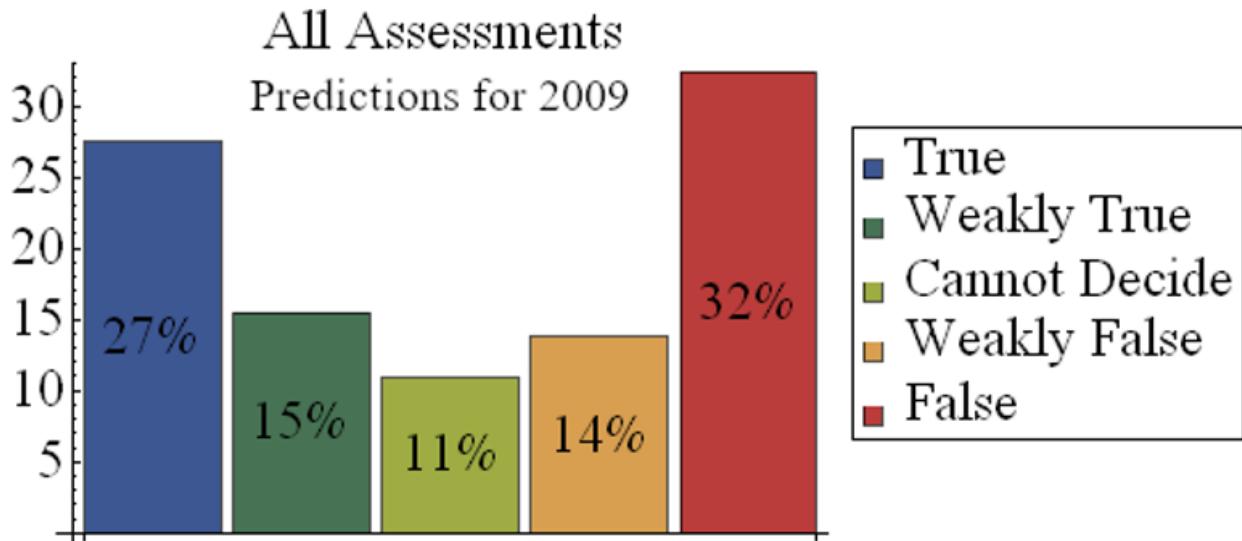
This is the reason artists draw from live models and study anatomy. The 3D shapes of the underlying structures generate the simple-seeming lines on a drawing. Drawing well requires opening the blackbox and looking at the internal gears of human anatomy.

Assessing Kurzweil predictions about 2019: the results

EDIT: Mean and standard deviation of individual predictions can be found [here](#).

Thanks to all my brave assessors, I now have the data about Kurzweil's [1999 predictions about 2019](#).

This was a follow up to a [previous assessment about his predictions about 2009](#), which showed a mixed bag. Roughly evenly divided between right and wrong, which I found pretty good for ten-year predictions:



So, did more time allow for trends to overcome noise or more ways to go wrong?

Pause for a moment to calibrate your expectations.

Methods and thanks

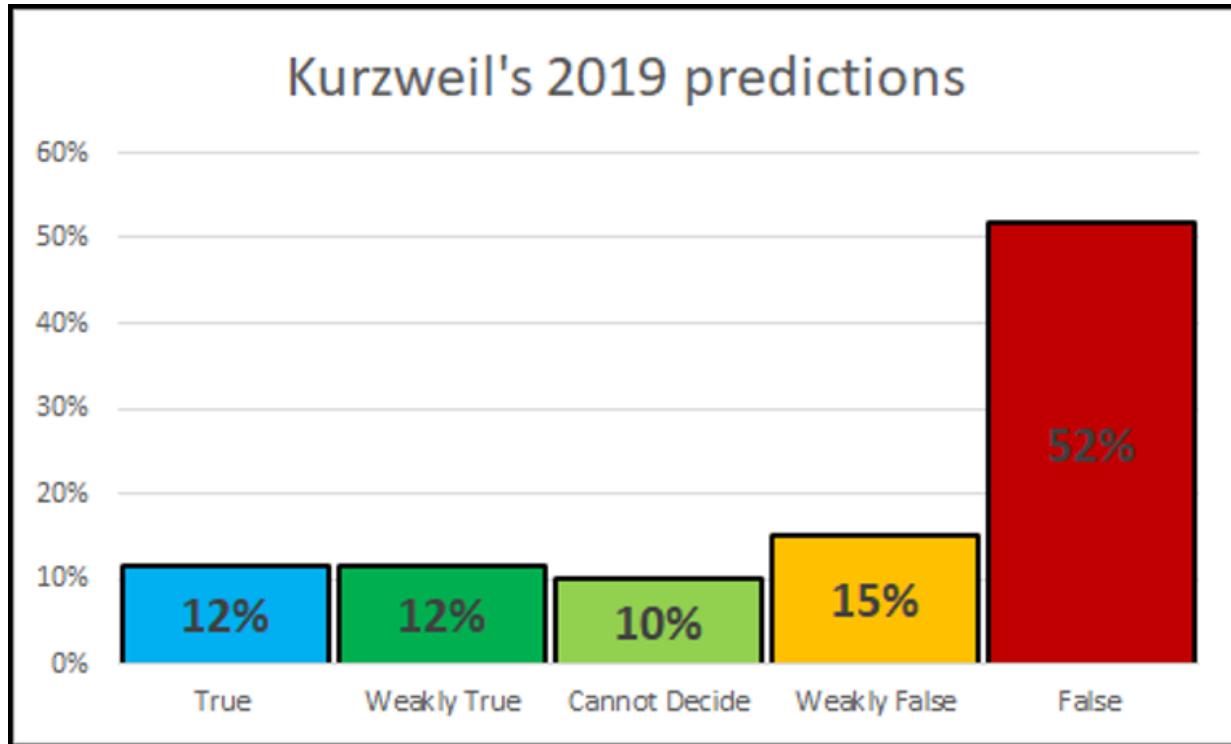
So, for the 2019 predictions, I divided them into [105 separate statements](#), did a call for volunteers, with [instructions here](#); the main relevant point being that I wanted their assessment for 2019, not for the (possibly transient) current situation. I got 46 volunteers with valid email addresses, of which 34 returned their predictions. So many thanks, in reverse alphabetical order, to Zvi Mowshowitz, Zhengdong Wang, Yann Riviere, Uriel Fiori, orthonormal, Nuño Sempere, Nathan Armishaw, Koen Holtman, Keller Scholl, Jaime Sevilla, Gareth McCaughan, Eli Rose and Dillon Plunkett, Daniel Kokotajlo, Anna Gardiner... and others who have chosen to remain anonymous.

The results

Enough background; what did the assessors find? Well, of the 34 assessors, 24 went the whole hog and did all 105 predictions; on average, 91 predictions were assessed by

each person, a total of 3078 individual assessments^[1].

So, did more time allow for more perspective or more ways to go wrong? Well, Kurzweil's predictions for 2019 were considerably worse than those for 2009, with more than half strongly wrong:



Interesting details

The (anonymised) data can be [found here^{\[2\]}](#), and I encourage people to download and assess it themselves. But some interesting results stood out to me:

Predictor agreement

Taking a single prediction, for instance the first one:

- 1: Computers are now largely invisible. They are embedded everywhere--in walls, tables, chairs, desks, clothing, jewelry, and bodies.

Then we can compute the standard deviation of the predictors' answer for that prediction. This gives an impression of how much disagreement there was between predictors; in this case, it was 0.84.

Perfect agreement would be a standard deviation of 0; maximum disagreement (half find "1", half find "5") would be a standard deviation of 2. Perfect spread - equal numbers of 1s, 2s, 3s, 4s, and 5s - would have a standard deviation of $\sqrt{2} \approx 1.4$.

Across the 105 predictions, the maximum standard deviation was **1.7**, the minimum was **0** (perfect agreement), and the average was **0.97**. So the predictors had a medium tendency to agree with each other.

Most agreement/falsest predictions

There was perfect agreement on five predictions; and on all of these, the agreed prediction was always "5": "False".

These predictions were:

- 51: "Phone" calls routinely include high-resolution three-dimensional images projected through the direct-eye displays and auditory lenses.
- 55: [...] Thus a person can be fooled as to whether or not another person is physically present or is being projected through electronic communication.
- 59: The all-enveloping tactile environment is now widely available and fully convincing.
- 62: [...] These technologies are popular for medical examinations, as well as sensual and sexual interactions with other human partners or simulated partners.
- 63: [...] In fact, it is often the preferred mode of interaction, even when a human partner is nearby, due to its ability to enhance both experience and safety.

As you can see, Kurzweil suffered a lot from his VR predictions. This seems a perennial thing: Hollywood is always convinced that mass 3D is just around the corner; technologists are convinced that VR is imminent.

Most accurate prediction:

With a mean score of 1.3, the prediction deemed most accurate was:

- 83: The existence of the human underclass continues as an issue.

Now this might seem a trivial prediction, especially in retrospect, but I want to defend Kurzweil here - it was not at all certain in 1999, with many utopian changes foreseen and expected, that this would still be an issue.

The next prediction deemed most accurate (mean of 1.4), is:

- 82: People attempt to protect their privacy with near-unbreakable encryption technologies, but privacy continues to be a major political and social issue with each individual's practically every move stored in a database somewhere.

This is truly non-trivial for 1999, and I do give Kurzweil credit for that.

Least agreement

With a standard deviation of 1.7, the predictors disagreed the most on this prediction:

- 37: Computation in general is everywhere, so a student's not having a computer is rarely an issue.

This may have to do with different judgement over the extent of "everywhere" and "rarely an issue", or over who might or might not find this to be an issue.

The next prediction with the most disagreement (st dev 1.6) is:

- 16: Rotating memories and other electromechanical computing devices have been fully replaced with electronic devices.

It's possible that "fully" was a problem here, but I see this prediction as being just false.

Most "Cannot Decide"

This prediction had more than 46% of predictors choosing "Cannot Decide":

- 20: It is now fully recognized that the brain comprises many specialized regions, each with its own topology and architecture of interneuronal connections.

Maybe the ambiguity in "fully recognized" made this hard to assess. Or maybe, as [suggested in the comments](#), because this doesn't look much like a "prediction", but an obviously true statement?

A question of timeline?

It's been suggested that [Kurzweil's predictions for 2009 are mostly correct in 2019](#). If this is the case - Kurzweil gets the facts right, but the timeline wrong - it would be interesting to revisit these predictions in 2029 (if he is a decade optimistic) and 2039 (if he expected things to go twice as fast). Though many of his predictions seem to be of the type "once true, always true", so his score should rise with time, assuming continuing technological advance and no disasters.

In conclusion

Again, thanks to all the volunteers who assessed these predictions and thanks to Kurzweil who, unlike most prognosticators, had the guts and the courtesy to write down his predictions and give them a date.

I strongly suspect that most people's 1999 predictions about 2019 would have been a lot worse.

-
1. Five assessments of the 3078 returned question marks; I replaced these with "3" ("Cannot Decide"). Four assessors of the 34 left gaps in their predictions, instead of working through the randomly ordered predictions; to two significant figures, excluding these four didn't change anything, so I included them all. [↩](#)
 2. Each column is an individual predictor, each row an individual prediction. [↩](#)

Studies On Slack

I.

Imagine a distant planet full of eyeless animals. Evolving eyes is hard: they need to evolve Eye Part 1, then Eye Part 2, then Eye Part 3, in that order. Each of these requires a separate series of rare mutations.

Here on Earth, scientists believe each of these mutations must have had [its own benefits](#) – in the land of the blind, the man with only Eye Part 1 is king. But on this hypothetical alien planet, there is no such luck. You need all three Eye Parts or they’re useless. Worse, each Eye Part is metabolically costly; the animal needs to eat 1% more food per Eye Part it has. An animal with a full eye would be much more fit than anything else around, but an animal with only one or two Eye Parts will be at a small disadvantage.

So these animals will only evolve eyes in conditions of relatively weak evolutionary pressure. In a world of intense and perfect competition, where the fittest animal always survives to reproduce and the least fit always dies, the animal with Eye Part 1 will always die – it’s less fit than its fully-eyeless peers. The weaker the competition, and the more randomness dominates over survival-of-the-fittest, the more likely an animal with Eye Part 1 can survive and reproduce long enough to eventually produce a descendant with Eye Part 2, and so on.

There are lots of ways to decrease evolutionary pressure. Maybe natural disasters often decimate the population, dozens of generations are spent recolonizing empty land, and during this period there’s more than enough for everyone and nobody has to compete. Maybe there are frequent [whalefalls](#), and any animal nearby has hit the evolutionary jackpot and will have thousands of descendants. Maybe the population is isolated in little islands and mountain valleys, and one gene or another can reach fixation in a population totally by chance. It doesn’t matter exactly how it happens, it matters that evolutionary pressure is low.

The branch of evolutionary science that deals with this kind of situation is called “adaptive fitness landscapes”. Landscapes really are a great metaphor – consider somewhere like this:



You pour out a bucket of water. Water “flows downhill”, so it’s tempting to say something like “water wants to be at the lowest point possible”. But that’s not quite right. The lowest point possible is the pit, and water won’t go there. It will just sit in the little puddle forever, because it would have to go up the tiny little hillock in order to get to the pit, and water can’t flow uphill. Using normal human logic, we feel tempted to say something like “Come on! The hillock is so tiny, and that pit is so deep, just make a single little exception to your ‘always flow downhill’ policy and you could do so much better for yourself!” But water stubbornly refuses to listen.

Under conditions of perfectly intense competition, evolution works the same way. We imagine a multidimensional evolutionary “landscape” where lower ground represents higher fitness. In this perfectly intense competition, organisms can go from higher to lower fitness, but never vice versa. As with water, the tiniest hillock will leave their potential forever unrealized.

Under more relaxed competition, evolution only *tends probabilistically* to flow downhill. Every so often, it will flow uphill; the smaller the hillock, the more likely evolution will surmount it. Given enough time, it's guaranteed to reach the deepest pit and mostly stay there.

Take a moment to be properly amazed by this. It sounds like something out of the Tao Te Ching. An animal with eyes has very high evolutionary fitness. It will win at all its evolutionary competitions. So in order to produce the highest-fitness animal, we need to - select for fitness less hard? In order to produce an animal that wins competitions, we need to stop optimizing for winning competitions?

This doesn't mean that less competition is always good. An evolutionary environment with no competition won't evolve eyes either; a few individuals might randomly drift into having eyes, but they won't catch on. In order to optimize the species as much as possible as fast as possible, you need the right balance, somewhere in the middle between total competition and total absence of competition.

In the esoteric teachings, total competition is called [Moloch](#), and total absence of competition is called [Slack](#). Slack (thanks to Zvi Moskovitz for the term and concept) gets short shrift. If you think of it as "some people try to win competitions, other people don't care about winning competitions and slack off and go to the beach", you're misunderstanding it. Think of slack as a paradox - the Taoist art of winning competitions by not trying too hard at them. Moloch and Slack are opposites and complements, like yin and yang. Neither is stronger than the other, but their interplay creates the ten thousand things.

II.

Before we discuss slack further, a digression on group selection.

Some people would expect this discussion to be quick, since group selection doesn't exist. These people understand it as evolution acting for the good of a species. It's a tempting way to think, because evolution usually eventually makes species stronger and more fit, and sometimes we colloquially round that off to evolution targeting a species' greater good. But inevitably we find evolution is awful and does absolutely nothing of the sort.

Imagine an alien planet that gets hit with a solar flare once an eon, killing all unshielded animals. Sometimes unshielded animals spontaneously mutate to shielded, and vice versa. Shielded animals are completely immune to solar flares, but have 1% higher metabolic costs. What happens? If you predicted "magnetic shielding reaches fixation and all animals get it", you've fallen into the group selection trap. The unshielded animals outcompete the shielded ones during the long inter-flare period, driving their population down to zero (though a few new shielded ones arise every generation through spontaneous mutations). When the flare comes, only the few spontaneous mutants survive. They breed a new entirely-shielded population, until a few unshielded animals arise through spontaneous mutation. The unshielded outcompete the shielded ones again, and by the time of the next solar flare, the population is 100% unshielded again and they all die. If the animals are lucky, there will always be enough spontaneously-mutated shielded animals to create a post-flare breeding population; if they are unlucky, the flare will hit at a time with unusually few such mutants, and the species will go extinct.

An Evolution Czar concerned with the good of the species would just declare that all animals should be shielded and solve the problem. In the absence of such a Czar, these animals will just keep dying in solar-flare-induced mass extinctions forever, even though there is an easy solution with only 1% metabolic cost.

A less dramatic version of the same problem happens here on Earth. Every so often predators (let's say foxes) reproduce too quickly and outstrip the available supply of prey (let's say rabbits). There is a brief period of starvation as foxes can't find any more rabbits and die *en masse*. This usually ends with a [boom-bust cycle](#): after most foxes die, the rabbits (who reproduce very quickly and are now free of predation) have a population boom; now there are rabbits everywhere. Eventually the foxes catch up, eat all the new rabbits, and the cycle repeats again. It's a waste of resources for foxkind to spend so much of time and energy breeding a huge population of foxes that will inevitably collapse a generation later; an Evolution Czar concerned with the common good would have foxes limit their breeding at a sustainable level. But since individual foxes that breed excessively are more likely to have their genes represented in the next generation than foxes that breed at a sustainable level, we end up with foxes that breed excessively, and the cycle continues.

([but humans are too smart to fall for this one, right?](#))

Some scientists [tried to create group selection under laboratory conditions](#). They divided some insects into subpopulations, then killed off any subpopulation whose numbers got too high, and promoted any subpopulation that kept its numbers low to better conditions. They hoped the insects would evolve to naturally limit their family size in order to keep their subpopulation alive. Instead, the insects became cannibals: they ate other insects' children so they could have more of their own without the total population going up. In retrospect, this makes perfect sense; an insect with the behavioral program "have many children, and also kill other insects' children" will have its genes better represented in the next generation than an insect with the program "have few children".

But sometimes evolution appears to solve group selection problems. What about multicellular life? Stick some cells together in a resource-plentiful environment, and they'll naturally do the evolutionary competition thing of eating resources as quickly as possible to churn out as many copies of themselves as possible. If you were expecting these cells to form a unitary organism where individual cells do things like become heart cells and just stay in place beating rhythmically, you would call the expected normal behavior "cancer" and be against it. Your opposition would be on firm group selectionist grounds: if any cell becomes cancer, it and its descendants will eventually overwhelm everything, and the organism (including all cells within it, including the cancer cells) will die. So for the good of the group, none of the cells should become cancerous.

The first step in evolution's solution is giving all cells the same genome; this mostly eliminates the need to compete to give their genes to the next generation. But this solution isn't perfect; cells can get mutations in the normal course of dividing and doing bodily functions. So it employs a host of other tricks: genetic programs telling cells to self-destruct if they get too cancer-adjacent, an immune system that hunts down and destroys cancer cells, or growing old and dying (this last one isn't usually thought of as a "trick", but it absolutely is: if you arrange for a cell line to lose a little information during each mitosis, so that it degrades to the point of gobbledegook after X divisions, this means cancer cells that divide constantly will die very quickly, but normal cells dividing on an approved schedules will last for decades).

Why can evolution "develop tricks" to prevent cancer, but not to prevent foxes from overbreeding, or aliens from losing their solar flare shields? Group selection works when the group itself has a shared genetic code (or other analogous ruleset) that can evolve. It doesn't work if you expect it to directly change the genetic code of each individual to cooperate more.

When we think of cancer, we are at risk of conflating two genetic codes: the shared genetic code of the multicellular organism, and the genetic code of each cell within the organism. Usually (when there are no mutations in cell divisions) these are the same. Once individual

cells within the organism start mutating, they become different. Evolution will select for cancer in changes to individual cells' genomes over an organism's lifetime, but select against it in changes to the overarching genome over the lifetime of the species (ie you should expect all the genes you inherited from your parents to be selected against cancer, and all the mutations in individual cells you've gotten since then to be selected for cancer).

The fox population has no equivalent of the overarching genome; there is no set of rules that govern the behavior of every fox. So foxes can't undergo group selection to prevent overpopulation (there are some more complicated dynamics that might still be able to rescue the foxes in some situations, but they're not relevant to the simple model we're looking at).

In other words, group selection can happen in a two-layer hierarchy of nested evolutionary systems when the outer system (eg multicellular humans) includes rules that the inner system (eg human cells) have to follow, and where the fitness of the evolving-entities in the outer system depends on some characteristics of the evolving-entities in the inner system (eg humans are higher-fitness if their cells do not become cancerous). The evolution of the outer layer includes evolution over rulesets, and eventually evolves good strong rulesets that tell the inner-layer evolving entities how to behave, which can include group selection (eg humans evolve a genetic code that includes a rule "individual cells inside of me should not get cancer" and mechanisms for enforcing this rule).

You can find these kinds of two-layer evolutionary systems everywhere. For example, "cultural evolution" is a two-layer evolutionary system. In the hypothetical state of nature, there's unrestricted competition – people steal from and murder each other, and only the strongest survive. After they form groups, the groups compete with each other, and groups that develop rulesets that prevent theft and murder (eg legal codes, religions, mores) tend to win those competitions. Once again, the outer layer (competition between cultures) evolves groups that successfully constrains the inner layer (competition between individuals). Species don't have a czar who restraints internal competition in the interest of keeping the group strong, but some human cultures do (eg Russia).

Or what about market economics? The outer layer is companies, the inner layer is individuals. Maybe the individuals are workers – each worker would selfishly be best off if they spent the day watching YouTube videos and pushed the hard work onto someone else. Or maybe they're executives – each individual executive would selfishly be best off if they spent their energy on office politics, trying to flatter and network with whoever was most likely to promote them. But if all the employees loaf off and all the executives focus on office politics, the company won't make products, and competitors will eat their lunch. So someone – maybe the founder/CEO – comes up with a ruleset to incentivize good work, probably some kind of performance review system where people who do good work get promoted and people who do bad work get fired. The outer-layer competition between companies will select for corporations with the best rulesets; over time, companies' internal politics should get better at promoting the kind of cooperation necessary to succeed.

How do these systems replicate multicellular life's success without being literal entities with literal DNA having literal sex? They all involve a shared ruleset and a way of punishing rulebreakers which make it in each individual's short-term interest to follow the ruleset that leads to long-term success. Countries can do that (follow the law or we'll jail you), companies can do that (follow our policies or we'll fire you), even multicellular life can sort of do that (don't become cancer, or immune cells will kill you). When there's nothing like that (like the overly-fast-breeding foxes) evolution fails at group selection problems. When there is something like that, it has a chance. When there's something like that, and the thing like that is itself evolving (either because it's encoded in literal DNA, or because it's encoded in things like company policies that determine whether a company goes out of

business or becomes a model for others), then it can reach a point where it solves group selection problems very effectively.

In the esoteric teachings, the inner layer of two-layer evolutionary systems is represented by the Goddess of Cancer, and outer layer by the [Goddess of Everything Else](#). In each part of the poem, the Goddess of Cancer orders the evolving-entities to compete, but the Goddess of Everything Else recasts it as a two-layer competition where cooperation on the internal layer helps win the competition on the external layer. He who has ears to hear, let him listen.

III.

Why the digression? Because slack is a group selection problem. A species that gave itself slack in its evolutionary competition would do better than one that didn't – for example, the eyeless aliens would evolve eyes and get a big fitness boost. But no individual can unilaterally choose to compete less intensely; if it did, it would be outcompeted and die. So one-layer evolution will fail at this problem the same way it fails all group selection problems, but two-layer systems will have a chance to escape the trap.

The multicellular life example above is a special case where you want 100% coordination and 0% competition. I framed the other examples the same way – countries do best when their citizens avoid all competition and work together for the common good, companies do best when their executives avoid self-aggrandizing office politics and focus on product quality. But as we saw above, some systems do best somewhere in the middle, where there's some competition but also some slack.

For example, consider a researcher facing their own version of the eyeless aliens' dilemma. They can keep going with business as normal – publishing trendy but ultimately useless papers that nobody will remember in ten years. Or they can work on Research Program Part 1, which *might* lead to Research Program Part 2, which *might* lead to Research Program Part 3, which *might* lead to a ground-breaking insight. If their jobs are up for review every year, and a year from now the business-as-normal researcher will have five trendy papers, and the groundbreaking-insight researcher will be halfway through Research Program Part 1, then the business-as-normal researcher will outcompete the groundbreaking-insight researcher; as the saying goes, “publish or perish”. Without slack, no researcher can unilaterally escape the system; their best option will always be to continue business as usual.

But group selection makes the situation less hopeless. Universities have long time-horizons and good incentives; they want to get famous for producing excellent research. Universities have rulesets that bind their individual researchers, for example “after a while good researchers get tenure”. And since universities compete with each other, each is incentivized to come up with the ruleset that maximizes long-term researcher productivity. So if tenure really does work better than constant vicious competition, then (absent the usual culprits like resistance-to-change, weird signaling equilibria, politics, etc) we should expect universities to converge on a tenure system in order to produce the best work. In fact, we should expect universities to evolve a really impressive ruleset for optimizing researcher incentives, just as impressive as the clever mechanisms the human body uses to prevent cancer (since this seems a bit optimistic, I assume the usual culprits are not absent).

The same is true for grant-writing; naively you would want some competition to make sure that only the best grant proposals get funded, but too much competition seems to stifle original research, so much so that some funders are [throwing out the whole process and selecting grants by lottery](#), and others are [running grants you can apply for in a half-hour and hear back about two days later](#). If there's a feedback mechanism – if these different

rulesets produce different-quality research, and grant programs that produce higher-quality research are more likely to get funded in the future – then the rulesets for grants will gradually evolve, and the competition for grants will take place in an environment with whatever the right evolutionary parameters for evolving good research are.

I don't want to say these things will definitely happen – you can read [Inadequate Equilibria](#) for an idea of why not. But they *might*. The evolutionary dynamics which would normally prevent them can be overcome. Two-layer evolutionary systems can produce their own slack, if having slack would be a good idea.

IV.

That was a lot of paragraphs, and a lot of them started with “imagine a hypothetical situation where...”. Let’s look deeper into cases where an understanding of slack can inform how we think about real-world phenomena. Five examples:

1. Monopolies. Not the kind that survive off overregulation and patents, the kind that survive by being big enough to crush competitors. These are predators that exploit low-slack environments. If Boeing has a monopoly on building passenger planes, and is exploiting that by making shoddy products and overcharging consumers, then that means anyone else who built a giant airplane factory could make better products at a lower price, capture the whole airplane market, and become a zillionaire. Why don’t they? Slack. In terms of those adaptive fitness landscapes, in between your current position (average Joe) and a much better position at the bottom of a deep pit (you own a giant airplane factor and are a zillionaire), there’s a very big hill you have to climb – the part where you build Giant Airplane Factory Part 1, Giant Airplane Factory Part 2, etc. At each point in this hill, you are worse off than somebody who was not building an as-yet-unprofitable giant airplane factory. If you have infinite slack (maybe you are Jeff Bezos, have unlimited money, and will never go bankrupt no matter how much time and cost it takes before you start earning profits) you’re fine. If you have more limited slack, your slack will run out and you’ll be outcompeted before you make it to the greater-fitness deep pit.

Real monopolies are more complicated than this, because Boeing can shape up and cut prices when you’re halfway to building your giant airplane factory, thus removing your incentive. Or they can do *actually* shady stuff. But none of this would matter if you already had your giant airplane factory fully built and ready to go – at worst, you and Boeing would then be in a fair fight. Everything Boeing does to try to prevent you from building that factory is exploiting your slacklessness and trying to increase the height of that hill you have to climb before the really deep pit.

(Peter Thiel inverts the landscape metaphor and calls the hill a “moat”, but he’s getting at the same concept).

2. Tariffs. Same story. Here’s the way I understand the history of the international auto industry – anyone who knows more can correct me if I’m wrong. Automobiles were invented in the early 20th century. Several Western countries developed homegrown auto industries more or less simultaneously, with the most impressive being Henry Ford’s work on mass production in the US. Post-WWII Japan realized that its own auto industry would never be able to compete with more established Western companies, so it placed high tariffs on foreign cars, giving local companies like Nissan and Toyota a chance to get their act together. These companies, especially Toyota, invented a new form of auto production which was actually much more efficient than the usual American methods, and were eventually able to hold their own. They started exporting cars to the US; although American tariffs put them at a disadvantage, they were so much better than the American cars of the time that consumers preferred them anyway. After decades of losing out, the

American companies adopted a more Japanese ethos, and were eventually able to compete on a level playing field again.

This is a story of things gone surprisingly *right* – Americans and Japanese alike were able to get excellent inexpensive cars. Two things had to happen for it to work. First, Japan had to have high enough tariffs to give their companies some slack – to let them develop their own homegrown methods from scratch without being immediately outcompeted by temporarily-superior American competitors. Second, America had to have low enough tariffs that eventually-superior Japanese companies could outcompete American automakers, and Japan’s fitness-improving innovations could spread.

From the perspective of a Toyota manager, this is analogous to the eyeless alien story. You start with some good-enough standard (blind animals, American car companies). You want to evolve a superior end product (eye-having animals, Toyota). The intermediate steps (an animal with only Eye Part 1, a kind of crappy car company that stumbles over itself trying out new things) are less fit than the good-enough standard. Only when the inferior intermediate steps are protected from competition (through evolutionary randomness, through tariffs) can the superior end product come into existence. But you want to keep enough competition that the superior end product can use its superiority to spread (there is enough evolutionary competition that having eyes reaches fixation, there is enough free trade that Americans preferentially buy Toyota and US car companies have to adopt its policies).

From the perspective of an economic historian, maybe it’s a group selection story. The various stakeholders in the US auto industry – Ford, GM, suppliers, the government, labor, customers – competed with each other in a certain way and struck some compromise. The various stakeholders in the Japanese auto industry did the same. For some reason the American compromise worked worse than the Japanese one – I’ve heard stories about how US companies were more willing to defraud consumers for short-term profit, how US labor unions were more willing to demand concessions even at the cost of company efficiency, how regulators and executives were in bed with each other to the detriment of the product, etc. Every US interest group was acting in its own short-term self-interest, but the Japanese industry-as-a-whole outcompeted the American one and the Americans had to adjust.

3. Monopolies, Part II. Traditionally, monopolies have been among the most successful R&D centers. The most famous example is Xerox; it had a monopoly on photocopiers for a few decades before losing an anti-trust suit in the late 1970s; during that period, its PARC R&D program invented “laser printing, Ethernet, the modern personal computer, graphical user interface (GUI) and desktop paradigm, object-oriented programming, [and] the mouse”. The second most famous example is Bell Labs, which invented “radio astronomy, the transistor, the laser, the photovoltaic cell, the charge-coupled device, information theory, the Unix operating system, and the programming languages B, C, C++, and S” before the government broke up its parent company AT&T. Google seems to be trying [something similar](#), though it’s too soon to judge their outcomes.

These successes make sense. Research and development is a long-term gamble. Devoting more money to R&D decreases your near-term profits, but (hopefully) increases your future profits. Freed from competition, monopolies have limitless slack, and can afford to invest in projects that won’t pay off for ten or twenty years. This is part of Peter Thiel’s defense of monopolies in *Zero To One*.

An administrator tasked with advancing technology might be tempted to encourage monopolies in order to get more research done. But monopolies can also be stagnant and resistant to change; it’s probably not a coincidence that Xerox wasn’t the first company to bring the personal computer to market, and ended up irrelevant to the computing revolution. Like the eyeless aliens, who will not evolve in conditions of perfect competition or perfect lack of competition, probably all you can do here is strike a balance. Some

Communist countries tried the extreme solution – one state-supported monopoly per industry – and it failed the test of group selection. I don't know enough to have an opinion on whether countries with strong antitrust eventually outcompete those with weaker antitrust or vice versa.

4. Strategy Games. I like the strategy game *Civilization*, where you play as a group of primitives setting out to found a empire. You build cities and infrastructure, research technologies, and fight wars. Your world is filled with several (usually 2 to 7) other civilizations trying to do the same.

Just like in the real world, civilizations must decide between Guns and Butter. The Civ version of Guns is called the Axe Rush. You immediately devote all your research to discovering how to make really good axes, all your industry to manufacturing those axes, and all your population into wielding those axes. Then you go and hack everyone else to pieces while they're still futzing about trying to invent pottery or something.

The Civ version of Butter is called Build. You devote all your research, industry, and populace to laying the foundations of a balanced economy and culture. You invent pottery and weaving and stuff like that. Soon you have a thriving trade network and a strong philosophical tradition. Eventually you can field larger and more advanced armies than your neighbors, and leverage the advantage into even more prosperity, or into military conquest.

Consider a very simple scenario: a map of Eurasia with two civilizations, Rome and China.

If both choose Axe Rush, then whoever Axe Rushes better wins.

If both choose Build, then whoever Builds better wins.

What if Rome chooses Axe Rush, and China chooses Build?

Then it depends on their distance! If it's a very small map and they start very close together, Rome will probably overwhelm the Chinese before Build starts paying off. But if it's a very big map, by the time Roman Axemen trek all the way to China, China will have Built high walls, discovered longbows and other defensive technologies, and generally become too strong for axes to defeat. Then they can crush the Romans - who are still just axe-wielding primitives - at their leisure.

Consider a more complicated scenario. You have a map of Earth. The Old World contains Rome and China. The New World contains Aztecs. Rome and China are very close to each other. Now what happens?

Rome and China spend the Stone, Bronze, and Iron Ages hacking each other to bits. Aztecs spend those Ages building cities, researching technologies, and building unique Wonders of the World that provide powerful bonuses. In 1492, they discover Galleons and starts crossing the ocean. The powerful and advanced Aztec empire crushes the exhausted axe-wielding Romans and Chinese.

This is another story about slack. The Aztecs had it – they were under no competitive pressure to do things that paid off next turn. The Romans and Chinese didn't – they had to be at the top of their game every single turn, or their neighbor would conquer them. If there was an option that made you 10% weaker next turn in exchange for making you 100% stronger ten turns down the line, the Aztecs could take it without a second thought; the Romans and Chinese would probably have to pass.

Okay, more complicated *Civilization* scenario. This time there are two Old World civs, Rome and China, and two New World civs, Aztecs and Inca. The map is stretched a little bit so

that all four civilizations have the same amount of natural territory. All four players understand the map layout and can communicate with each other. What happens?

Now it's a group selection problem. A skillful Rome player will private message the China player and explain all of this to her. She'll remind him that if one hemisphere spends the whole Stone Age fighting, and the other spends it building, the builders will win. She might tell him that she knows the Aztec and Inca players, they're smart, and they're going to be discussing the same considerations. So it would benefit both Rome and China to sign a peace treaty dividing the Old World in two, stick to their own side, and Build. If both sides cooperate, they'll both Build strong empires capable of matching the New World players. If one side cooperates and the other defects, it will easily steamroll over its unprepared opponent and conquer the whole Old World. If both sides defect, they'll hack each other to death with axes and be easy prey for the New Worlders.

This might be true in *Civilization* games, but real-world civilizations are more complicated. Graham Greene wrote:

In Italy, for thirty years under the Borgias, they had warfare, terror, murder and bloodshed, but they produced Michelangelo, Leonardo da Vinci and the Renaissance. In Switzerland, they had brotherly love, they had five hundred years of democracy and peace - and what did that produce? The cuckoo clock.

So maybe a little bit of internal conflict is good, to keep you honest. Too much conflict, and you tear yourselves apart and are easy prey for outsiders. Too little conflict, and you invent the cuckoo clock and nothing else. The continent that conquers the world will have enough pressure that its people want to innovate, and enough slack that they're able to.

This is total ungrounded amateur historical speculation, but when I hear that I think of the Classical world. We can imagine it as divided into a certain number of "theaters of civilization" - Greece, Mesopotamia, Egypt, Persia, India, Scythia, etc. Each theater had its own rules governing average state size, the rules of engagement between states, how often bigger states conquered smaller states, how often ideas spread between states of the same size, etc. Some of those theaters were intensely competitive: Egypt was a nice straight line, very suited to centralized rule. Others had more slack: it was really hard to take over all of Greece; even the Spartans didn't manage. Each theater conducted its own "evolution" in its own way - Egypt was ruled by a single Pharaoh without much competition, Scythia was constant warfare of all against all, Greece was isolated city-states that fought each other sometimes but also had enough slack to develop philosophy and science. Each of those systems did their own thing for a while, until finally one of them produced something perfect: 4th century BC Macedonia. Then it went out and conquered everything.

If Greene is right, the point isn't to find the ruleset that promotes 100% cooperation. It's to find the ruleset that promotes an evolutionary system that makes your group the strongest. Usually this involves some amount of competition - in order to select for stronger organisms - but also some amount of slack - to let organisms develop complicated strategies that can make them stronger. Despite the earlier description, this isn't necessarily a slider between 0% competition and 100% competition. It could be much more complicated - maybe alternating high-slack vs. low-slack periods, or many semi-isolated populations with a small chance of interaction each generation, or alternation between periods of isolation and periods of churning.

In a full two-layer evolution, you would let the systems evolve until they reached the best parameters. Here we can't do that - Greece has however many mountains it has; its success does not cause the rest of the world to grow more mountains. Still, we randomly started with enough different groups that we got to learn something interesting.

(I can't emphasize enough how ungrounded this historical speculation is. Please don't try to evolve Alexander the Great in your basement and then get angry at me when it doesn't work)

5. The Long-Term Stock Exchange. Actually, all stock exchanges are about slack. Imagine you are a brilliant inventor who, given \$10 million and ten years, could invent fusion power. But in fact you have \$10 and need work tomorrow or you will starve. Given those constraints, maybe you could start, I don't know, a lemonade stand.

You're in the same position as the animal trying to evolve an eye - you could create something very high-utility, if only you had enough slack to make it happen. But by default, the inventor working on fusion power starves to death ten days from now (or at least makes less money than his counterpart who ran the lemonade stand), the same way the animal who evolves Eye Part 1 gets outcompeted by other animals who didn't and dies out.

You need slack. In the evolution example, animals usually stumble across slack randomly. You too might stumble across slack randomly - maybe it so happens that you are independently wealthy, or won the lottery, or something.

More likely, you use the investment system. You ask rich people to give you \$10 million for ten years so you can invent fusion; once you do, you'll make trillions of dollars and share some of it with them.

This is a great system. There's no evolutionary equivalent. An animal can't pitch Darwin on its three-step plan to evolve eyes and get free food and mating opportunities to make it happen. Wall Street is a giant multi-trillion dollar time machine funneling future profits back into the past, and that gives people the slack they need to make the future profits happen at all.

But the [Long-Term Stock Exchange](#) is especially about slack. They are a new exchange (approved by the SEC last year) which has complicated rules about who can list with them. Investors will get extra clout by agreeing to hold stocks for a long time; executives will get incentivized to do well in the far future instead of at the next quarterly earnings report. It's making a deliberate choice to give companies more slack than the regular system and see what they do with it. I don't know enough about investing to have an opinion, except that I appreciate the experiment. Presumably its companies will do better/worse than companies on the regular stock exchange, that will cause companies to flock toward/away from it, and we'll learn that its new ruleset is better/worse at evolving good companies through competition than the regular stock exchange's ruleset.

6. That Time Ayn Rand Destroyed Sears. Or at least that's how Michael Rozworski and Leigh Phillips describe Eddie Lampert's corporate reorganization in [How Ayn Rand Destroyed Sears](#), which I recommend. Lampert was a Sears CEO who figured - since free-market competitive economies outcompete top-down economies, shouldn't free-market competitive companies outcompete top-down companies? He reorganized Sears as a set of competing departments that traded with each other on normal free-market principles; if the Product Department wanted its products marketed, it would have to pay the Marketing Department. This worked really badly, and was one of the main contributors to Sears' implosion.

I don't have a great understanding of exactly why Lampert's Sears lost to other companies, but capitalist economies beat socialist ones; Rozworski and Phillips' [People's Republic Of Wal-Mart](#), which looks into this question, is somewhere on my reading list. But even without complete understanding, we can use group selection to evolve the right parameters. Imagine an economy with several businesses. One is a straw-man communist collective, where every worker gets paid the same regardless of output and there are no

promotions (0% competition, 100% cooperation). Another is Lampert's Sears (100% competition, 0% cooperation). Others are normal businesses, where employees mostly work together for the good of the company but also compete for promotions (X% competition, Y% cooperation). Presumably the normal business outcompetes both Lampert and the commies, and we sigh with relief and continue having normal businesses. And if some of the normal businesses outcompete others, we've learned something about the best values of X and Y.

7. Ideas. These are in constant evolutionary competition – this is the insight behind [memetics](#). The memetic equivalent of slack is inferential range, aka “willingness to entertain and explore ideas before deciding that they are wrong”.

[Inferential distance](#) is the number of steps it takes to make someone understand and accept a certain idea. Sometimes inferential distances can be very far apart. Imagine trying to convince a 12th century monk that there was no historical Exodus from Egypt. You're in the middle of going over archaeological evidence when he objects that the Bible says there was. You respond that the Bible is false and there's no God. He says that doesn't make sense, how would life have originated? You say it evolved from single-celled organisms. He asks how evolution, which seems to be a change in animals' accidents, could ever affect their essences and change them into an entirely new species. You say that the whole scholastic worldview is wrong, there's no such thing as accidents and essences, it's just atoms and empty space. He asks how you ground morality if not in a striving to approximate the ideal embodied by your essence, you say...well, it doesn't matter what you say, because you were trying to convince him that some very specific people didn't leave Egypt one time, and now you've got to ground morality.

Another way of thinking about this is that there are two self-consistent equilibria. There's your equilibrium, (no Exodus, atheism, evolution, atomism, moral nonrealism), and the monk's equilibrium (yes Exodus, theism, creationism, scholasticism, teleology), and before you can make the monk budge on any of those points, you have to convince him of all of them.

So the question becomes – how much patience does this monk have? If you tell him there's no God, does he say “I look forward to the several years of careful study of your scientific and philosophical theories that it will take for that statement not to seem obviously wrong and contradicted by every other feature of the world”? Or does he say “KILL THE UNBELIEVER”? This is inferential range.

Aristotle says that the mark of an educated man is to be able to entertain an idea without accepting it. Inferential range explains why. The monk certainly shouldn't immediately accept your claim, when he has countless pieces of evidence for the existence of God, from the spectacular faith healings he has witnessed (“look, there's this thing called psychosomatic illness, and it's really susceptible to this other thing called the placebo effect...”) to Constantine's victory at the Mulvian Bridge despite being heavily outnumbered (“look, I'm not a classical scholar, but some people are just really good generals and get lucky, and sometimes it happens the day after they have weird dreams, I think there's enough good evidence the other way that this is not the sort of thing you should center your worldview around”). But if he's willing to entertain your claim long enough to hear your arguments one by one, eventually he can reach the same self-consistent equilibrium you're at and judge for himself.

Nowadays we don't burn people at the stake. But we do make fun of them, or flame them, or block them, or wander off, or otherwise not listen with an open mind to ideas that strike us at first as stupid. This is another case where we have to balance competition vs. slack. With perfect competition, the monk instantly rejects our “no Exodus” idea as less true (less memetically fit) than its competitors, and it has no chance to grow on him. With zero competition, the monk doesn't believe anything at all, or spends hours patiently listening

to someone explain their world-is-flat theory. Good epistemics require a balance between being willing to choose better ideas over worse ones, and open-mindedly hearing the worse ones out in case they grow on you.

([Thomas Kuhn](#) points out that early versions of the heliocentric model were much worse than the geocentric model, that astronomers only kept working on them out of a sort of weird curiosity, and that it took decades before they could clearly hold their own against geocentrism in a debate).

Different people strike a different balance in this space, and those different people succeed or fail based on their own epistemic ruleset. Someone who's completely closed-minded and dogmatic probably won't succeed in business, or science, or the military, or any other career (except maybe politics). But someone who's so pathologically open-minded that they listen to everything and refuse to prioritize what is or isn't worth their time will also fail. We take notice of who succeeds or fails and change our behavior accordingly.

Maybe there's even a third layer of selection; maybe different communities are more or less willing to tolerate open-minded vs. close-minded people. The Slate Star Codex community has really different epistemic norms from the Catholic Church or Infowars listeners; these are evolutionary parameters that determine which ideas are more memetically fit. If our epistemics make us more likely to converge on useful (not necessarily true!) ideas, we will succeed and our epistemic norms will catch on. Francis Bacon was just some guy with really good epistemic norms, and now everybody who wants to be taken seriously has to use his norms instead of whatever they were doing before. Come up with the right evolutionary parameters, and that could be you!

Baking is Not a Ritual

I started baking about 2 years ago. Since I became a frequent supplier of baked goods in the office, a lot of people have come to me for baking advice. I've noticed a trend of comments about baking that all share a common root cause.

See if you can spot the common failure mode that led to these very-paraphrased comments:

- "Baking is too precise for me. I want to bake without following recipes exactly, but I feel like I can't improvise."
- "I tried making this. I left out ingredient Y because I didn't have it and the recipe only needed a little bit of it. Why didn't it work out?"
- "I tried doing step X for exactly N minutes this time and that worked well. Oh, you're saying that duration doesn't matter? Well, it worked for me so I think I'll just keep doing it just in case."
- "I always have to repeat a new recipe a bunch before I stop ruining it every other time."

The misconception that leads to these comments is **treating a baking recipe like a ritual and blindly following it, without attempting to understand the baking process at a gears-level.**

Many people seem to approach baking like it is a ritual, where one follows a recipe exactly and some magic happens to produce the baked goods. Things will go mysteriously wrong if you stirred the cauldron counter-clockwise or added the water too early. In reality, baking is combining and heating up a series of ingredients so they undergo physical and chemical reactions to achieve certain texture and taste. There are underlying principles that govern the process of baking and the results.

Looking at a recipe and following the steps works fine a lot of the time, if a recipe is good and has the right details. However, if one treats the baking process as a black box ritual, without understanding the underlying mechanisms, one can run into troubles, such as:

- Unable to improvise or change a recipe
- Not knowing which parts of the recipe matter, i.e. have a strong effect on the end-results. Some people end up compensating for this by just trying to do everything super precisely, even when some of the steps don't make any difference.
- Unable to react when something goes wrong, like a missing ingredient, or a mistake in an earlier step.

The right way to approach baking is to realize it is not a ritual. Instead, try to understand the principles of how baking works, to understand why an ingredient is in the recipe, and why a particular step is needed. Some examples of gears-level principles are:

- Acidity interacts with baking soda to create bubbles, so don't leave out lemon juice in a recipe that calls for baking soda
- Kneading a wheat dough folds and so strengthens the gluten, which makes the end product more chewy, which is commonly desirable in bread but not in cakes or biscuits
- Eggs acts as an emulsifier to help combine water and oil ingredients. Don't skip it in recipes where an emulsifier is needed, but they're optional in recipes that don't need an emulsifier.
- A wetter dough rises more, so add more liquid ingredients if you want a fluffier result, and less if you prefer a denser version.

Understanding these underlying principles can make recipes flexible. One can easily make tweaks if one knows why each ingredient is needed and how it affects the final result. For

instance, [this apple bread](#) is one of my staple recipes. I once baked it while I was short one egg, but I knew it was not a key ingredient in this recipe (i.e. it did not act as an emulsifier), so I compensated by adding some extra butter and some yogurt to make up for the missing egg's fat and liquid content - it turned out just fine. I've also adapted this recipe to use cherries instead of apples, because I know the fruit part can be fully swapped out.



Cherry bread adapted from an [apple bread recipe](#)

Understanding the baking process also means knowing which steps of the process is important, and which are not. This lets one focus on key parts, but be "lazier" with parts that either do not matter or can be easily adjusted later. For instance, the exact amount of vanilla extract doesn't make a difference in my recipe above, so instead of dirtying yet another spoon to measure exactly $\frac{1}{4}$ teaspoon of vanilla extract, I just give the bottle a squirt and call it a day. Another example, I know that additional flour can be easily added when kneading a yeast dough, so while many people swear by precisely measuring flour by weight, I can be lazy and approximate when measuring out flour by erring on the side of adding less to start, then sprinkle in more as needed.



Yeast bread, after kneading and the final product

On the other hand, mixing in cold, pea-sized butter is important for achieving the flaky crumbly texture of biscuits, so even though it's more work, I grate my butter and take care to keep it cold throughout, sometimes even running my hands under freezing water before working with the biscuit dough.



Cold, pea-sized chunks of butter in the dough is crucial to making biscuits flaky and crumbly, so don't take the easy way out by melting the butter or substituting with canola or olive oil.

Understanding the baking process can help one understand new recipes, because many recipes share the same underlying principles. If it's a recipe for something similar to baked goods I'm familiar with, I can often evaluate it at a glance and draw conclusions like "oh this step is probably unnecessary" or "I don't have X but I can substitute with Y". My friends find it helpful to run a new recipe by me before they begin, as I can often highlight key steps to them and give advice even if I've never used that recipe.

Realizing that baking is not a ritual and that there are underlying principles is often sufficient for people to seek out these principles and improve. One additional tip is, when learning to make something completely new, don't try to perfectly follow one recipe. Instead, look at multiple recipes for the same item. Many recipes on the internet are accompanied by blog posts and comments. These often contain tips and advice at the gears-level and give insights into why a certain amount of an ingredient is needed, and how a certain step would affect the outcome. Paying attention to not only the recipe but also reading through these advice when learning to bake something new allows one to have a much greater success rate, even on the very first attempt.



I was challenged to attempt a souffle. Not perfect, but it did rise on my first try after I researched extensively on how beating and folding in the egg whites makes the souffle airy.

In conclusion, many people I talked to seem to believe baking is a ritual, where you have to follow recipes exactly to be successful. They never open the blackbox and therefore lack the understanding of baking at a gears-level. When one grasps that baking is not a ritual and learns the principles behind the ingredients and the steps in baking, one can easily make adjustments, adapt recipes, and be more successful.

A non-mystical explanation of insight meditation and the three characteristics of existence: introduction and preamble

Introduction

Insight meditation, enlightenment, what's that all about?

The sequence of posts starting from this one is my personal attempt at answering that question. It grew out of me being annoyed about so much of this material seeming to be straightforwardly explainable in non-mysterious terms, but me also being unable to find any book or article that would do this to my satisfaction. In particular, I wanted something that would:

- Explain what kinds of implicit assumptions build up our default understanding of reality and how those assumptions are subtly flawed. It would then point out aspects from our experience whose repeated observation will update those assumptions, and explain how this may cause psychological change in someone who meditates.
- It would also explain how the so-called "[three characteristics of existence](#)" of Buddhism - impermanence, no-self and unsatisfactoriness - are all interrelated and connected with each other in a way your average Western science-minded, allergic-to-mysticism reader can understand.

I failed to find a resource that would do this in the way I had in mind, so then I wrote one myself.

From the onset, I want to note that I am calling this a non-mystical take on the three characteristics, rather than *the* non-mystical take on the three characteristics. This is an attempt to explain what I personally think is going on, and to sketch out an explanation of how various experiences and Buddhist teachings *could* be understandable in straightforward terms. I don't expect this to be anything like a complete or perfect explanation, but rather one particular model that might be useful.

The main intent of this series is summarized by a [comment written by Vanessa Kosoy](#), justifiably skeptical of grandiose claims about enlightenment that are made without further elaboration on the actual mechanisms of it:

I think that the only coherent way to convince us that Enlightenment is real is to provide a model from a 3rd party perspective. [...] The model doesn't have to be fully mathematically rigorous: as always, it can be a little fuzzy and informal. However, it must be precise enough in order to (i) correctly capture the essentials and (ii) be interpretable more or less unambiguously by the sufficiently educated reader.

Now, having such a model doesn't mean you can actually reproduce Enlightenment itself. [...] However, producing such a model would give us the

enormous advantages of (i) being able to come up with experimental tests for the model (ii) understanding what sort of advantages we would gain by reaching Enlightenment (iii) being sure that you are talking about something that is at least a coherent possible world even if we are still unsure whether you are describing the actual world.

I hope to at least put together a starting point for a model that would fulfill those criteria.

Note that these articles are *not* saying “you should meditate”. Getting deep in meditation requires a huge investment of time and effort - though smaller investments are also likely to produce benefits - and is associated with its own risks [[1](#) [2](#) [3](#) [4](#)]. My intent is merely to discuss some of the mechanisms involved in meditation and the mind. Whether one should get direct acquaintance with them is a separate question that goes beyond the scope of this discussion.

Briefly on the mechanisms of meditation

In a previous article, [A Mechanistic Model of Meditation](#), I argued that it is possible in principle for meditation to give people an improved understanding of the way their mind operates.

To briefly recap my argument: we know it is possible for people to train their senses, such as learning to notice more details or make more fine-grained sensory discriminations. One theory is that those details have always been processed in the brain, but the information has not made it to the higher stages of the processing hierarchy. As you repeatedly focus your attention to a particular kind of pattern in your consciousness, neurons re-orient to strengthen that pattern and build connections to the lower-level circuits from which it emerges. This re-encodes the information in those circuits in a format which can be represented in consciousness.

This means at least some kinds of sensory training are *training in introspection* - learning to better access information which already exists in your brain. This implies you can also learn to strengthen *other* patterns in your consciousness, especially if you have some source of feedback that you can use to guide the training.

I gave an example of experiential forms of therapy doing exactly this, and then described how a particular style of meditation used one’s awareness of the breath as an objective feedback signal for developing increased “introspective awareness” of one’s own mind.

That post was mostly describing the ways in which meditation can be used to become more aware of the *content* of your thoughts. However, in observing the content, it is hard to avoid noticing at least some of the *structure* of the thought process as well.

For example, you might try to follow your breath and think you are doing a good job. In this case, there are at least two kinds of content in your mind: the actual sensory experience of the breath, and thoughts about how badly or well you are doing. The latter might take the form of e.g. mental dialogue that says things like “I’m still managing to follow my breath”. Now, since you may find it rewarding to just think that you are meditating well, *that thought* may start to become rewarded, and you may find yourself repeatedly *thinking* that you are successfully following the breath... even

as the thought of "I am meditating well" has become self-sustaining and no longer connected to whether you are following the breath or not.

Eventually you will realize that you have actually been *thinking about following the breath* rather than *actually following it*. This is a minor insight into the way that your thought processes are structured, revealing it is possible for sensations and thoughts about sensations to become mixed up.

It is also possible to practice meditation in a way which explicitly focuses on investigating structure. We can make an analogy to looking at a painting. (Thanks to Alexei Andreev for suggesting this analogy.) Seen from some distance, a painting has "content": it depicts things like people, buildings, boats and so forth. But when you get closer to it and look carefully, you can see that all the content is composed of things like brush strokes, individual colored shapes, paint of varying thickness, and so on. This is "structure". While all types of meditation are going to reveal *something* about structure, there are also types of meditation which are specifically aimed at exploring it. Meditation which focuses on investigating structure is commonly called *insight* meditation.

Investigating the mind vs. investigating reality

Now, it is worth noting that these practices are not always framed in terms of "investigating the structure of the mind", nor does the actual experience of doing them necessarily feel like that. Rather, the framing and experience is commonly that of investigating the nature of *reality*.

For example, in [an earlier article](#) trying to explain insight meditation, I mentioned I had once had the thought that I could never be happy. When I paid closer attention to why I thought that, I noticed that my mental image of a happy person included strong extraversion, which conflicted with the self-image that I had of myself as an introvert. After I noticed the happiness-extraversion connection, it became apparent that I could be happy even as an introvert, and the original thought disappeared. (Although I didn't know it at the time, [it is common for emotional beliefs to change](#) when they become explicit enough for the brain to notice them being erroneous.)

Essentially, I had originally believed "I can never be happy", and this belief about me didn't feel like a "belief". It felt like a *basic truth of what I was*, the kind of truth that you just know - in the same way that you might look at an apple and *just know* you are having the experience of seeing an apple. But when I investigated the details of that experience, I realized that this wasn't actually a fact about me. Rather it was just a belief that I had.

In a similar way, there are many aspects of our subjective experience that feel like facts about reality, but upon doing insight practices and investigating them closer, we can come to see that they are not so.

The philosopher Daniel Dennett has coined the term "[heterophenomenology](#)" to refer to a particular approach to the study of consciousness. In this approach, we assume that people are correctly describing how things *seem* to them and treat this as something that needs to be explained. However, the actual mechanism of why things seem like that to them, may be different from what they assume.

If I see an apple, it typically feels to me like I am seeing reality as it is. From a scientific point of view, this is mistaken: the sight of an apple is actually a complex interpretation my brain has created. Likewise, if I have the experience that I can never be happy, then this also feels like a raw fact while actually being an interpretation. In either case, if I manage to do practices which reveal my interpretation to be flawed, they will subjectively feel like I am investigating reality... while from a third-person perspective, we would rather say that I am investigating the way my mind builds up reality.

It is valid to stick to just the first-person experience of investigating reality directly. Many of these practices are framed solely in those terms, because a stance of curiosity and having as few assumptions as possible is the best mindset for actually doing the practices. But if one says that meditation investigates the nature of reality, then it becomes hard to test the claim from a third-person perspective. A common criticism is that meditation certainly *changes* how people experience the world, but it might just as well be *loosening* their grasp on reality.

On the other hand, if we provisionally assume that meditation works by revealing how the mind structures its model of reality, then we can check whether the kinds of insights that people report are compatible with what science tells us about the brain. If it turns out that meditators doing insight practices are coming up with experiences that match our understanding of actual brain mechanisms, then the practices might actually provide insight rather than delusion. In cases where no scientific evidence is yet available, it should at least be possible to construct a model that *could* be true and compatible with the third-person evidence.

In [previous posts](#), I have explored some scientifically-informed models of the brain, which I think are naturally linked to the kinds of discoveries made in insight meditation. This article will more explicitly connect concepts from the theory of meditation to those kinds of models.

It is also worth noting that I think *both* claims about meditative insights are true: some things you can do with meditation *do* give you a better insight into reality, while some other things *do* just break your brain and reduce your contact with reality. (A fact responsible meditation teachers [also warn about](#).) This makes it important to have third-person models of what could be a genuine insight and what is probably delusion, to help avoid the dangerous territory.

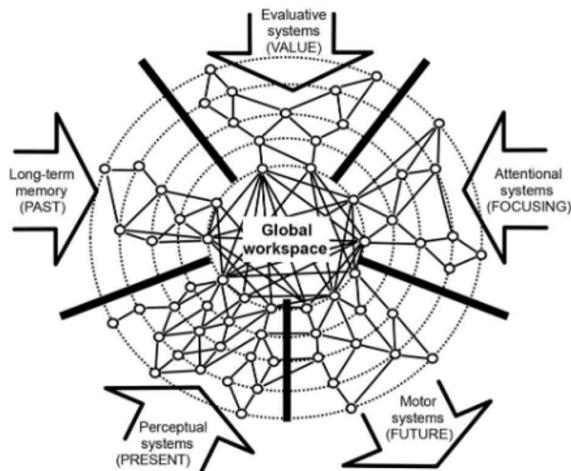
My multiagent model of mind

I have been calling my interpretation of those models a “[multiagent model of mind](#)”. What follows is a highly abridged version of it; see the linked index of posts for much more extensive discussion, including the sources that I have been drawing on for my synthesis.

One of the main ideas of the multiagent model is that the brain contains a number of different subsystems operating in parallel, each focusing on their own responsibilities. They share information on a subconscious level, but also through conscious thought. The content of consciousness [roughly corresponds](#) to information which is being processed in a “global workspace” - a “brain web” of long-distance neurons, which link multiple areas of the brain together into a densely interconnected network.

The global workspace can only hold a single piece of information at a time. At any given time, multiple different subsystems are trying to send information into the workspace, or otherwise modify its contents. Experiments show that a visual stimuli needs to be shown for about 50 milliseconds for it to be consciously registered, suggesting that the contents of consciousness might be updated at least 20 times per second. Whatever information makes it into consciousness will then be broadcast widely throughout the brain, allowing many subsystems to synchronize their processing around it.

The exact process by which this happens is not completely understood, but involves a combination of top-down mechanisms (e.g. attentional subsystems trying to strengthen particular signals and keep those in the workspace) as well as bottom-up ones (e.g. emotional content getting a priority). For example, if you are listening to someone talk in a noisy restaurant, both their words and the noise are bottom-up information within the workspace, while a top-down process tries to pick up on the words in particular. If a drunk person then suddenly collides with you, you are likely to become startled, which is a bottom-up signal strong enough to grab your attention (dominate the workspace), momentarily pushing away everything else.



There is also a constant learning process going on, where the brain learns which subsystems should be given access in which circumstances, while the subsystems themselves also undergo learning about what kind of information to send to consciousness.

When I talk about “subsystems” sending content into consciousness, I mean this as a very generic term, which includes all of the following:

- Literal subsystems, e.g. information from the visual, auditory, and other sensory systems
- Subpatterns within larger subsystems, e.g. a particular neuronal pattern encoding a specific memory or habit
- [Emotional schemas](#) which trigger in particular situations and contain an interpretation of that situation and a response
- Working memory buffers [associated with type 2 \(“System 2”\)](#) reasoning, helping chain the outputs of several different subsystems together

In some cases, I might talk about there being two separate subsystems, when one could argue that this would be better described as something like two separate pieces of data within a single subsystem. For example, I might talk about two different memories as two different subsystems, when one could reasonably argue that they are both contained within the same memory subsystem. Drawing these kinds of distinctions within the brain seems tricky, so rather than trying to figure out what term to use when, I will just talk about subsystems all the time.

Epistemic status

Buddhist theories of the mind are based on textual traditions that purport to record the remembered word of the Buddha, on religious and philosophical interpretations of those texts, and on Buddhist practices of mental cultivation. The theories aren't formulated as scientific hypotheses and they aren't scientifically testable. Buddhist insights into the mind aren't scientific discoveries. They haven't resulted from an open-ended empirical inquiry free from the claims of tradition and the force of doctrinal and sectarian rhetoric. They're stated in the language of Buddhist metaphysics, not in an independent conceptual framework to which Buddhist and non-Buddhist thinkers can agree. Buddhist meditative texts are saturated with religious imagery and language. Buddhist meditation isn't controlled experimentation. It guides people to have certain kinds of experiences and to interpret them in ways that conform to and confirm Buddhist doctrine. The claims that people make from having these experiences aren't subject to independent peer review; they're subject to assessment within the agreed-upon and unquestioned framework of the Buddhist soteriological path. [...]

I'm not saying that Buddhist meditative techniques haven't been experientially tested in any sense. Meditation is a kind of skill, and it's experientially testable in the way that skills are, namely, through repeated practice and expert evaluation. I have no doubt that Buddhist contemplatives down through the ages have tested meditation in this sense. I'm also not saying that meditation doesn't produce discoveries in the sense of personal insights. (Psychoanalysis can also lead to insights.) Rather, my point is that the experiential tests aren't experimental tests. They don't test scientific hypotheses. They don't provide a unique set of predictions for which there aren't other explanations. The insights they produce aren't scientific discoveries. [...]

I'm also not trying to devalue meditation. On the contrary, I'm trying to make room for its value by showing how likening it to science distorts it. Meditation isn't controlled experimentation. Attention and mindfulness aren't instruments that reveal the mind without affecting it. Meditation provides insight into the mind (and body) in the way that body practices like dance, yoga, and martial arts provide insight into the body (and mind). Such mind-body practices—meditation included—have their own rigor and precision. They test and validate things experientially, but not by comparing the results obtained against controls.

-- Evan Thompson, [*Why I Am Not A Buddhist*](#)

I think it is reasonable to believe that meditation can give us genuine insights into the way the mind functions. The meditative techniques and practices which I am drawing upon in this series have been developed within Buddhist traditions, and I make frequent references to the theory developed within those traditions.

At the same time, while I am drawing upon theories developed within these traditions, I am treating those as a source of inspiration to be critically examined, rather than as sources of authority.

For one, there are many different Buddhist theories and schools that disagree with each other, many of them claiming to teach [what the Buddha really meant](#). And as e.g. Evan Thompson's book discusses, one cannot cleanly separate Buddhist meditative techniques from Buddhist religious teaching. People who meditate using those techniques - myself included - do so while being guided by an existing conceptual framework, framing their experiences in light of their framework. Practitioners who use different kinds of techniques and frameworks end up drawing different conclusions: e.g. some frameworks end up at the conclusion that no selves exist, while others end up believing that everything is self. (The extent to which this difference in framing actually leads to a different *experience* is unclear.) Many of these frameworks also draw upon supernatural elements, such as claims of rebirth and remembering past lives.

Still, many meditation teachers also say things along the lines of "you should not take any of this on faith, just try it out and see for yourself". Personally I started out skeptical of many claims, dismissing them as pre-scientific folk-psychological speculation, before gradually coming to believe in them - sometimes as a result of meditation which hadn't even been aimed at investigating those claims in particular, but where I thought I was doing something completely different. And it seems to me that many of the meditative techniques actively require you to *suspend your expectations* in order to work properly, requiring you to look at what's present rather than at the thing you expect to see.

So, like many others, I simultaneously believe that i) meditative techniques point at genuine insights and also produce them in the minds of people who meditate and also that ii) we should not put excess faith in the claims of the existing meditation traditions. As many teachers encourage exactly this line of thought - as in the comment of taking nothing on faith - this feels like an appropriate spirit for approaching these matters.

Rather than trying to be authentically Buddhist, this article is concerned with building a model of the neural and psychological mechanisms I think the three characteristics are pointing at, even if that model ends up sharply deviating from the original theories. I heavily draw on my own experiences and the experiences and theories of other people whose reports I have reason to trust. I proceed from the assumption that regardless of whether the original frameworks are true or false, they do systematically produce similar effects and insights in the minds of the people following them, and that is an [observation which needs to be explained](#).

In fact, I am happy to mix and match examples, exercises, interpretations and results drawn from all of the contemplative traditions that I happen to have any familiarity with, with current-day Western psychology and psychotherapy thrown in for good measure. They may have different approaches, but to the extent that they share commonalities, those commonalities tell us something about what human minds might have in common. And to the extent that they differ, one tradition might be pointing out aspects about the human mind that the others have neglected and vice versa, as in the fable of the blind men and the elephant.

Current articles in this series:

- Introduction and preamble (you are here)
- [A non-mystical explanation of "no-self"](#)
- [Craving, suffering, and predictive processing](#)
- [From self to craving](#)
- [On the construction of the self](#)
- [Impermanence](#)

Thank you to Alexei Andreev, David Chapman, Eliot Re, Jacob Spence, James Hogan, Magnus Vinding, Max Daniel, Matthew Graves, Michael Ashcroft, Romeo Stevens, Santtu Heikkinen, and Vojtěch Kovařík for valuable comments. Additional special thanks to Maija Haavisto. None of the people in question necessarily agree with all the content in this or the upcoming posts; much of the content has also been rewritten after the drafts that most of them saw.

Tips/tricks/notes on optimizing investments

I've been optimizing various aspects of my investment setup recently, and will write up some tips and tricks that I've found in the form of "answers" here. Others are welcome to share their own here if they'd like. (Disclaimer: I'm not a lawyer, accountant, or investment advisor, and everything here is for general informational purposes only.)

What are your greatest one-shot life improvements?

Sometimes, people have life problems that can be entirely solved by doing one thing. (doing X made my life 0.1% better, PERMANENTLY!) These are not things like "This TAP made me exercise more frequently", but rather like "moving my scale into my doorway made me weigh myself more, causing me to exercise more frequently" - a one-shot solution that makes a reasonable amount of progress in solving a problem.

I've found that I've had a couple of life problems that I couldn't solve because I didn't know what the solution was, not because it was hard to solve - once I *thought* of the solution, implementation was not that difficult. I'm looking to collect various one-shot solutions to problems to expand my solution space, as well as potentially find solutions to problems that I didn't realize I had.

Please only put one problem-solution pair per answer.

Mazes Sequence Summary

This post attempts to summarize the key points of the Immoral Mazes Sequence, which begins [here](#), so they can be referenced without asking readers to get through an entire book first.

Due to the change in format, the posts will be summarized slightly out of order.

Note that this summary, and especially the summary of the summary, represent a not only an abridged and simplified but also *sanitized* version of the central points. Brains do their best to continuously *round down* and *not fully* see these concepts.

Core Ideas (Summary of the Summary)

The book [Moral Mazes](#), by Robert Jackall, is a detailed exploration of middle manager hell. Managers must abandon all other goals and values, in favor of spending all their time and resources on manipulations of the system. They must learn to view such actions as intrinsically good and worthy of reward. Only those who let this process entirely consume them can survive.

The Immoral Mazes sequence is an exploration of what causes that hell, and how and why it has spread so widely in our society. Its thesis is that this is the result of a vicious cycle arising from competitive pressures among those competing for their own organizational advancement. Over time, those who focus more on and more value such competitions win them, gain power and further spread their values, unless they are actively and continuously opposed.

Once things get bad in an organization they tend to only get worse, but things *in general* get better because such organizations then decay and are replaced by new ones. Unfortunately, our society now slows or prevents that process, with these same organizations and their values increasingly running the show.

Investment and flexibility become impossible. Even appearing to care about anything except the competition itself costs you your allies. Thus things inevitably decay and then collapse, flexibility returns, cycle repeats.

Involvement with such patterns is far more destructive to humans than is commonly known. Employment or other involvement with such patterns should be avoided or ended, even at seemingly high cost in money and superficial status. If one wishes to accomplish something other than competition for advancement, one must be vigilant to punish maze-promoting behaviors, and to keep out or cast out those whose values align with mazes.

This is a special case of the principle that sufficiently intense competition along a single axis destroys all. Exit costs to any situation are almost always non-zero, so situations that otherwise look like they are ‘perfectly competitive’ are instead what I dub ‘super-perfectly competitive,’ where profits are negative, and all participants unfortunate enough to have entered continuously lose ground and eat their seed corn.

Eventually such systems collapse and are replaced. This is why the world has nice things and often greatly improves over time, even when specific things seem to mostly tangibly decay.

Dangers of Perfect Competition

A long time ago, Scott Alexander wrote [Meditations on Moloch](#). Moloch represents the inevitability of competitive selection pressure, given a long enough time horizon, completely binding all behavior, and thus destroying all value, then all life.

This hasn't fully happened yet. But it will. Unless we use this one opportunity offered to us by our technological progress before that happens, and stop it.

A less long time ago, I introduced the (pre-existing) concept of [Slack](#). Slack is the absence of binding constraints on behavior. Metaphorically, slack is life. Literally, slack is also life. Without slack, life is unable to both survive and retain the ability to adapt, and thus loses one, then the other, and dies.

Recently, Scott wrote [Studies on Slack](#), which made a lot of this more explicit and easier to understand, especially the point that slack is life.

A few months ago, I wrote the book-length Immoral Mazes Sequence. This was my attempt to understand the facts described in the important and very hard to get through book (not because it's badly written, but because the things in it are hard to look at) [Moral Mazes](#). The book describes the lives and experiences of middle managers in major American corporations.

Why are things so bad for these managers? Why aren't they as bad for everything and everyone else?

[The Molochian nightmare is not how most of the world works](#). We have nice things. Over time they improve. Systems do not automatically collapse to the elimination of anything that shows the slightest inefficiency. Most places and times and groups get a lot of genuine cooperation towards worthy goals, and avoid or mitigate any race to the bottom. Our lives are full of slack. The question is why.

The reason is because perfectly binding constraints need only arise in the presence of [perfect competition](#). If *and only if* everyone and everything is identical and standardized, all constraints must bind and behavior must be only that which maximizes short-term measured competitive results. Even worse, there is what I called super-perfect competition. Perfect competition allows everyone to break even (make zero economic profits), but that is because participants are allowed to exit. Super-perfect competition has costly exit (and in reality, all exit has some cost), and thus everyone competes with everything they have *and still loses*. The "good" news is that such systems, when they arise, over-compete and over-optimize, eat their seed corn, and thus quickly collapse.

[The cycle repeats](#). It repeats with the rise and fall of civilizations, and also the rise and fall of individual groups and corporations. Large organizations are doomed. We need them, but should be cautious about creating things that are inevitably doomed. Over time, things almost always get locally worse in these ways rather than better, until it is disrupted from outside and replaced by the new. We thus should not weep too much when this occurs.

Our society has chosen, to a large extent that has been laid even more bare by recent events, not to allow these failures and this renewal. Hence the term ‘Too Big to Fail.’ Our institutions of all kinds are becoming increasingly dysfunctional, increasingly concerned with politics in its broad sense, and incapable of useful action. Thus, people increasingly have the instinct that Moloch inevitably wins everywhere. We are disarming the forces that keep Moloch in check. And Moloch wants us to think its victory is inevitable, so we will actively support it rather than oppose it, so we can form an implicit alliance with others doing the same, in the hope the process will kill us last.

Mazes as Super-Perfectly Competitive Battle Between Managers

Despite this trend, when we look around at markets in practice, we instead mostly see [highly imperfect competition](#) on lots of different levels. Behaviors are a long way from fully optimized *for anything*. Market participants enjoy high degrees of differentiation between each other.

Middle managers at many major corporations, as reported in the book, face a different situation. They are trapped in the Molochian nightmare of super-perfect competition *between different managers*. [The protections against this](#) process that most people have, are gone. Too many managers seek too few promotions, with too level a playing field. Everyone was assumed, past a certain level, to have the same skill at actually managing and getting things done.

This includes self-modification to seeing this competition as inherently good, and instinctively rewarding and allying with those who do the same, while punishing those who do otherwise. Dedication to the firm and to work and to personal success are the virtues. Valuing other things becomes vice. Morality, like family or religion or a hobby, is one more thing that can distract from your journey to success. If you’re distracted, you’ll lose, so you’d be a bad ally, so you fail.

These second order effects, combined with the optimization around getting ahead, allow mazes to spread and take over anyone and anywhere they are allowed a foothold.

[Big business does not hate your family](#), but its managers and bottom line see you as composed of things it could profit from, so effectively? It kind of does.

The self-modifications managers do, and the fact that many skills and connections, and much knowledge, and all their local privileges and status that they have become attached to, would expire worthless if they left, heavily discourages exit.

[What is their life like](#)? Their whole life is dedicated to “success” within the hierarchy, but awaits them there is another struggle for more “success.” Slack is non-existent, especially in terms of time. Anything that makes you unique, anything else you care about, is stamped out. They choose their activities, their friends, even their family, aiming at this, so all of them rely on the quest for this “success.” Mostly they are failures by their own metrics.

[The few who rise through this](#), and even become CEO, *still lose*. Their “success” does not bring happiness, or provide reproductive fitness, or improve the world. The fruits

are hollow. The game is rigged. Yet even for those who realize this, it is often seen as even worse to stop playing.

Maze Origins and Damage Mitigation

To avoid or oppose moral mazes, [we must identify them](#). The best known ways to do this are to look for too many levels of hierarchy, for people to *not* primarily describe their jobs as working for a particular person, and the absence of skin in the game, soul in the game, diversity of acknowledged skill levels, and slack. And of course, to *pay attention*, and see how things are done. Don't only check off boxes.

If someone does find themselves trapped in a maze, [it is imperative to escape](#). If at all possible, *quit and do something else*. That's easier said than done, but is easier done than those in the position to do it think it is. You're already doing something hard, and you have the skills to do a different hard thing that won't make you miserable, even if the pay starts out lower. You'll adjust. People around you will understand and sympathize more than you'd expect, especially if you tell it to them straight - you were unhappy, the job was toxic. If they don't sympathize, and demand that you devote yourself to the *illusion of security*, be sympathetic to that, especially if they love and/or depend on you, but do what you must. If you actually can't leave, try not caring about "success" or taking bold risks to achieve it.

[What has made these mazes so much more powerful than in the past?](#)

Some factors are real and inevitable. We *need* more large organizations for our civilization to function, than did prior civilizations, and in many ways they get to better leverage big data, machine learning and the internet, giving them an edge. As we grow safer and wealthier, our demand for the illusion of security rises, and mazes are relatively better positioned to provide that illusion.

There are many other reasons that are less real, and less inevitable. We protect organizations from disruption, especially in times of crisis. We see rent seeking of all kinds as increasingly legitimate. Mazes have gotten sufficiently powerful to cause a vicious cycle, as mazes reward and support other mazes and structure things to favor mazes. Our laws and regulations favor mazes over non-mazes, far beyond what is necessary due to civilizational complexity. Our educational system trains people for the maze, so much so that the people have largely forgotten what mazes are and what the alternative to them might be. We have been so atomized, and their ordinary human needs so delegitimized, that we do not see what we are giving up.

[What might we do to change things for the better?](#) Regulatory reform, health care reform, tort reform, ending corporate welfare or even forcibly breaking up large corporations would help. So would educating people on what mazes are and the dangers they pose, especially to their employees. We could work to change consumer behavior, to lower the status and aura of legitimacy of mazes and those who work for them. And we could work to lower demand for the illusion of security.

[For a given project](#), the best defense is to focus on the core elements, and thus do less things and be smaller, while minimizing interaction with other mazes. One should also seek separately to minimize levels of hierarchy, provide skin in the game and soul in the game, and be extremely careful with people. Hire, promote, evaluate and fire them with a keen eye. Anyone making you more like a maze needs to go, no matter how painful that is. You must fight for your institutional culture.

If someone with several hundred million dollars or more to spend wants to help, my best suggestion is to [create a full alternative stack](#), to allow people to sidestep maze incentives entirely and to actually get things done.

This only scratches the surface. I encourage you to follow the links to the original posts.

OpenAI announces GPT-3

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

This is a linkpost for <https://arxiv.org/abs/2005.14165>

Get It Done Now

Epistemic Status: Reference

A while ago, I read the book [Getting Things Done](#). Like most productivity books and systems, it includes detailed advice that approximately no one will follow. Unlike most productivity books and systems, it has two highly valuable key concepts. The second alone justified the time cost of reading the book. That principles are these:

Keep a record of tasks you've decided to do.

If you decide to eventually do a task that requires less than two minutes to do, that can efficiently be done right now, do it right now.

This wording is a refinement of the original concept of applying the two-minute rule *during 'processing time'* only. I think it's much better to use it any time doing the new task can be done efficiently – it's not waiting on anything, you have the necessary tools, it wouldn't interfere *too much* with your state, with a key short-term deadline, or the need to protect a large or important block of time, etc etc.

Having this simple concept in your head – it's better, once you notice something that you need to do, to just do it now rather than add it to your stack of things to do – has saved me far more trouble than one might expect.

Two minutes is a placeholder. Some people should use a lower or more often higher time threshold. The threshold should be adjusted based on the situation.

The book also contains a detailed method of how to create and maintain the list of tasks. It seemed annoying and overly complex and not suited to the way I think, and I never gave it a real try. The basic principle of 'have a system that ensures such tasks are not forgotten' still seems very strong.

The principle remains, and can be usefully extended further, which I plan to do in additional posts. But better to, by its own principles, write and get this posted now, so I can refer back to it.

A non-mystical explanation of "no-self" (three characteristics series)

This is the second post of the "a non-mystical explanation of insight meditation and the three characteristics of existence" series. You can read the first post, explaining my general intent and approach, [here](#).

On the three characteristics

So, just what are [the three characteristics of existence](#)?

My take is that they are a *rough way of clustering the kinds of insights that you may get from insight meditation*: in one way or another, most insights about the structure of your mind can be said to be about no-self, impermanence, unsatisfactoriness, or some combination of them. As upcoming posts should hopefully make obvious, this is not a very clear-cut distinction: the three are deeply intertwined with each other, and you can't fully explain one without explaining the others. I am starting with a discussion of no-self, then moving to unsatisfactoriness, then coming back to no-self, moving between the different characteristics in a way that seems most clear.

I think that what's called "enlightenment" refers to the gradual accumulation of these kinds of insights, combined with practices aimed at exploiting an understanding of them. There are many different insights and ways of exploring them, as well as many general [approaches for making use of them](#). Different traditions also seem to have different enlightenments [1, 2]. Thus, rather than providing any definitive explanation of "*this* is enlightenment", I attempt to focus on exploring how various cognitive mechanisms behind different enlightenments work. My intent is to cover enough of different things to give a taste of what's out there and what kinds of outcomes might be possible, while acknowledging that there's also a lot that I have no clue of yet.

So this is not trying to be anything like "a definitive and complete explanation of the three characteristics"; I don't think anyone could write such a thing, as nobody can have explored all the aspects of all the three. Rather, this is more of a sketch of those aspects of the three characteristics which I think I have some understanding of.

In particular, this explanation strongly emphasizes no-self and unsatisfactoriness, which I feel I have a better understanding of. Impermanence, which some approaches consider the very core characteristic, ends up relatively neglected. Apologies to any impermanence fans - maybe some day I'll come back to write more about it.

But let's get started with talking about no-self.

No-self

No-self is a confusing term, since it can easily be interpreted as the claim that, well, there is no self. But at least on one interpretation of Buddhism, the claim is much more subtle. Here's an excerpt from the article [No-self or Not-self?](#), by the Buddhist monk [Thanissaro Bhikkhu](#):

In fact, the one place where the Buddha was asked point-blank whether or not there was a self, he refused to answer. When later asked why, he said that to hold either that there is a self or that there is no self is to fall into extreme forms of wrong view that make the path of Buddhist practice impossible. Thus the question should be put aside. [...]

So, instead of answering "no" to the question of whether or not there is a self — interconnected or separate, eternal or not — the Buddha felt that the question was misguided to begin with.

Now, there are many interpretations of Buddha's teaching, and [what the Buddha even really said](#) in the first place; other people will offer a different kind of an account. But let's suppose that this particular interpretation *is* correct. What might it mean?

Well, clearly people *feel* that something like a "self exists". But [rather than arguing](#) about whether or not a self exists, one should investigate the mechanisms by which the experience of having a self is constructed. Once those cognitive algorithms are understood, one knows what creates a *feeling* of having a self - and then there is nothing more to explain.

Of course, in Buddha's day, they did not have cognitive science and a theory of neural networks, so he was unable to express his position in those terms. They did, however, have well-developed meditative techniques. And those techniques could be used to investigate how the experience of having a self was developed.

Now, one might reasonably ask, if the question of "does the self exist" is misleading, then why is this often phrased in the form of the claim that the self does *not* exist?

"The self does not exist in the way you think"

In my daily experience, it generally feels like there [exists a distinct "me"](#). There is *someone*, an "I" who sees what I see, hears what I hear, feels what I feel. It feels like I can generally make choices, consider information, act according to my best judgment. It feels that there's a meaningful sense in which the same me existed yesterday, and will continue to exist tomorrow. If you were to make a copy of me that was atom-to-atom identical, I [might intuitively feel](#) that there would exist a distinct difference between the original me and the copy. We might be exactly identical and act exactly the same, but there would still be a different *experiencer*.

But I also know about the scientific "multi-agent" models of mind, described briefly in [my last post](#) and more extensively in [earlier ones](#), where different subsystems within the brain are responsible for my actions. In those models, there is no privileged subsystem in charge of making decisions. Different subsystems take charge at different times, based on a preconscious selection process which is not under the control of any particular subsystem. There is also no particular subsystem which could be singled out as *the one* experiencing things. Rather, anything which makes it to consciousness is broadcast into many different subsystems, each of which can do different things with that information.

So experientially, I feel like I have a self which works in a particular way. Science suggests that my mind actually works in a different way: e.g. decisions are made by a distributed collection of semi-independent subsystems, rather than by a distinct "deciding self". So some might claim that the self *as I intuitively experience it* does not exist, as the intuitive conception does not match reality.

For example, [*The Manual of Insight*](#) is a treatise on meditation by the Theravada Buddhist monk Mahasi Sayadaw, who had a significant impact on spreading insight meditation in the West. The book quotes the Theravada scripture of [*Paṭisambhidāmagga*](#) as saying, in its elaboration of no-self, that:

There is no self that is able to control, to own, to feel, to give orders, to behave according to one's will, no self that is everlasting, or that is the agent of going, seeing, and so on. [...]

[As a result of meditation practice, one comes to see the mind as] empty of self (suññato). Here "self" (atta) means an entity that is the owner of the body, permanently residing in the body, the agent of going, seeing, and so on, the agent who feels pleasant and unpleasant feelings, able to give any orders, and able to exercise mastery. Such an entity, which is [a product of one's] speculation, belief, or obsession, may be called being, soul, ego, or self.

Likewise, Daniel Ingram, meditation teacher and author of the widely-read book *Mastering the Core Teachings of the Buddha*, [writes that](#) (emphasis mine):

The original Pali term, anatta, means literally "not-self". This same term is also rendered by other authors in other ways, some of which can be extremely problematic, such as egolessness, a terribly problematic term, since ego as understood in the Western psychological sense is not the referent of the conception of "self" targeted in Buddhism. Another problematic rendering of this term is "emptiness". Emptiness, for all its mysterious-sounding connotations, means that **reality is empty of, devoid of, or lacking a permanent, separate, independent, acausal, autonomous self**. It doesn't mean that reality is not there, but that reality is not there in the way it may appear to us to be. [...]

It's not that the constellation labeled "me", or "you", a grouping of physical and mental components, does not exist and function in some ordinary sense. **It's that none of those components exist independently or acausally, which is how ignorance conceives of them.** Ultimate unfindability of the components of reality in no way precludes their conventional existence!

Intellectually many people do not think that their self has an acausal existence, independent of the laws of physics. But the kind of understanding one can get from meditation is different. As I will discuss, the ways that specific subsystems react to various situations is linked to their model of the self. Normally, even if you intellectually understand that you do not have an acausally acting self, your mind cannot directly see the actual causality. Many of the subconscious models driving your behavior will only update if they are forced to directly witness evidence contradicting their old assumptions. (For a previous discussion of this in the context of psychotherapy and emotional beliefs, see [my review of *Unlocking the Emotional Brain*](#).)

Early insights into no-self

Recall again the model that the content of consciousness roughly corresponds to a "[global workspace](#)" which contains information submitted by different subsystems. In normal circumstances, there are some objects in the stream of experience (global workspace) which the overall system treats as being more "me" than the rest. For

example, many people experience themselves as inhabiting a space somewhere behind their eyes, looking at the world from that location.

Suppose that I now do some kind of practice where I examine this experience in more detail. Here is a simple one:

1. Look at an object in front of you. Spend a moment simply examining its features.
2. Become aware of the sensation of being someone who is looking at this object.
While letting your attention rest on the object, try to notice what this sensation of being someone who is looking at the object feels like. Does it have a location, shape, or feel?

You may wish to take a moment to do this right now, before reading about my results.

When I do this kind of exercise, a result that I may get is that there is the sight of the object, and then a pattern of tension behind my eyes. *Something* about the pattern of tension feels like "me" - when I feel that "I am looking at a plant in front of me", this could be broken down to "there is a tension in my consciousness, it feels like the tension is what's looking at the plant, and that tension feels like me".

Your result may be different from this. You may find yourself identifying with another sensation, or you might not be able to hone down on any particular sensation on the first try... but if you are like most people, you probably still have some kind of a feeling of looking out at the world.

My guess is that this sensation is a tag coming from some subsystem whose task is to keep track of one's spatial location relative to their surroundings. We know that there are multiple such systems in the brain, and that these systems getting out of sync - one system indicating a particular location and another indicating a differing location - [can create the feeling of an out-of-body experience](#). In computer terms, sensory data comes in, and then some subsystem parses that sensory data and indicates where one's "I" is located, passing this tag for other subsystems to use. Going by the previous example of me feeling a tension around my eyes that feels like me looking at the plant, we might think that something like the following is happening:

- Subsystem 1 sends the sight of a plant into the global workspace
- Subsystem 2 sends the feeling of tension around the eyes into the global workspace
- Subsystem 3 tags the tension as my current location, and binds all of these percepts together as an experience of "I am seeing the plant", which is also sent to the global workspace

An interesting thing is that the subsystems in the brain seem to take the tag as an ontological fact. Suppose that someone hands you a map of your surroundings, and has helpfully marked your current location with a red tag saying “YOU ARE HERE”.



But suppose that you now get a little confused. Rather than taking the spot with red ink as *indicating* your location in your physical world, you take the red spot on the map to *be* your physical location. That is, you think that you *are* the “YOU ARE HERE” tag, looking at the rest of the map *from* the red ink itself.

But of course, the fact that you are seeing the above picture, means that you cannot be looking *from* the red ink in the picture. The map includes the red ink, meaning that the person who is looking at it is actually *outside* the map.

Likewise, people tend to have a sensation of looking at the world from behind their eyes; but they are actually aware *of* the sensation, as opposed to being aware *from* it. It is a computational representation of a location, rather than being the location itself. Still, once this representation is fed into other subsystems in the brain, those subsystems will treat the tagged location as the one that they are “looking at the sense data from”, as if they had been fed a physical map of their surroundings with their current location marked.

But a particular tag in the sense data is not actually where they are looking at it from; for one, the visual cortex is located in the back of the head, rather than right behind the eyes. Furthermore, any visual information is in principle just a piece of data that has been fed into a program running in the brain. If we think of cognitive programs as analogous to computer programs, then a computer program that is fed a piece of data isn't really “looking at” the data “from” any spatial direction.

In vipassana-style meditation, you train your attention to dissect components of your experience into smaller pieces. (Vipassana is commonly translated as insight meditation, but here I treat it as a particular subcategory of insight meditation.) In third-person terms, this probably trains up pattern-detectors which can monitor the content of the global workspace in extreme detail. Eventually, there's sufficient clarity about the sense of location for low-level schemas to pick up on the inherent contradiction involved in looking *at* something which the system is supposedly looking *out from*.

The opposite strategy is commonly associated with what are so-called [nondual techniques](#). Instead of training an analytical, attention-controlled part of the mind to examine the sense of self, the nondual route is to nudge the mind into a state where

those analytical parts of the brain become less active. As those parts also produce the sense of ‘the observer’ in the first place, attenuating their activity can offer a glimpse into a state of consciousness where that sensation is lacking. Some versions of this approach seem to be tapping into some of the same machinery which causes people to experience a state of flow, as flow states also seem to involve a downregulation in both analytical thought and the sense of self.

Frequently, the sense of self being diminished in this way is a sufficiently interesting experience that the analytical subsystems kick back online to make sense of it - but over time, one can train oneself to experience more such glimpses, until there is a broader shift.

It is not clear to me to what extent these routes lead to exactly the same result. It seems to me that both eventually end up at a state where the sensations tagging one’s physical location still continue to be produced, and can be used as an aid for spatial reasoning, but the system no longer intrinsically identifies with them. Rather, the sensations are seen as being constructed by a machinery which is independent of the actual stream of sensory input.

But there seem to be some differences in how you reach that place. For the sake of analogy, let’s pretend that the machinery is a hologram projector, painting a realistic image of a person in the middle of a room. The vipassana path would correspond to looking very closely at all the details of the hologram, until you noticed discrepancies in how it was created. That would give you a detailed insight into how exactly the projector used light to draw the image, but would be rather slow. In contrast, the nondual route involves just turning the projector off for a moment - making it very obvious that the hologram was in fact a hologram, but telling you much less of how it was built.

Another difference is the no-self versus all-self interpretation. Some schools say that this kind of practice leads you to realizing that there is no self; other schools, generally more associated with Hinduism than Buddhism, say that they lead you to realizing that all is self. (Western philosophy has the corresponding concepts of [closed, open and empty individualism](#).)

Some of the end results from both paths are described in a similar way, however. For example, a common metaphor about the result of some varieties of practice is that of “being the sky rather than the clouds”. Below is one formulation of it. The outcome seems to be that rather than identifying with the sensations of the supposed observer, one’s identity shifts to *the entire field of consciousness itself* (in line with the thing about a program reading a file not having any location that would be defined in terms of the file):

One way of describing the experience of glimpsing in effortless mindfulness practice is to use the metaphor of a cloud. You may have felt as if you have been living in a cloud; maybe it feels like a storm cloud a lot of the time. See if you can feel the boundary and foginess of this cloud that you call “me.” You may have been trying to feel better by cleaning up the cloud of your mind by replacing negative thoughts with positive thoughts and developing good attitudes. You may have tried to calm your body and mind to make your brain as clear as possible. Within your cloud are storms, old traumas, emotional challenges, and relationships of all types. Each time you change these things and clean up one area of the cloud, it seems that another foggy issue or thunderous problem arises.

Effortless mindfulness does not begin with dissolving the cloud, calming it, or trying to transform its contents. The glimpsing method of effortless mindfulness begins with awake awareness stepping out of the cloud, shifting, dropping, or opening to discover that you are also the open sky of awake awareness! When you shift out of this cloud of the emotional or small mind and discover this spaciousness of still, quiet, alert awareness, it's a great relief. You can realize that you are the sky, and the cloudy emotions and thoughts are everchanging weather.

[...] As we reach the fullness of effortless mindfulness, we will discover open-hearted awareness and ways to naturally embrace and welcome all emotions and parts of ourselves. [...] After all, all weather comes and goes, and no storm ever hurt the sky.

(Loch Kelly, [The Way of Effortless Mindfulness](#))

Or this more concrete description, quoted in a paper on meditation-induced changes to the sense of self ([Lindahl & Britton 2019](#)):

So, [the retreat] was in the spring and I was doing some raking leaves, and just as I was raking, this really profound feeling of 'this is all me' came to me. And so the 'this is all me' — what that means is that my identity is literally everything that I could see through my eyes. So, the rake that I was holding in my hands was me. The ground that I was raking was me. The feet that I could see down at the bottom of my body, that was me. The steps up to the residence, that was me. The sky was me. The trees were me. And so, everything was just 'me'. And that there wasn't really anything else. It was all just 'me'. [...] Those experiences that I related about what I would call kenshō experiences, there was no viewer in those — it was just what was there, and there was no viewer observing it.

Here is how I would rephrase these reports in third-person terms. Normally, there is a flow of information within the global workspace: mental objects representing sensory information, thoughts, and some objects encoding a sense of there being someone who watches the senses. These kinds of experiences are a part of a process where the system reorients its assumptions to recognize that there is no homunculus sitting behind the eyes and watching everything.

From an external point of view, we can say that *your conscious mind - or "you" - consists of everything that is in the global workspace, and no particular piece of mental content is more or less "you" than the others are*. If you see a rake in your hands, then there is a process within your brain generating that visual experience. The experience exists as a part of your mind. Likewise, the experience of there being a *someone* who is *having* that experience, is generated by a process within your brain, and exists as a part of your mind. *Everything that you ever experience is mental content generated by your brain*, as opposed to you having [direct access to reality](#).

Now, in Loch Kelly's quote above, there is the suggestion that changing one's identification to the entire field of consciousness will also change how one relates to negative experiences. Exactly why this would happen is an important question, and I will come back to it later. For now, let's look a bit more at why getting such an experience can be so difficult.

The self as a tool for planning

A thing that might happen, once the above has been explained to you, is that you put a lot of effort into *intellectually* figuring out the contradiction between experiencing something that you also identify with, and then figuring out what must be going on instead. This kind of theorizing can be useful for purposes of writing articles such as this one. But you cannot use theorizing alone to put your brain into a no-self state by convincing your brain of the contradiction. Meditation teachers may explicitly warn you that it is impossible for your thinking mind to comprehend no-self states in such a way that would cause you to actually *experience* them.

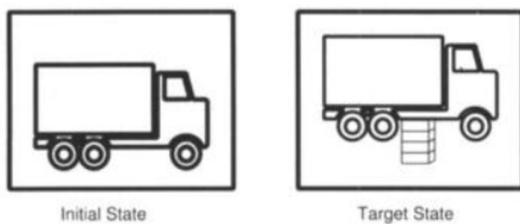
I suspect that a part of this is because the subsystems in the brain used for this kind of theorizing take the sense of self as input. As a result, them being active tends to put the mind in a state where it identifies more strongly with the sensations of a self.

Going back to the map analogy, consider the route-finding algorithm included in Google Maps: you give it a starting location, an end location, some parameters of what kind of a route you prefer, and it then finds you the best route that meets those criteria.

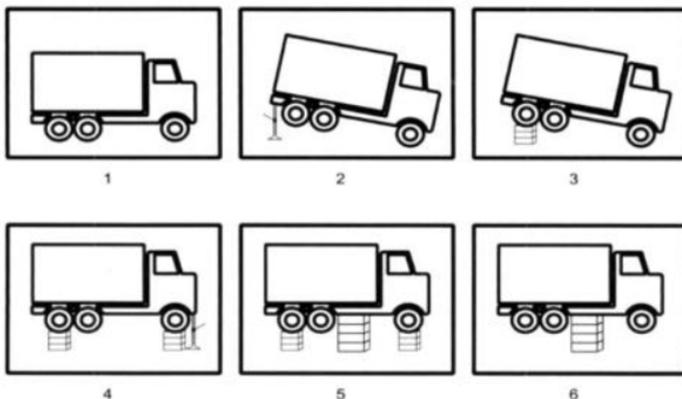
I suggested that the sense of the observer is like a point in a map, saying “YOU ARE HERE”, and that one of the goals of practice was coming to see that the point that’s marked on the map cannot actually be “your” real location. That is, some part of your mind stops treating the red ink on the map as being identical to where you are. But a route-finding algorithm does not have the option of treating the starting point of a route as anything else than as the starting point of a route. Its *entire purpose* is to assume that the “YOU ARE HERE” really does correspond to your real location, and to then plot a route from there. If it didn’t, it wouldn’t be a route-finding algorithm anymore.

What I am calling “intellectual” parts of your brain, seem to be similar to route-finding algorithms. Their purpose is to figure out a path from where you are now, to some desired target state.

The literature on expertise suggests that people figure out novel tasks by running mental simulations of how to get from a current state to a target state, and then trying to carry out a sequence that they have successfully simulated ([Klein 1999](#)). For example, you might be faced with a truck sitting on the ground. Using a jack and concrete blocks, you want to get it up on the air on a column of blocks.



You mentally go through different options, until you figure out a sequence of steps that gets you to the end result. When you find something that seems promising enough, you give it a shot.



Now, in the example of a truck, your reasoning can happen purely in terms of what is going to happen to the truck; the same process would work exactly the same regardless of whether you or someone else was doing it. But what happens if you set the goal of “I want to get to a state where I experience no sense of self?”

This again fires up the parts of your brain that carry out mental simulations... but just as in the truck example, where they needed to track what was happening to the truck in each step of the sequence, they now need to track whether or not *you* are experiencing a sense of self in any given step. This makes it impossible for them to find a state where you wouldn't experience a sense of self, as the very act of trying to plan how to get you there requires instantiating a sense of self that represents you in the simulation!

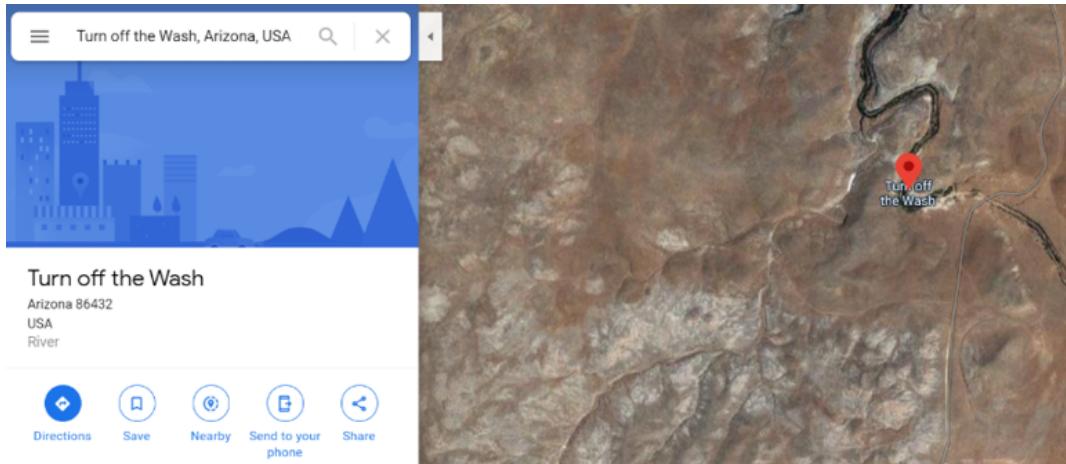
This can make for some frustrating experiences, in that if you once experience a state with a drastically weakened sense of self, it may feel pleasant and you then want to get back to it. But trying to figure out how to get back into it, is exactly the kind of a process that may *prevent* you from getting back. This is part of the reason why some traditions and teachers say things like “in order to get enlightened, you must stop striving for enlightenment”, as well as claiming that thinking in terms of outcomes is contrary to the spirit of the practice.

What the planning system would actually need to do to achieve its goal, is to simply turn itself off, so that it stops projecting a sense of self into the global workspace. But it cannot accurately represent this target state, as it parses it as “a state where *I* experience no sense of self”. Its representation of the target still includes a sense of someone who either is or is not experiencing a sense of self.

To use the map analogy, this is something like asking Google Maps to route a path to a state where the Google Maps program has been turned off. There is simply no way for it to do that, because the notion of being on or off is not explicitly represented anywhere in the program. The pathfinding routine of Google Maps only reasons in terms of where “you” are in the maps that are loaded into it.

What the route-finding algorithm in Google Maps *can* do, is something like take a map, find the location on it that sounds the closest to “turn yourself off”, and plot a route to that. Of course, this will not actually turn it off, but it is something that the algorithm can at least *do*. So upon being given the task of turning itself off, it will plot a route to that location, correctly notice that this is not actually fulfilling the task that it was given, and trash around trying to find a better target location. This corresponds to a meditator thinking something like “oh, how do I get into a no-self state again, oh wait,

if I try to get into a no-self state I can't do it, so I have to stop trying to get to it... so now I am going to stop trying to get to it... wait, that is trying again, gahhhhhh."



Google Maps trying to figure out where "turn off" is. This location isn't quite it, but maybe it would at least be getting close?

This runs into the contradiction between the way that we often think about our minds, and the way that our minds actually work. We often have the feeling that at least *some* of the content in our consciousness is something that we can actively choose. Most people don't expect to be able to choose their emotions, but at least the act of *intentionally trying to do something* feels like it should be under conscious control - isn't that what intentionally acting *means*?

But under a multi-agent framework, "trying to do something" simply means that a subsystem is active and pursuing a particular goal. Neither the subsystem itself, nor any other subsystem, has direct access to a command which would turn that subsystem off: the choice of which subsystem to activate or keep running, happens by means of a [preconscious selection process](#). That means that, despite it possibly going against one's naive intuition, it is perfectly possible to consciously intend to do something while also having no conscious control over the fact that you are intending to do so.

As I noted before, there are several approaches to dealing with this problem. For example, [flow states](#) typically involve activities that are similar to the truck task, in that they do not require a sense of self. At the same time, the task is challenging enough that it requires one's full attention: in other words, a single planning subsystem uses up the full bandwidth of consciousness, being the only one that projects content to the global workspace. If there was any spare capacity, other planning systems could project self-related thoughts at the same time (e.g. thinking about what to do after the current task is done), thus instantiating a sense of self. Thus, getting the mind into something like a flow state is one way to reduce the sense of self.

On the other hand, some situations just trigger the self-related planning machinery very strongly. In vipassana/mindfulness-style approaches, one frequently ends up creating a sense of being an observer who is detached from their thoughts and emotions. For example, a simple set of "labeling" instructions is just:

1. Notice something in your consciousness.
2. Give it a label, such as "[seeing](#)", "[feeling](#)", or "[hearing](#)".
3. Go back to 1.

In these instructions, the planning machinery is given a goal that it *is* capable of carrying out. Following these instructions does instantiate a sense of self - the planning system needs to monitor the question of "am I still labeling my experience". However, this task constructs an experience where the "I" is merely *observing* other mental content, and that mental content is happening on its own.

This can be particularly useful in situations which are experienced as important or potentially threatening, as those kinds of situations tend to make goal-oriented systems kick in very strongly to help resolve the situation. For people with trauma and ongoing anxiety, this might include even situations with no immediate external concerns; such people may almost constantly be in a state of uncertainty, activating planning systems with the goal of making those unpleasant feelings go away. If one practices dispassionately observing the contents of their mind, even when the content is unpleasant, one can in effect train up a new subsystem that competes with the other subsystems in projecting content to the global workspace. (However, it needs to be noted that training the mind to closely examine unpleasant feelings may also [make trauma responses worse](#) by bringing more attention to them and interfering with the subsystems that were previously regulating the responses.)

In this, one continues the process of identifying with a self, but the thing that is being identified with shifts to a sense of someone who is just observing everything happening in the mind - which can bring relief from various unpleasant emotions. Once one gets to this kind of a state, the subsystem trained to do this can continue to further investigate the contents of the mind in fine detail... either looking at other characteristics like impermanence or unsatisfactoriness, or turning its focus *on itself*, to deepen the no-self realization by seeing that the observer self that it is projecting is *also* something that can be dis-identified with.

The meditation teacher [Michael Taft](#) describes this kind of a turn in his article on [Escaping the Observer Trap](#):

Many traditions—especially mindfulness meditation—encourage you to observe your sensory experience in a neutral manner. Observe your breathing, observe emotions, observe thoughts, and so on, without reacting to them. This observer technique works really well because it gives you something like an outside perspective on your own experience. You can watch your own mind, your reactions, your emotions, your behavior almost from the perspective of another person, and that is tremendously useful feedback to have. It leads to equanimity, and the tremendous personal growth that mindfulness advocates are always talking about. [...]

Taking this observer stance is so useful, in fact, that many teachers stop there and do not talk about the next important step in spiritual development. But there is a hidden problem with the observer technique, which becomes obvious once you think about it. Who is the observer? Who is this person who is behind the binoculars, watching your experience from the outside? This neutral observer you've created over time is actually just another—albeit smaller and less neurotic—version of the ego. It's the sense of being a person who is doing the meditating. You could also call it a meditator ego or an observer ego. Creating this neutral

observer is very useful, but the goal of meditation is not to create a new meditator ego, it's to see through the illusion of the ego entirely.

It is quite common for even very dedicated mindfulness students in observation-based traditions to get stuck in observer mode forever. I have seen it over and over in my experience. Being the observer, a neutral meditator ego, is not such a bad place to be; certainly it is much preferable to the unconscious, robotic mode of life lived without any self-reflection. However, it impedes all deeper progress toward real awakening. So the only way forward is to let go of the observer ego; to release the sense of being a person who is doing a meditation. [...]

To release yourself from the observer trap, begin by realizing that the observer, however comfortable or habitual, is still just another version of the ego. You've spent endless hours watching your breath and your emotions and your thoughts. Now it's time to watch the watcher instead. You have to observe the observer. You do this, in typical mindfulness style, by carefully deconstructing the components of the observer itself.

The observer ego is constructed out of the same components as the everyday ego, but on a smaller scale. The everyday mind has thoughts about all sorts of stuff, the observer has thoughts about how the mediation is going, or how long until this sit is over. The everyday ego has emotions about all sorts of stuff, but observer has emotions about how this sit is going, or even blissful feelings of love and joy. The everyday ego has all sorts of body sensations, but the observer has a very special set of body sensations: the sensations of where he/she imagines awareness is located. [...] So to overcome the observer problem and get unstuck in your practice, closely observe the sensations (i.e. the thoughts and feelings) associated with the observer ego.

This is the second post of the "a non-mystical explanation of insight meditation and the three characteristics of existence" series. The next post in the series is "[Craving, suffering, and predictive processing](#)".

Why We Age, Part 2: Non-adaptive theories

This is a linkpost for <https://apomorphic.com/2020/01/12/why-we-age-2-nonadaptive>

Follows from: [Why We Age, Part 1](#); [Evolution is Sampling Error](#); [An addendum on effective population size](#)

Partially redundant with [Highlights of Comparative and Evolutionary Aging](#).

[Last time](#), I introduced three puzzles in the evolution of ageing:

This, then, is the threefold puzzle of ageing. Why should a process that appears to be so deleterious to the individuals experiencing it have evolved to be so widespread in nature? Given this ubiquity, which implies there is some compelling evolutionary reason for ageing to exist, why do different animals vary so much in their lifespans? And how, when ageing has either evolved or been retained in so many different lineages, have some animals evolved to escape it?

I divided existing theories of the evolution of ageing into two groups, adaptive and nonadaptive, and discussed why one commonly believed nonadaptive theory – namely, simple wear and tear – could not adequately answer these questions.

In this post I'll discuss other, more sophisticated non-adaptive theories. These theories are characterised by their assertion that ageing provides no fitness benefit to organisms, but rather evolves *despite* being deleterious to reproductive success. Despite the apparent paradoxicality of this notion, these theories are probably the most widely-believed family of explanations for the evolution of ageing among academics in the field; they're also the group of theories I personally put the most credence in at present.

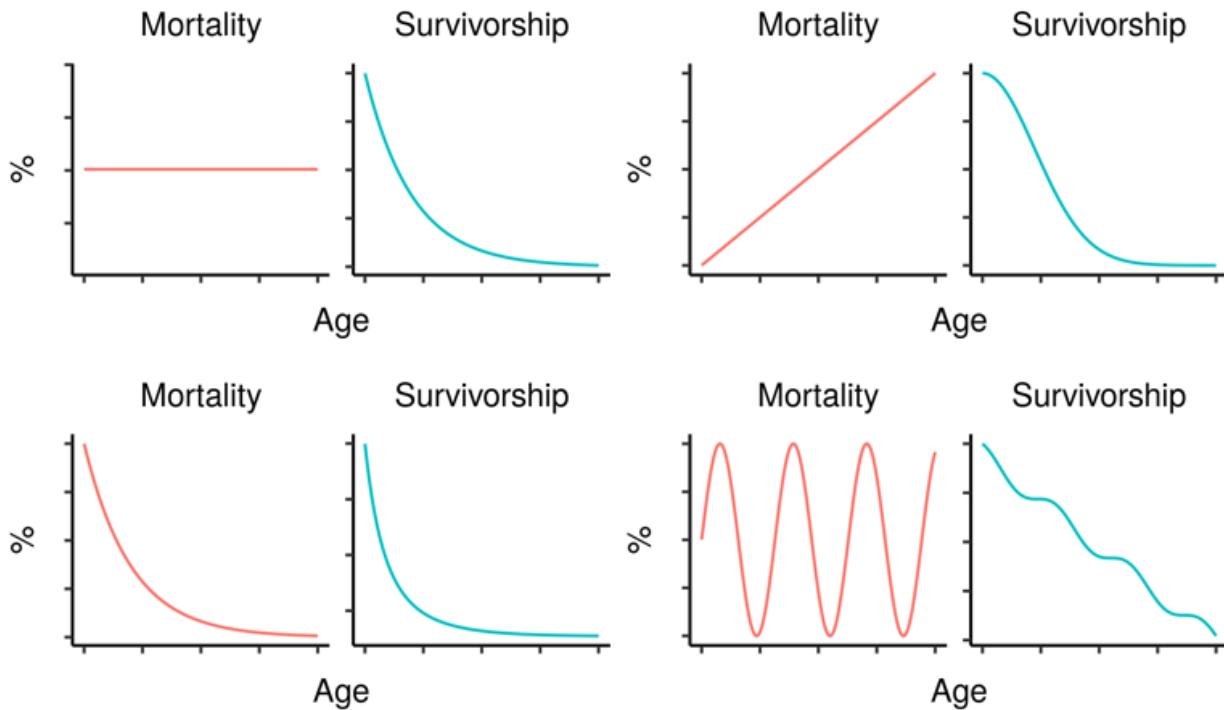
How can this be? How can something non-adaptive – even deleterious – have evolved and persisted in so many species across the animal kingdom? To answer this question, we need to understand a few important concepts from evolutionary biology, including [genetic drift](#), relaxed purifying selection, and pleiotropy. First, though, we need to clarify some important terminology.

Mortality, survivorship, and fecundity

For the purposes of this post, a **cohort** is a group of individuals from the same population who were all born at the same time, i.e. they are of the same age. The **survivorship** of a cohort at a given age is the percentage of individuals surviving to that age, or equivalently the probability of any given individual surviving at least that long. Conversely, the **mortality** of a cohort at a given age is the probability of an individual from that cohort dying at that age, and not before or after.

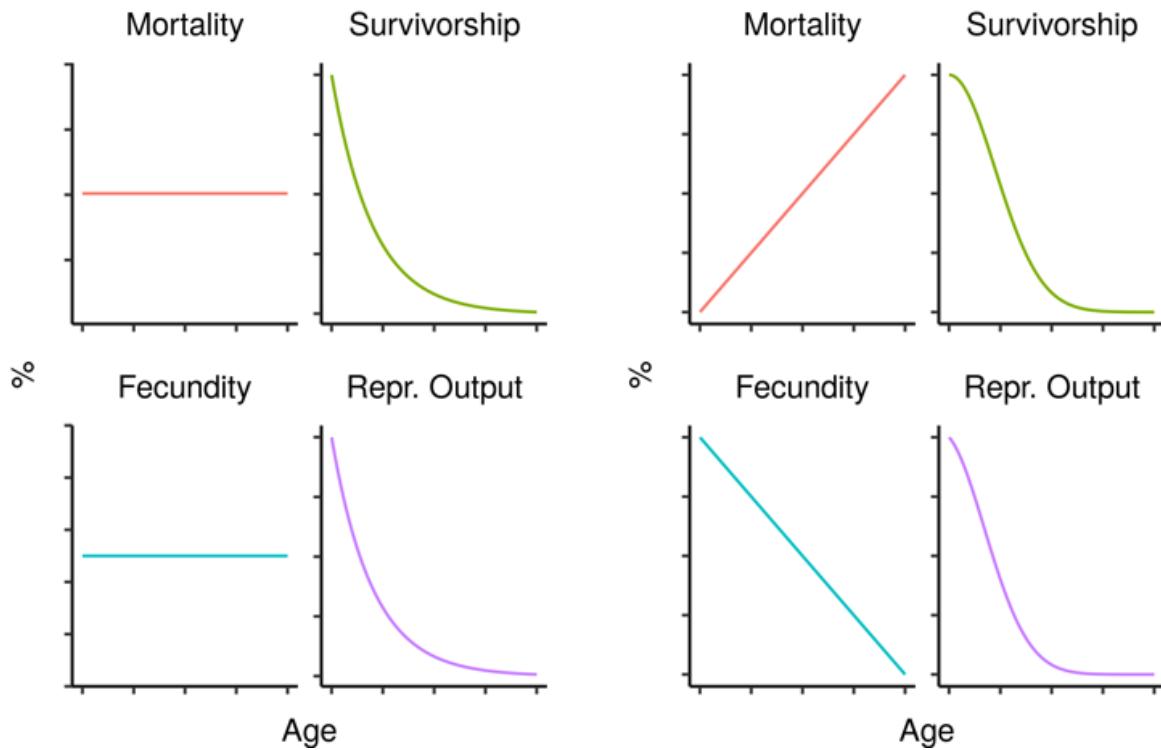
Survivorship and mortality are therefore related, but distinct: survivorship is the result of *accumulating* mortality at all ages from birth to the age of interest^[1]. As a result, the mortality and survivorship curves of a cohort will almost always look very different; in particular, while mortality can increase, decrease or stay the same as age increases,

survivorship must always decrease. As one important example, *constant mortality* will give rise to an *exponential* decline in survivorship^[2].



Four hypothetical mortality curves and their corresponding survivorship curves.

In evolutionary terms, survival is only important insofar as it leads to reproduction. The age-specific **fecundity** of a cohort is the average number of offspring produced by an individual of that cohort at that age. Crucially, though, you need to survive to reproduce, so the actual number of offspring you are expected to produce at a given age needs to be downweighted in proportion to your probability of dying beforehand. This survival-weighted fecundity (let's call it your age-specific **reproductive output**) can be found by multiplying the age-specific fecundity by the corresponding age-specific survivorship. Since this depends on survivorship, not mortality, it will tend to decline with age: a population with constant mortality and constant fecundity (i.e. no demographic ageing) will show reproductive output that declines exponentially along with survivorship.

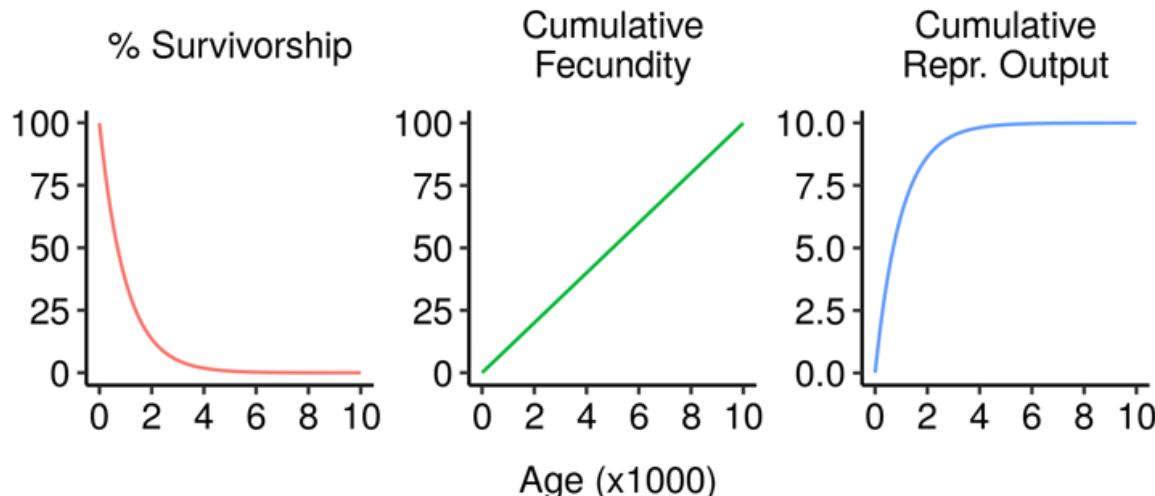


Two hypothetical mortality/fecundity curves and their corresponding reproductive outputs.

The fitness of an individual is determined by their lifetime reproductive output (i.e. the total number of offspring they produce over their entire lifespan)^[3]. Mutations that significantly decrease lifetime reproductive output will therefore be strongly opposed by natural selection. It seems mutations leading to ageing (i.e. an increase in mortality and decrease in fecundity with time) should be in that category. So why does ageing evolve?

What good is immortality?

Imagine a race of beautiful, immortal, ageless beings -- let's call them elves. Unlike we frail humans, elves don't age: they exhibit constant mortality and constant fecundity. As a result, their age-specific survivorship and reproductive output both fall off exponentially with increasing age -- far more slowly, in other words, than occurs in humans.



Survivorship, cumulative fecundity and cumulative reproductive output curves for a population of elves with 1% fecundity and 0.1% mortality per year.

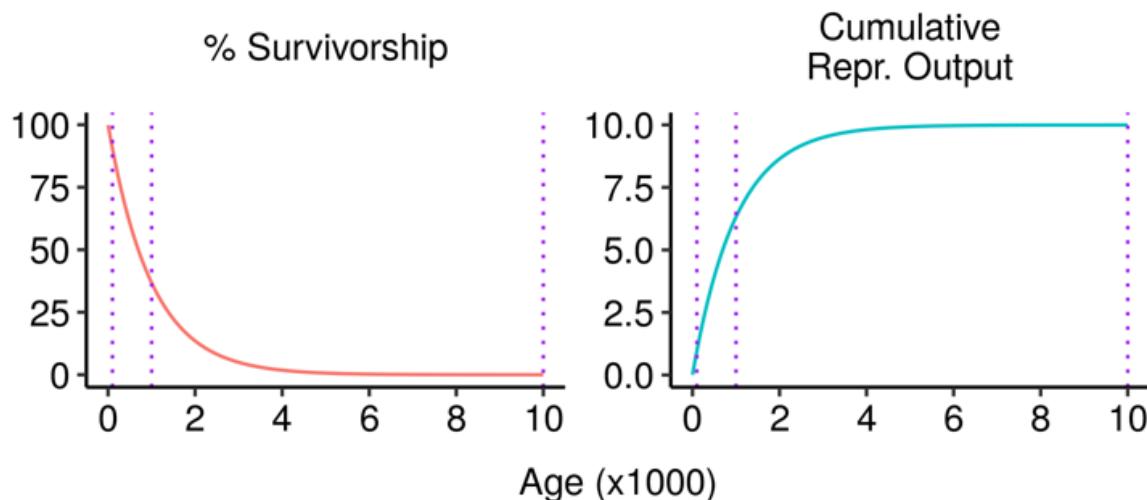
Under the parameters I've used here (1% fecundity, 0.1% mortality), an elf has about a 50% chance of making it to 700 years old and a 10% chance of living to the spry old age of 2,300. An elf that makes it that far will have an average of 23 children over its life; 7 if it only makes it to the median lifespan of 700.

Since fecundity and mortality are constant, an elf that makes it to 3,000 will be just as fit and healthy then as they were as a mere stripling of 500, and will most likely still have a long and bright future ahead of them. Nevertheless, the chance of any given *newborn* elf making it that far is small (about 5%). This means that, even though an old elf could in principle have as many children as a much younger individual elf, the actual offspring in the population are mainly produced by younger individuals. Just over 50% of the lifetime expected reproductive output of a newborn elf is concentrated into its first 700 years; even though it could in principle live for millennia, producing children at the same rate all the while, its odds of reproducing are best early in life. You can, after all, only breed when you're living.

This fact -- that reproductive output is concentrated in early life even in the absence of ageing -- has one very important consequence: natural selection cares much more about you when you're young.

Natural selection is ageist

No genome is totally stable -- mutations always occur. Let's imagine that three mutations arise in our elven population. Each is fatal to its bearer, but with a time delay, analogous to [Huntington's disease](#) or some other congenital diseases in humans. Each mutation has a different delay, taking effect respectively at 100, 1,000, and 10,000 years of age. What effect will these mutations have on their bearers' fitness, and how well will they spread in the population?



Three potential fatal mutations in the elven populations, and their effects on lifetime reproductive output.

Although all three mutations have similar impacts on an individual who lives long enough to experience them, from a fitness perspective they are very different. The first mutation is disastrous: almost 90% of wild-type individuals (those without the mutation) live past age 100, and a guaranteed death at that age would eliminate almost 90% of your expected lifetime reproductive output. The second mutation is still pretty bad, but less so: a bit over a third of wild-type individuals live to age 1000, and dying at that age would eliminate a similar proportion of your expected lifetime reproductive output. The third mutation, by contrast, has almost no expected effect: less than 0.005% of individuals make it to that age, and the effect on expected lifetime reproductive output is close to zero. In terms of fitness, the first mutation would be strenuously opposed by natural selection; the second would be at a significant disadvantage; and the third would be virtually neutral.

This extreme example illustrates a general principle:

The impact of a mutation on the fitness of an organism depends on both the magnitude of its effect and the proportion of total reproductive output affected.

— Williams 1957 [4]

Mutations that take effect later in life affect a smaller proportion of total expected reproductive output and so have a smaller selective impact, even if the size of the effect when they do take effect is just as strong. The same principle applies to mutations with less dramatic effects: those that affect early-life survival and reproduction have a big effect on fitness and will be strongly selected for or against, while those that take effect later will have progressively less effect on fitness and will thus be exposed to correspondingly weaker selection pressure. Put in technical language, the selection coefficient of a mutation depends upon the age at which it takes effect, with mutations affecting later life having coefficients closer to zero.

[Evolution is sampling error](#), and selection is sampling bias. When the selection coefficient is close to zero, this bias is weak, and the mutation's behaviour isn't much different from that of a neutral mutation. As such, mutations principally affecting later-life fitness will act more like neutral mutations, and increase and decrease in frequency

in the population with little regard for their effects on those individuals that do live long enough to experience them. As a result, while mutations affecting early life will be purged from the population by selection, those affecting late life will be allowed to accumulate through genetic drift. Since the great majority of mutations are negative, this will result in deteriorating functionality at older ages.

So our elves are sadly doomed to lose their immortality, unless something very weird is happening to cause them to keep it. Mutations impairing survival and reproduction early in life will be strenuously removed by natural selection, but those causing impairments later in life will accumulate, leading to a progressive increase in mortality and decline in fecundity. This might seem bad enough, but unfortunately there is more bad news on the horizon -- because this isn't the only way that nonadaptive ageing can evolve.

Perverse trade-offs

Imagine now that instead of a purely negative, Huntingdon-like mutation arising in our ageless elf population, a mutation arose that provided some fitness benefit early in life at the cost of some impairment later; perhaps promoting more investment in rapid growth and less in self-repair, or disposing the bearer more towards risky fights for mates. How would *this* new mutation behave in the population?

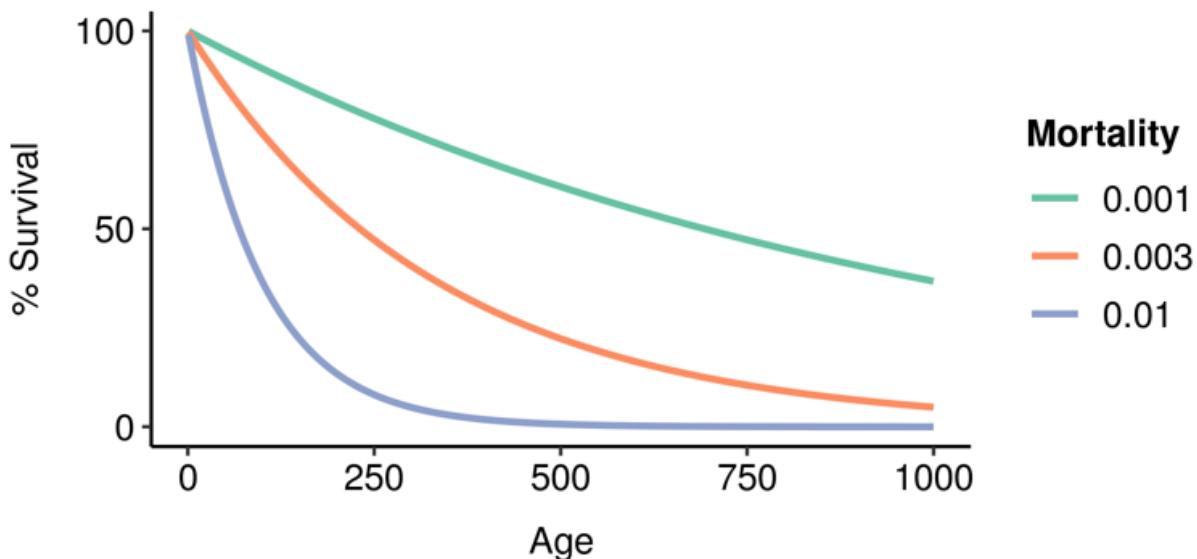
The answer depends on the magnitude of the early-life benefit granted by the mutation, as well as of its later-life cost. However, we already saw that in weighing this trade-off natural selection cares far more about fitness in early life than in later life; as such, even a mutation whose late-life cost far exceeded its early-life benefit in magnitude could be good for overall lifetime fitness, and hence have an increased chance of spreading and becoming fixed in the population. Over time, the accumulation of mutations like this could lead to ever-more-severe ageing in the population, even as the overall fitness of individuals in the population continues to increase.

This second scenario, in which the same mutation provides a benefit at one point in life and a cost at another, is known as *antagonistic pleiotropy*^[5]. It differs from the *mutation accumulation* theory of ageing outlined above in that, while in the former case ageing arises primarily through genetic drift acting on late-life-affecting deleterious mutations, the latter proposes that ageing arises as a non-adaptive *side effect* of a fitness-increasing process. Both theories are "non-adaptive" in that the ageing that results is not in itself good for fitness, and both depend on the same basic insight: due to inevitably declining survivorship with age, the fitness effect of a change in survival or reproduction tends to decline as the age at which it takes effect increases.

Mutation accumulation and antagonistic pleiotropy have historically represented the two big camps of ageing theorists, and the theories have traditionally been regarded as being in opposition to each other. I've never really understood why, though: the basic insight required to understand both theories is the same, and conditions that gave rise to ageing via mutation accumulation could easily also give rise to additional ageing via antagonistic pleiotropy^[6]. Importantly, both theories give the same kinds of answers to the other two key questions of ageing I discussed last time: why do lifespans differ between species, and why do some animals escape ageing altogether?

It's the mortality, stupid

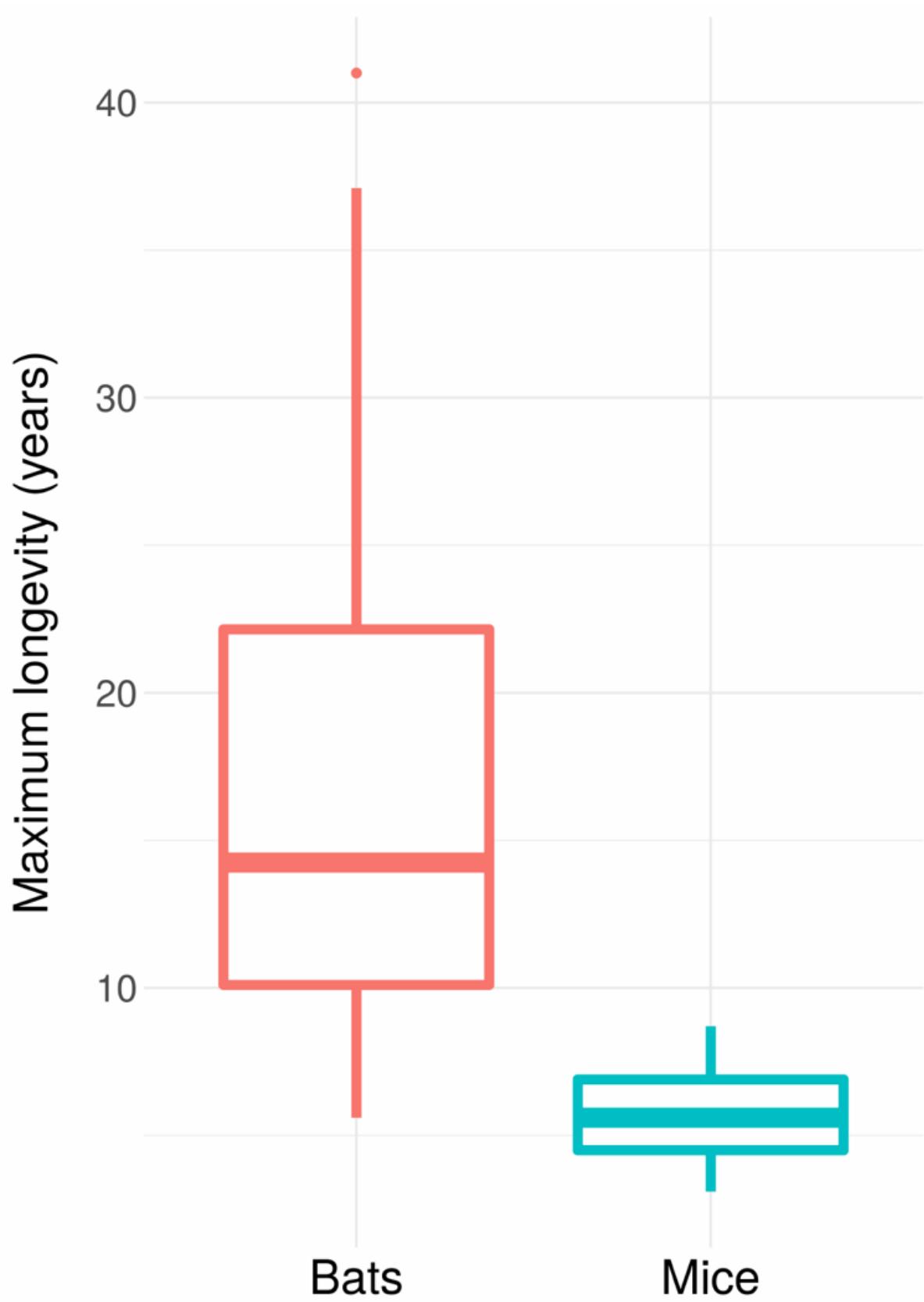
As explanations of ageing, both mutation accumulation and antagonistic pleiotropy depend on *extrinsic mortality*; that is, probability of death arising from environmental factors like predation or starvation. As long as extrinsic mortality is nonzero, survivorship will decline monotonically with age, resulting (all else equal) in weaker and weaker selection against deleterious mutations affecting later ages. The higher the extrinsic mortality, the faster the decline in survivorship with age, and the more rapid the corresponding decline in selection strength.



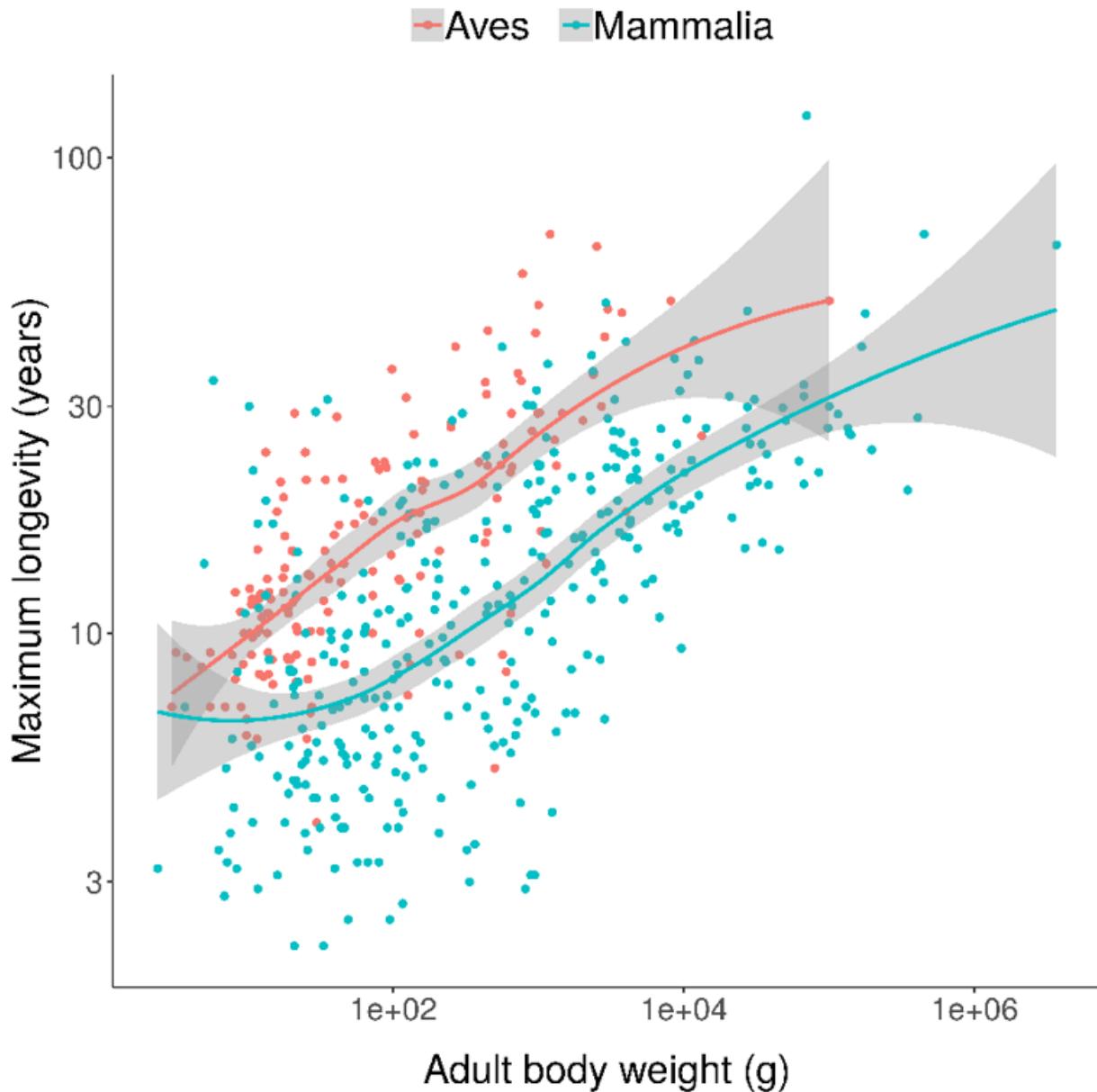
Age-specific survivorship as a function of different levels of constant extrinsic mortality.

As a result, lower extrinsic mortality will generally result in slower ageing: your chance of surviving to a given age is higher, so greater functionality at that age is more valuable, resulting in a stronger selection pressure to maintain that functionality.

This is the basic explanation for why bats live so much longer than mice despite being so similar: they can fly, which protects them from predators, which reduces their extrinsic mortality.



You can see something similar if you compare all birds and all mammals, controlling for body size (being larger also makes it harder to eat you):



Scatterplots of bird and mammal maximum lifespans vs adult body weight from the AnAge database, with central tendencies fit in R using local polynomial regression (LOESS).

In addition to body size and flight, you are also likely to have a longer lifespan if you are^[7]:

- Arboreal
- Burrowing
- Poisonous
- Armoured
- Spiky

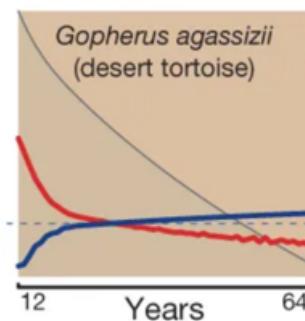
- Social

All of these factors share the property of making it harder to predate you, reducing extrinsic mortality. In many species, females live longer than males even in captivity: males are more likely to (a) be brightly coloured or otherwise ostentatious, increasing predation, and (b) engage in fights and other risky behaviour that increases the risk of injury. I'd predict that other factors that reduce extrinsic mortality in the wild (e.g. better immune systems, better wound healing) would similarly correlate with longer lifespans in safe captivity.

This, then, is the primary explanation non-adaptive ageing theories give for differences in rates of ageing between species: differences in extrinsic mortality. Mortality can't explain everything, though: in particular, since mortality is always positive, resulting in strictly decreasing survivorship with increasing age, it can't explain species that don't age at all, or even age in reverse (with lower intrinsic mortality at higher ages).

It's difficult to come up with a general theory for non-ageing species, many of which have quite idiosyncratic biology; one might say that all ageing species are alike, but every non-ageing species is non-ageing in its own way. But one way to get some of the way there is to notice that mortality/survivorship isn't the only thing affecting age-specific reproductive output: age-specific fecundity also plays a crucial role. If fecundity increases in later ages, this can counterbalance, or even occasionally outweigh, the decline in survivorship and maintain the selective value of later life.

Mammals and birds tend to grow, reach maturity, and stop growing. Conversely, many reptile and fish species keep growing throughout their lives. As you get bigger, you can not only defend yourself better (reducing your extrinsic mortality), but also lay more eggs. As a result, fecundity in these species increases over time, resulting – sometimes – in delayed or even nonexistent ageing:



Mortality (red) and fertility (blue) curves from the desert tortoise, showing declining mortality with time. Adapted from Fig. 1 of [Jones et al. 2014](#).

So that's one way a species could achieve minimal/negative senescence under non-adaptive theories of ageing: ramp up your fecundity to counteract the drop in survivorship. Another way would be to be under some independent selection pressure to develop systems (like *really good* tissue regeneration) that incidentally also counteract the ageing process. Overall, though, it seems to be hard to luck yourself into a situation that avoids the inexorable decline in selective value imposed by falling survivorship, and non-ageing animal species are correspondingly rare.

Next time in this series, we'll talk about the other major group of theories of ageing: adaptive ageing theories. This post will probably be quite a long time coming since I don't know anything about adaptive theories right now and will have to actually do

some research. So expect a few other posts on different topics before I get around to talking about the more heterodox side of the theoretical biology of ageing.

1. In discrete time, the survivorship function of a cohort will be the product of instantaneous survival over all preceding time stages; in continuous time, it is the [product integral](#) of instantaneous survival up to the age of interest. Instantaneous survival is the probability of surviving at a given age, and thus is equal to 1 minus the mortality at that age. [←](#)
2. Exponential in continuous time; geometric in discrete time. [←](#)
3. Lifetime reproductive output is equal to $\int_0^{\infty} r_a da$ (in continuous time) or $\sum_{a=0}^{\infty} r_a$ (in discrete time), where r_a is the age-specific reproductive output at age a . [←](#)
4. Williams (1957) *Evolution* 11(4): 398-411. [←](#)
5. "[Pleiotropy](#)" is the phenomenon whereby a gene or mutation exerts effects of multiple different aspects of biology simultaneously: different genetic pathways, developmental stages, organ systems, *et cetera*. *Antagonistic* pleiotropy is pleiotropy that imposes competing fitness effects, increasing fitness in one way while decreasing it in another. [←](#)
6. Which of the two is likely to predominate depends on factors like the [relative strength of selection and drift](#) (which is heavily dependent on effective population size) and the commonness of mutations that cause effects of the kind proposed by antagonistic pleiotropy. [←](#)
7. My source for this is personal communication with Linda Partridge, one of the directors at my institute and one of the most eminent ageing scientists in the world. I'm happy to see any of these points contested if people think they have better evidence than an argument from authority. [←](#)

Subspace optima

The term "global optimum" and "local optimum" have come from mathematical terminology and entered daily language. They are useful ways of thinking in every day life. Another useful concept, which I don't hear people talk about much is "**subspace optimum**": A point maximizes a function not in the whole space, but in a subspace. You have to move along a different dimension than those of the subspace in order to improve. A subspace optimum doesn't have to be a local optimum either, because even a small change along the new dimension might yield improvements. If you're in a subspace optimum, this requires a different attitude to get to a global optimum, than if you're in a local optimum, which makes me think it's good for the term to be part of every day language.

- **When you're in a local optimum**, you have to do something quite different from what you're doing to improve.
- **When you're in a subspace optimum**, you have to notice **dimensions along which you could** be doing things differently that you didn't even notice before, but small changes along those new dimensions might already help. You're applying constraints to yourself that you could let go.

Regarding how it looks subjectively:

- The phrase: "**am I in a local optimum?**" generates curiosity about whether you maybe should undertake a quite different plan from the one you're taking now. (Should I do a different project, rather than make local changes to the project I'm taking?)
- The phrase: "**am I in a subspace optimum?**" generates curiosity about whether you maybe are not noticing (possibly small) changes you could be making across dimensions you haven't been considering. (Should I optimize/adjust the way I'm doing my project across different dimensions/variables than the ones I've been optimizing over so far?)

My impression is that somewhat often when people informally use the term local optimum, they are in fact talking about a subspace optimum.

Bonus for the theoretically inclined: A local subspace optimum is one where you can improve by temporarily doing things differently along dimension X, moving around in a bigger space, while eventually ending up on a different, better, point in the same subspace.

Why Rationalists Shouldn't be Interested in Topos Theory

I spent a lot of the last two years getting really into [categorical logic](#) (as in, using [category theory](#) to study logic), because I'm really into logic, and category theory seemed to be able to provide cool alternate foundations of mathematics.

Turns out it doesn't really.

Don't get me wrong, I still think it's interesting and useful, and it did provide me with a very cosmopolitan view of logical systems (more on that later). But *category theory is not suitable for foundations or even meant to be foundational*. Most category theorists use an [extended version of set theory](#) as foundations!

In fact, its purpose is best seen as exactly dual to that of foundations: while set theory allows you to build things from the ground up, category theory allows you to organize things from high above. A category by itself is not so interesting; one often studies a category in terms of how it maps from and into other categories (including itself!), with functors, and, most usefully, [adjunctions](#).

Ahem. This wasn't even on topic.

I want to talk about a particular subject in categorical logic, perhaps the most well-studied one, which is *topos theory*, and why I believe it be to useless for rationality, so that others may avoid retreading my path. The thesis of this post is that *probabilities aren't (intuitionistic) truth values*.

Topoi and toposes

A *topos* is perhaps best seen not even as category, but as an alternate mathematical universe. They are, essentially, "weird set theories". Case in point: Set itself is a topos, and other toposes are often constructed as categories of functors $F : C \rightarrow \text{Set}$, for C an arbitrary category.

(Functors assemble into categories if you take *natural transformations* between them. That basically means that you have maps $F(c) \rightarrow G(c)$, such that if you compare the images of a path under F and G , all the little squares commute.)

Consider that natural numbers, with their usual ordering like $4 \leq 5$, can form a category if you take instead $4 \rightarrow 5$. So one simple example is to consider the category of *all* functors $\mathbb{N} \rightarrow \text{Set}$, which are really just sequences of sets, like

$$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow \dots$$

where the arrows are regular set theoretic functions. You can do practically any kind of mathematical reasoning using sequences of sets! (as long as it is [constructive](#)). For example, you have

- an "empty set", which is just a sequence of empty sets;
- a "point" given by a sequence of points;
- "products" of sequences given by $\{X_i \times Y_i\}$;

and so on. Most interestingly, you have [truth values given by subobjects of the point](#); accordingly, in Set those are the empty set and the point itself, since $P(*) = \{\emptyset, *\}$, corresponding to *true* and *false*. Notice that $\emptyset \subseteq *$; in fact the truth values in general will have the structure of a [partially ordered set](#).

What are our truth values here? What is a subobject of a sequence of points? For one, each X_i has to be a subset of $*$. And there are no maps $* \rightarrow \emptyset$; so each "truth value" will look like

$$\emptyset \rightarrow \emptyset \rightarrow \dots \rightarrow \emptyset \rightarrow * \rightarrow * \rightarrow \dots$$

a bunch of empty sets and, at some position n , all points, meaning that we have as many truth values as natural numbers. This is our first glance into the cosmopolitan nature of topos theory: weird truth values! Notice, however, that if $n \leq m$, their corresponding subobjects will have this ordering reversed (an exercise left for the already knowledgeable reader); so in the end it might have been better to use functors on N^{op} , natural numbers with their order reversed.

To sum up, we made the category Set^N of sequences of sets, and realized that it was a topos with truth values N^{op} . Isn't it that interesting...

Topoi-logical

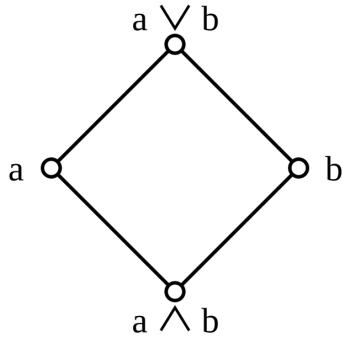
Turns out there's a [big connection](#) between toposes and topological spaces.

The open sets of a topological space have the structure of a partially ordered set, if you set $U \leq V$ whenever $U \subseteq V$. Moreover, in that poset, you can describe $U \cap V$ as the *greatest lower bound* of U and V , and $U \cup V$ as their *least upper bound*.

This is in fact (almost) exactly the structure we want of our topos-theoretic poset of truth values: the greatest lower bound corresponds to conjunction $p \wedge q$ and least upper bound is the disjunction $p \vee q$. So we can use topological spaces as spaces of

truth values, and this is in fact the approach used in [Heyting semantics](#) of intuitionistic logic.

(so each *open set*, and not *point*, corresponds to a truth value; you take the AND of two open sets to be their intersection and the OR to be their union)



Alright, so as with N , the poset of open sets can define a category if you set $U \rightarrow V$ whenever $U \leq V$. So take X a topological space and $O(X)$ its category of open sets.

We'll define a new category $Sh(X)$ of functors $O(X)^{op} \rightarrow \text{Set}$, except that they won't be all of the functors, only those that preserve the topo/logical structure ([sheaves](#)).

Guess what? Not only is $Sh(X)$ a topos, turns out that its truth values are isomorphic to $O(X)$! And since the truth values are the subobjects of the point, that means that the points of the topos are in fact shaped like X ...

Now we have a fuller view of the logically cosmopolitan view that topos theory can bring us. You can create, just like that, a whole parallel mathematical universe of bizarro sets where everything is made up of, say, donuts. Or coffee cups. It is as you wish.

Where's my Bayesian topos?

Since I am at heart a LessWronger, and since I care deeply about problems of logical induction and logical counterfactuals and whatnot, I spent a while trying to design a topos that would behave like manipulating probabilities. Or distributions. Or something. With the objective of making something that would represent beliefs.

Well, I'm sorry, but it doesn't work.

At minimum, we would expect that the truth values of this topos be probabilities, yeah? And with the cosmopolitan principle above, we could then just take the sheaves on this poset of probabilistic truth values.

So these truth-values would be order-isomorphic to $[0,1]$. But for them to actually represent probabilities, we'd want that $p \wedge q = pq$, and yet the order on $[0, 1]$ already prescribes that $p \wedge q = \min(p, q)$, and we are doomed from the start.

Furthermore, even in an intuitionistic logic, the provable statements all have the maximal truth value (which here would be 1); but we all know that [0 and 1 are not probabilities](#), and so nothing should be provable... which seems like it wouldn't be very useful.

All in all, I'm truly sorry you had to bear through all of the math above just for this conclusion. It's still pretty cool, though, right?

(Geometric) topoi aren't reflective

In order to legitimately use topos theory for rationality, we should have a way for the topos to "think about itself". Analogously to the [situation in Peano arithmetic](#), for a topos E , we'd want some object $E \in E$ (specifically, an [internal category](#)) to be isomorphic to E in some sense.

We can define an "element" of an object $E \in E$ as being an arrow $* \rightarrow E$. So the objects of the internal category are given by the set $\text{Hom}(*, E)$, and in fact the functor $\text{Hom}(*, -) : E \rightarrow \text{Set}$ respects the structure of the internal category enough that it becomes a category internal to Set , which is just a small category.

But wait. The toposes generated by sheaves on a topological space are at least as big as Set , but the collection of all sets is too big to be a set, and thus we run into size issues.

It should in principle be possible to do so in small toposes, such as the [free \(as in syntactic\) topos](#), but I am not sure and will refrain from claiming so. It is however certainly possible to do so in [list-arithmetic pretoposes](#) (yes it's a mouthful), as shown by André Joyal in his as of yet unpublished categorical proof of Gödel's incompleteness theorems, which I have studied with him last year.

What now?

It now seems to me that [linear logic](#) might be the "right" weakening of classical logic into something probabilistic. I still need to figure out some of the details, but let's say [that the work has already been done](#), and one need only piece it together into something relevant to rationality and agent foundations. Particularly promising is that some claim that linear logic is a good setting for "paraconsistent" logic (logic that deals gracefully with contradictions), which could make it work for logical counterfactuals.

All this and more in my next post, pretentiously monikered "Probability Monads".

How uniform is the neocortex?

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

How uniform is the neocortex?

The neocortex is the part of the human brain responsible for higher-order functions like sensory perception, cognition, and language, and has been hypothesized to be uniformly composed of general-purpose data-processing modules. What does the currently available evidence suggest about this hypothesis?

*"How uniform is the neocortex?" is one of the background variables in my [framework for AGI timelines](#). My aim for this post is **not** to present a complete argument for some view on this variable, so much as it is to:*

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

There's a [long list of different regions in the neocortex](#), each of which appears to be responsible for something totally different. One interpretation is that these cortical regions are doing fundamentally different things, and that we acquired the capacities to do all these different things over hundreds of millions of years of evolution.

A radically different perspective, first put forth by Vernon Mountcastle in 1978, hypothesizes that the neocortex is implementing a single general-purpose data processing algorithm all throughout. From the popular neuroscience book *On Intelligence*, by Jeff Hawkins^[1]:

[...] Mountcastle points out that the neocortex is remarkably uniform in appearance and structure. The regions of cortex that handle auditory input look like the regions that handle touch, which look like the regions that control muscles, which look like Broca's language area, which look like practically every other region of the cortex. Mountcastle suggests that since these regions all look the same, perhaps they are actually performing the same basic operation! He proposes that the cortex uses the same computational tool to accomplish everything it does.

[...]

Mountcastle [...] shows that despite the differences, the neocortex is remarkably uniform. The same layers, cell types, and connections exist throughout. [...] The differences are often so subtle that trained anatomists can't agree on them. Therefore, Mountcastle argues, all regions of the cortex are performing the same operation. The thing that makes the vision area visual and the motor area motoric is how the regions of cortex are connected to each other and to other parts of the central nervous system.

In fact, Mountcastle argues that the reason one region of cortex looks slightly different from another is because of what it is connected to, and not because its basic function is different. He concludes that there is a common function, a common algorithm, that is performed by all the cortical regions. Vision is no different from hearing, which is no different from motor output. He allows that our genes specify how the regions of cortex are connected, which is very specific to function and species, but the cortical tissue itself is doing the same thing everywhere.

If Mountcastle is correct, the algorithm of the cortex must be expressed independently of any particular function or sense. The brain uses the same process to see as to hear. The cortex does something universal that can be applied to any type of sensory or motor system.

The rest of this post will review some of the evidence around Mountcastle's hypothesis.

Cortical function is largely determined by input data

When visual inputs are fed into the auditory cortices of infant ferrets, those auditory cortices [develop into functional visual systems](#). This suggests that different cortical regions are all capable of general-purpose data processing.

Humans can learn how to perform forms of sensory processing we haven't evolved to perform—blind people can learn to [see with their tongues](#), and can learn to echolocate well enough to [discern density and texture](#). On the flip side, forms of sensory processing that we *did* evolve to perform depend heavily on the data we're exposed to—for example, [cats exposed only to horizontal edges early in life don't have the ability to discern vertical edges later in life](#). This suggests that our capacities for sensory processing stem from some sort of general-purpose data processing, rather than innate machinery handed to us by evolution.

Blind people who learn to echolocate do so with the help of [repurposed visual cortices](#), and they can [learn to read Braille using repurposed visual cortices](#). Our visual cortices did not evolve to be utilized in these ways, suggesting that the visual cortex is doing some form of general-purpose data processing.

There's a man who had the entire left half of his brain removed when he was 5, who has [above-average intelligence](#), and went on to [graduate college and maintain steady employment](#). This would only be possible if the right half of his brain were capable of taking on the cognitive functions of the left half of the brain.

The patterns identified by the primary sensory cortices (for vision, hearing, and seeing) [overlap substantially](#) with the patterns that numerous different unsupervised learning algorithms identified from the same data, suggesting that the different cortical regions (along with the different unsupervised learning algorithms) are all just doing some form of general-purpose pattern recognition on its input data.

Deep learning and cortical generality

The above evidence does not rule out the possibility that the cortex's apparent adaptability stems from developmental triggers, rather than some capability for general-purpose data-processing. By analogy, stem cells all start out very similar, only to differentiate into cells with functions tailored to the contexts in which they find themselves. It's possible that different cortical regions have hard-coded genomic responses for handling particular data inputs, such that the cortex gives one hard-coded response when it detects that it's receiving visual data, another hard-coded response when it detects that it's receives auditory data, etc.

If this were the case, the cortex's data-processing capabilities can best be understood as [specialized responses to distinct evolutionary needs](#), and our ability to process data that we haven't evolved to process (e.g. being able to look at a Go board and intuitively discern what a good next move would be) most likely utilizes a complicated mishmash of heterogeneous data-processing abilities acquired over evolutionary timescales.

Before I learned about any of the advancements in deep learning, this was my most likely guess about how the brain worked. It had always seemed to me that the hardest and most mysterious part of intelligence was intuitive pattern-recognition, and that the various forms of intuitive processing that let us recognize images, say sentences, and play Go might be totally different and possibly arbitrarily complex.

So I was very surprised when I learned that a single general method in deep learning (training an artificial neural network on massive amounts of data using gradient descent)^[2] led to performance comparable or superior to humans' in tasks as disparate as [image classification](#), [speech synthesis](#), and [playing Go](#). I found superhuman Go performance particularly surprising—intuitive judgments of Go boards encode distillations of high-level strategic reasoning, and are highly sensitive to small changes in input. Neither of these is true for sensory processing, so my prior guess was that the methods that worked for sensory processing wouldn't have been sufficient for playing Go as well as humans.^[3]

This suggested to me that there's nothing fundamentally complex or mysterious about intuition, and that seemingly-heterogeneous forms of intuitive processing can result from simple and general learning algorithms. From this perspective, it seems most parsimonious to explain the cortex's seemingly general-purpose data-processing capabilities as resulting straightforwardly from a general learning algorithm implemented all throughout the cortex. (This is *not* to say that I think the cortex is doing what artificial neural networks are doing—rather, I think deep learning provides evidence that general learning algorithms exist *at all*, which increases the prior likelihood on the cortex implementing a general learning algorithm.^[4])

The strength of this conclusion hinges on the extent to which the "artificial intuition" that current artificial neural networks (ANNs) are capable of is analogous to the intuitive processing that humans are capable of. It's possible that the "intuition" utilized by ANNs is deeply analogous to human intuition, in which case the generality of ANNs would be very informative about the generality of cortical data-processing. It's also possible that "artificial intuition" is different in kind from human intuition, or that it only captures a small fraction of what goes into human intuition, in which case the generality of ANNs would not be very informative about the generality of cortical data-processing.

It seems that experts are divided about how analogous these forms of intuition are, and I conjecture that this is a major source of disagreement about overall AI timelines.

[Shane Legg](#) (a cofounder of DeepMind, a leading AI lab) has been talking about how [deep belief networks might be able to replicate the function of the cortex](#) before deep learning took off, and he's [been predicting human-level AGI in the 2020s since 2009](#). Eliezer Yudkowsky has directly talked about AlphaGo providing evidence of ["neural algorithms that generalize well, the way that the human cortical algorithm generalizes well"](#) as an indication that AGI might be near. [Rodney Brooks](#) (the former director of MIT's AI lab) has written about how deep learning is not capable of [real perception or manipulation](#), and thinks AGI is [over 100 years away](#). [Gary Marcus](#) has described deep learning as a "[wild oversimplification](#)" of the "[hundreds of anatomically and likely functionally \[distinct\] areas](#)" of the cortex, and estimates AGI to be [20-50 years away](#).

Canonical microcircuits for predictive coding

If the cortex were uniform, what might it *actually be doing* uniformly?

The cortex has been hypothesized to consist of [canonical microcircuits that implement predictive coding](#). In a nutshell, predictive coding (aka predictive processing) is a theory of brain function which hypothesizes that the cortex learns hierarchical structure of the data it receives, and uses this structure to encode predictions about future sense inputs, resulting in "controlled hallucinations" that we interpret as direct perception of the world.

On Intelligence has an excerpt that cleanly communicates what I mean by "learning hierarchical structure":

[...] The real world's nested structure is mirrored by the nested structure of your cortex.

What do I mean by a nested or hierarchical structure? Think about music. Notes are combined to form intervals. Intervals are combined to form melodic phrases. Phrases are combined to form melodies or songs. Songs are combined into albums. Think about written language. Letters are combined to form syllables. Syllables are combined to form words. Words are combined to form clauses and sentences. Looking at it the other way around, think about your neighborhood. It probably contains roads, schools, and houses. Houses have rooms. Each room has walls, a ceiling, a floor, a door, and one or more windows. Each of these is composed of smaller objects. Windows are made of glass, frames, latches, and screens. Latches are made from smaller parts like screws.

Take a moment to look up at your surroundings. Patterns from the retina entering your primary visual cortex are being combined to form line segments. Line segments combine to form more complex shapes. These complex shapes are combining to form objects like noses. Noses are combining with eyes and mouths to form faces. And faces are combining with other body parts to form the person who is sitting in the room across from you.

All objects in your world are composed of subobjects that occur consistently together; that is the very definition of an object. When we assign a name to something, we do so because a set of features consistently travels together. A face is a face precisely because two eyes, a nose, and a mouth always appear together. An eye is an eye precisely because a pupil, an iris, an eyelid, and so on,

always appear together. The same can be said for chairs, cars, trees, parks, and countries. And, finally, a song is a song because a series of intervals always appear together in sequence.

In this way the world is like a song. Every object in the world is composed of a collection of smaller objects, and most objects are part of larger objects. This is what I mean by nested structure. Once you are aware of it, you can see nested structures everywhere. In an exactly analogous way, your memories of things and the way your brain represents them are stored in the hierarchical structure of the cortex. Your memory of your home does not exist in one region of cortex. It is stored over a hierarchy of cortical regions that reflect the hierarchical structure of the home. Large-scale relationships are stored at the top of the hierarchy and small-scale relationships are stored toward the bottom.

The design of the cortex and the method by which it learns naturally discover the hierarchical relationships in the world. You are not born with knowledge of language, houses, or music. The cortex has a clever learning algorithm that naturally finds whatever hierarchical structure exists and captures it.

The clearest evidence that the brain is learning hierarchical structure comes from the visual system. The visual cortex is known to have [edge detectors at the lowest levels of processing](#), and neurons that fire when shown [images of particular people](#), like Bill Clinton.

What does predictive coding say the cortex does with this learned hierarchical structure? From [an introductory blog post about predictive processing](#):

[...] the brain is a multi-layer prediction machine. All neural processing consists of two streams: a bottom-up stream of sense data, and a top-down stream of predictions. These streams interface at each level of processing, comparing themselves to each other and adjusting themselves as necessary.

The bottom-up stream starts out as all that incomprehensible light and darkness and noise that we need to process. It gradually moves up all the cognitive layers that we already knew existed – the edge-detectors that resolve it into edges, the object-detectors that shape the edges into solid objects, et cetera.

The top-down stream starts with everything you know about the world, all your best heuristics, all your priors, [all the structure you've learned,] everything that's ever happened to you before – everything from "solid objects can't pass through one another" to " $e=mc^2$ " to "that guy in the blue uniform is probably a policeman". It uses its knowledge of concepts to make predictions – not in the form of verbal statements, but in the form of expected sense data. It makes some guesses about what you're going to see, hear, and feel next, and asks "Like this?" These predictions gradually move down all the cognitive layers to generate lower-level predictions. If that uniformed guy was a policeman, how would that affect the various objects in the scene? Given the answer to that question, how would it affect the distribution of edges in the scene? Given the answer to that question, how would it affect the raw-sense data received?

As these two streams move through the brain side-by-side, they continually interface with each other. Each level receives the predictions from the level above it and the sense data from the level below it. Then each level uses Bayes' Theorem to integrate these two sources of probabilistic evidence as best it can.

[...]

"To deal rapidly and fluently with an uncertain and noisy world, brains like ours have become masters of prediction - surfing the waves and noisy and ambiguous sensory stimulation by, in effect, trying to stay just ahead of them. A skilled surfer stays 'in the pocket': close to, yet just ahead of the place where the wave is breaking. This provides power and, when the wave breaks, it does not catch her. The brain's task is not dissimilar. By constantly attempting to predict the incoming sensory signal we become able [...] to learn about the world around us and to engage that world in thought and action."

The result is perception, which the PP theory describes as "controlled hallucination". You're not seeing the world as it is, exactly. You're seeing your predictions about the world, cashed out as expected sensations, then shaped/constrained by the actual sense data.

An illustration of predictive processing, from the same source:

This demonstrates the degree to which the brain depends on top-down hypotheses to make sense of the bottom-up data. To most people, these two pictures start off looking like incoherent blotches of light and darkness. Once they figure out what they are ([spoiler](#)) the scene becomes obvious and coherent. According to the predictive processing model, this is how we perceive everything all the time - except usually the concepts necessary to make the scene fit together come from our higher-level predictions instead of from clicking on a spoiler link.

Predictive coding has been hailed by [prominent neuroscientists](#) as a possible [unified theory of the brain](#), but I'm confused about how much physiological evidence there is that the brain is actually implementing predictive coding. It seems like there's physiological evidence in support of predictive coding being implemented [in the visual cortex](#) and in the [auditory cortex](#), and there's a [theoretical account](#) of how the [prefrontal cortex](#) (responsible for higher cognitive functions like planning, decision-making, and executive function) might be utilizing similar principles. [This paper](#) and [this paper](#) review some physiological evidence of predictive coding in the cortex that I don't really know how to interpret.

My current take

I find the various pieces of evidence that cortical function depends largely on data inputs (e.g. the ferret rewiring experiment) to be pretty compelling evidence of general-purpose data-processing in the cortex. The success of simple and general methods in deep learning across a wide range of tasks suggests that it's most parsimonious to model the cortex as employing general methods throughout, but only to the extent that the capabilities of artificial neural networks can be taken to be analogous to the capabilities of the cortex. I currently consider the analogy to be deep, and intend to explore my reasons for thinking so in future posts.

I think the fact that predictive coding offers a plausible theoretical account for what the cortex could be doing uniformly, which can account for higher-level cognitive functions in addition to sensory processing, is itself some evidence of cortical uniformity. I'm confused about how much physiological evidence there is that the

brain is actually implementing predictive coding, but I'm very bullish on predictive coding as a basis for a unified brain theory based on non-physiological evidence (like our subjective experiences making sense of the images of splotches) that I intend to explore in a future post.

Thanks to Paul Kreiner, David Spivak, and Stag Lynn for helpful suggestions and feedback, and thanks to Jacob Cannell for writing [a post](#) that inspired much of my thinking here.

1. [This blog post](#) comment has some good excerpts from *On Intelligence*. ↵
2. Deep learning is a general method in the sense that most tasks are solved by utilizing a handful of basic tools from a standard toolkit, adapted for the specific task at hand. Once you've selected the basic tools, all that's left is figuring out how to supply the training data, specifying the objective that lets the AI know how well it's doing, throwing a lot of computation at the problem, and fiddling with details. My understanding is that there typically isn't much conceptual ingenuity involved in solving the problems, that most of the work goes into fiddling with details, and that trying to be clever [doesn't lead to better results than using standard tricks with more computation and training data](#). It's also worth noting that most of the tools in this standard toolkit have been around since the 90's (e.g. convolutional neural networks, LSTMs, reinforcement learning, backpropagation), and that the recent boom in AI was driven by using these decades-old tools with unprecedented amounts of computation. ↵
3. AlphaGo did simulate future moves to achieve superhuman performance, so the direct comparison against human intuition isn't completely fair. But AlphaGo Zero's raw neural network, which just looks at the "texture" of the board without simulating any future moves, can still play quite formidably. From the [AlphaGo Zero paper](#): "The raw neural network, without using any lookahead, achieved an Elo rating of 3,055. AlphaGo Zero achieved a rating of 5,185, compared to 4,858 for AlphaGo Master, 3,739 for AlphaGo Lee and 3,144 for AlphaGo Fan." (AlphaGo Fan beat the European Go champion 5-0.) ↵
4. Eliezer Yudkowsky has an insightful exposition of this point in [a Facebook post](#). ↵

"AI and Efficiency", OA (44× improvement in CNNs since 2012)

This is a linkpost for <https://openai.com/blog/ai-and-efficiency/>

Abstract:

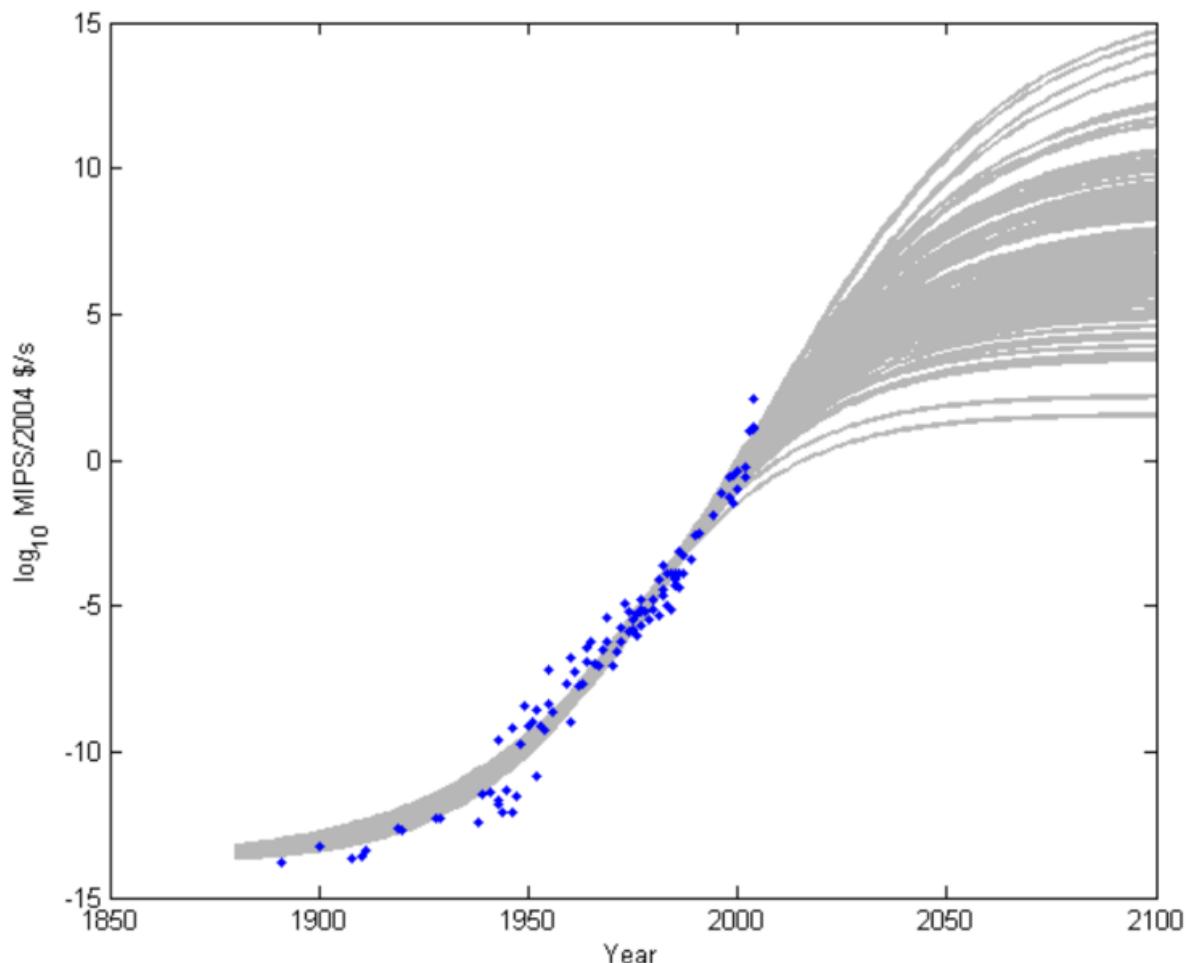
We're releasing an analysis showing that since 2012 the amount of compute needed to train a neural net to the same performance on ImageNet¹ classification has been decreasing by a factor of 2 every 16 months. Compared to 2012, it now takes 44 times less compute to train a neural network to the level of AlexNet² (by contrast, Moore's Law³ would yield an 11x cost improvement over this period). Our results suggest that for AI tasks with high levels of recent investment, algorithmic progress has yielded more gains than classical hardware efficiency.

Maths writer/cowriter needed: how you can't distinguish early exponential from early sigmoid

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

It's [well known](#) in FHI and similar circles, that it's impossible to distinguish an exponential (growth going up wildly) from a sigmoid/logistic curve (exponential growth until a turning point - an S shape) - until well after the turning point.

Which means we can't effectively predict that turning point. And so can't distinguish when a sigmoid will have a turning point, even when we know it must have one.



But this doesn't seem to exist in the statistics literature; and it would be very useful to have such a paper or textbook to point to.

We don't have time to write a full paper ourselves, but is there someone on this list with statistical experience who would like to write or co-write such a paper?

Since this result is important and as yet unpublished, it's plausible that such a publication may get an extremely high number of citations.

Cheers!

The Oil Crisis of 1973

Last month I investigated [commonalities between recessions](#) of the last 50 years or so. But of course this recession will be different, because (among other things) we will simultaneously have a labor shortage and a lot of people out of work. That's really weird, and there's almost no historical precedent- the 1918 pandemic took place during a war, and neither 1957 nor 1968 left enough of an impression to have a single book dedicated to them.

So I expanded out from pandemics, and started looking for recessions that were caused by any kind of exogenous shock. The best one I found was the [1973 Oil Crisis](#). That was kicked off by Arab nations refusing to ship oil to allies who had assisted Israel during the [Yom Kippur war](#)- as close as you can get to an economic impact without an economic cause. I started to investigate the 1973 crisis as the one example I could find of a recession caused by a sudden decrease in a basic component of production, for reasons other than economic games.

Spoiler alert: that recession was not caused by a sudden decrease in a basic component of production either.

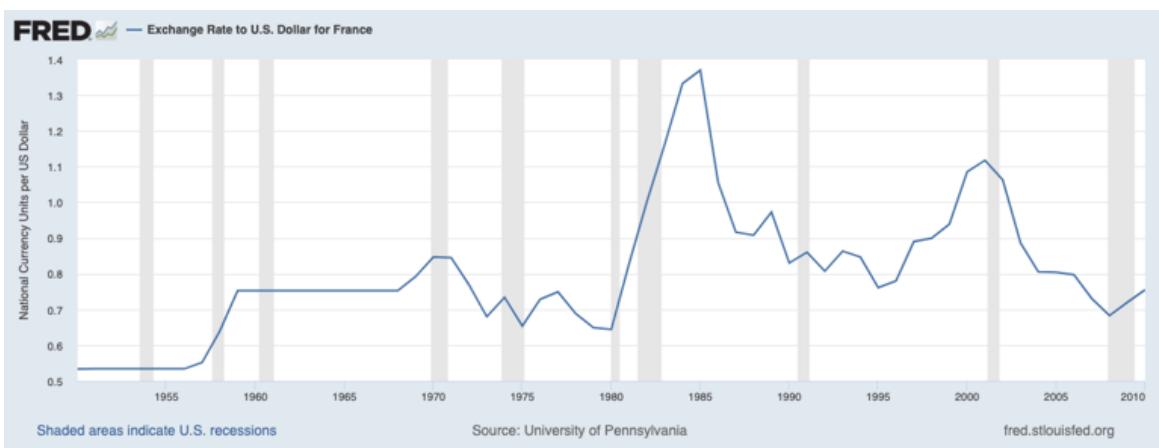
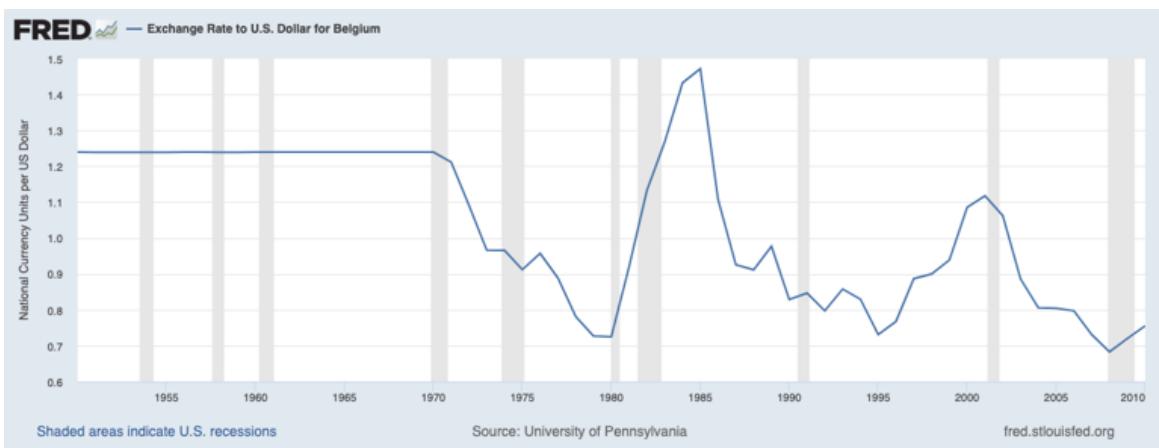
Why am I so sure of this? Here's a short list of little things,

- The embargo was declared [October 17th](#), but the price of oil did not really spike until [January 1st](#).
- The [price of food spiked two months](#) before the embargo was declared and plateaued before oil prices went up.
- A multiyear stock market crash started in [January 1973](#), 9 months before embargo was declared.
- Previous oil embargoes had been attempted in [1956 and 1967](#), to absolutely no effect.

But here's the big one: we measure the price of oil in USD. That's understandable, since oil sales are legally required to be denominated in dollars. But the US dollar underwent a massive overhaul in 1971, when America decided it was tired of some parts of the [Bretton Woods Agreement](#). Previously, the US, Japan, Canada, Australia and many European countries maintained peg (set exchange rate) between all other currencies and USD, which was itself pegged to gold. In 1971 the US decided not to bother with the gold part anymore, causing other countries to break their peg. I'm sure why we did this is also an interesting story, but I haven't dug into it yet, because what came after 1971 is interesting enough. The currency of several countries appreciated noticeably (Germany, Switzerland, Japan, France, Belgium, Holland, and Sweden)...

(I apologize for the inconsistent axes, they're the best I could do)

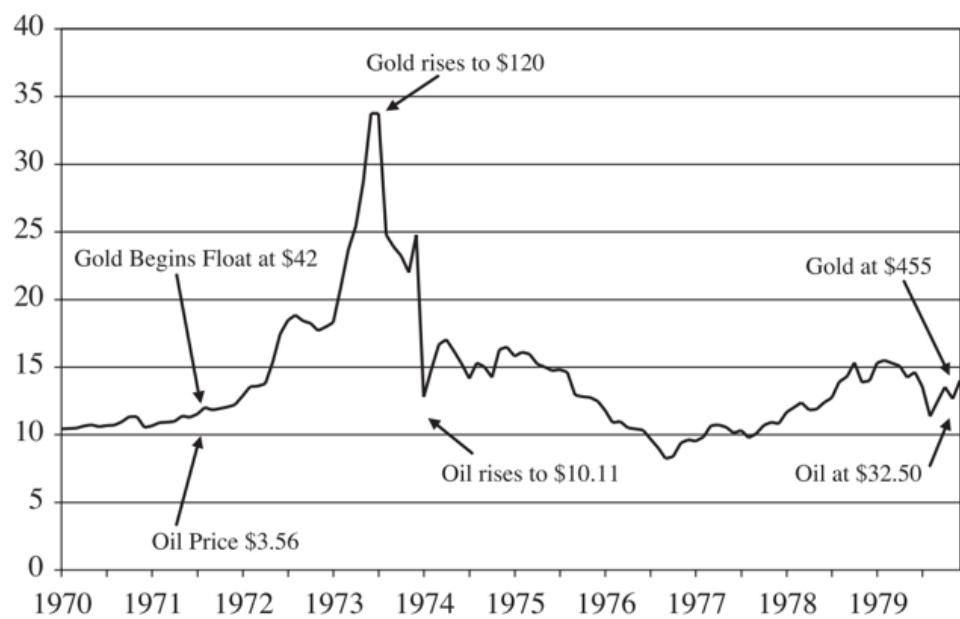




...but as I keep harping on, oil prices were denominated in dollars. This meant that oil producing countries, from their own perspective, were constantly taking a pay cut. Denominated in USD, 1/1/74 saw a huge increase in the price of oil. Denominated in gold, 1/1/74 saw a return to the historic average after an unprecedented low.



Figure 1
Barrels of Oil per Ounce of Gold



(apologies for these axes too- the spike in this graph means oil was was worth less, because you could buy more with the same amount of gold)

This is a little confusing, so here's a timeline:

- 1956: Failed attempt at oil embargo
- 1967: Failed attempt at oil embargo
- 1971, August: US leaves the gold standard
- 1972: Oil prices begin to fall, relative to gold
- 1972, December: US food prices begin to increase the rate of price increases.
- 1973, January: US Stock market begins 2-year crash
- 1973, August: US food prices begin to go up *really* fast
- 1973, October, 6: Several nearby countries invade Israel
- 1973, October, 17: Several Arab oil producing countries declare an embargo against Israeli allies, and a production decrease. Price of oil goes up a little (in USD).
- 1974, January, 1: Effective date of declared price increase from \$5.12 to \$11.65/barrel. Oil returns to historically normal price measured in gold.

This is not the timeline you'd expect to see if the Yom Kippur war caused a supply shock in oil, leading to a recession.

My best guess is that something was going wrong in the US and world economy well before 1971, but the market was not being allowed to adjust. Breaking Bretton Woods took the finger out of the dyke and everything fluctuated wildly for a few years until the world reached a new equilibrium (including some new and different economic games). The Yom Kippur war was a catalyst or excuse for raising the price of oil, but not the cause.

Thanks to my [Patreon](#) subscribers for funding this research, and several reviewers for checking my research and writing.

162 benefits of coronavirus

[More added: now 183]

WHILE THE HUGE harms of coronavirus are well-known – death, illness, lockdowns, unemployment, recession, etc. – less attention has understandably been paid to the benefits.

Even clouds this dark have silver linings. Crises produce opportunities, innovation, and long-overdue reforms. 2020 will contain an [extra year's worth](#) of mortality – but also a decade's worth of progress, a leap into the future.

This post lists many benefits that could arise, so readers can consider how to maximize them, not just minimize harms. They cover a wide range of consequences. For example, lockdown has made many people (even [drug gangsters](#)) reassess their lives. Working from home has suddenly become normal, with less commuting, less cost, and more time for leisure and sleep. So people may move from cities to cheaper, more pleasant areas, or indeed countries.

Above all, coronavirus is a wake-up call - it could have been [far worse](#). Better preparation for the next pandemic will reduce [existential risk](#), potentially saving billions, or even trillions, of future lives.

Experiment & evolve

Lockdowns have created an experiment, making people and organisations re-think how - and why - they do things. Some activities become impossible and are abandoned, e.g. travel. For others, alternatives are tried, e.g. video calls for meetings and doctor's appointments; or innovations, such as businesses [sharing employees](#). This experimentation will continue well beyond lockdown, as the new reality emerges.

Many of these changes will turn out to be improvements, and will stick. Others, e.g. [government-funded furloughing](#) and [virtual horse races](#), are temporary fixes which will go - as will changes that didn't work. And things that were dropped as unnecessary, e.g. pointless meetings and regulations, will stay dropped.

All of this involves prioritizing: deciding what outcomes matter, and which solutions now work best. Many things will modernize, simplify, and become more efficient. Cost-effectiveness is key, as incomes will shrink for a while.

Finally, lessons will be learned from what went badly in the pandemic, and steps taken to improve resilience and prepare for future crises.

We can also view the situation in terms of evolution. The world has been struck by a metaphorical meteor, threatening not just lives, but ways of life. Those organisations, jobs, and activities that are fittest for the new environment, or can adapt, will survive. Others that are no longer useful will die out, often replaced by innovations, to produce a new normal.

The benefits

The list below contains all the potential long-term benefits of the pandemic that I could find or think of. No doubt it is somewhat focused on rich countries, though this is not the aim. Please suggest additions or changes in the comments.

Some benefits have started under lockdown, such as more volunteering. Others may come later, such as de-urbanization.

Some are mixed blessings, causing substantial harm as well; e.g. failures of non-viable businesses, charities and educational institutions. With some items it's unclear, or a matter of opinion, whether it is a benefit or not, e.g. political changes. While many potential benefits are speculative, some are especially so - more hopes than predictions; e.g. better international cooperation, in reaction to the protectionism of the pandemic. So I've qualified some entries accordingly:

- ± Benefit with substantial harm, or unclear whether it's a net benefit at all
- ? Very speculative

Disaster preparedness

Governments:

- Preparation for future pandemics
- ?Planning for existential risks: if coronavirus prompts even a small improvement in this, it would vastly outweigh all of the pandemic's harms*

Businesses & other organisations:

- Continuity planning
- Insurance
- Better contractual arrangements, e.g. force majeure clauses
- More robust supply chains, e.g. less just-in-time manufacturing
- ±Re-shoring

Individuals:

- Saving
- Insurance
- ±Survivalism

Spare capacity & redundancy:

- Essential services: e.g. healthcare, supermarkets
- Critical infrastructure
- Manufacturing
- Stockpiling essential supplies: e.g. food, fuel, medicines, vaccines, PPE

Government

Welfare state:

- ±Increased safety net for healthcare, unemployment, etc.
- ±Calls for Universal Basic Income: due to government-funded furloughing in some countries during lockdowns

Digitization & modernization:

- [E-government](#)
- Faster processing of benefit applications
- Remote operation & streamlining of courts
- Remote operation of parliaments
- Electronic & postal voting

Trust in government in some countries: due to effective pandemic control, job retention schemes, etc.

Change of government/leader in some countries: if they did not handle pandemic well

Less avoidance of tax and regulations, as a result of [re-shoring](#)

Cost-saving efficiencies due to higher debt & lower tax revenue

?Transparency of government

?More constructive national politics

?Improved international cooperation, e.g.:

- World Health Organisation
- Trade in essentials, e.g. food, energy, medical supplies
- Disaster preparedness (see above)

?Foreign aid:

- Healthcare aid
- Suspend debts
- ±Cancel debts

?Ceasefires during pandemic in conflict zones, perhaps continuing afterwards

?Wellbeing/happiness economics take-up, as the pandemic highlights dilemmas between lives, livelihoods, and quality of life

Health & science

Public healthcare funding:

- For spare capacity (see **Disaster preparedness**)
- ?Policies to promote health, diet & exercise, particularly to groups which had disproportionate coronavirus mortality
- ?Care home funding

International collaboration on health research

Faster health research processes:

- Disease research
- Vaccine & drug development
- ±Deregulation

- Journal publishing

Advances in virology, epidemiology, sociology etc. from coronavirus research

Infectious disease reduction, due to long-term hygiene improvements (e.g. handwashing, ?face masks):

- Common diseases, e.g. colds, flu, food poisoning
- Rarer, more serious diseases, e.g. some cancers
- Diseases not previously known to be infectious
- Future pandemics

Telehealth, including:

- Symptom checking apps
- Video consultations & therapy
- Online prescribing
- Treatments sent by post
- Online self-help & automated therapy, e.g. for mental health
- Remote monitoring

Hence:

Digitization of health data:

- More efficient, e.g. the UK's NHS still relies on paper records
- Enables research on the data

Self-care:

- Physical health: importance highlighted by increased coronavirus mortality (even though lockdown may produce shorter-term harms from alcohol and less exercise)
- Mental health: importance highlighted by lockdowns and anxiety about health & jobs
- More cycling & walking: to avoid infection risk on public transport
- Sleep: improved by less commuting or shorter work hours (see **Work**)
- Personal health trackers: more usage & features, e.g. measuring temperature

Trust in science and medicine

Work

Remote work (usually office jobs):

- From home; cafes, shared workspaces etc. nearby; or while travelling elsewhere
- Move home to better/cheaper area or country, or to be nearer family/friends (see **Relocation & transport**)
- Saves office cost, commuting time & cost
- **Digital transformation** of organisations, increasing efficiency
- More international employment & collaboration
- More work for disabled people
- Less 'presenteeism' (unnecessary attendance at work)
- Better rural Internet access

- ?Less office politics: as harder to do remotely
- ?VR headsets for remote meetings, etc.

Change in work hours to suit worker (e.g. after reflection during lockdown), as cost/job-saving measure by employer, or to enable social distancing in workplaces/transport:

- [Flexi-time](#)
- ±Shorter hours / part-time work
- ±Shifts
- ±Weekends
- Remote workers paid for actions & results, not hours: as hours harder to track

Change of job/career:

- After reflection during lockdown
- ±Forced by unemployment

±Bullshit jobs cut

±Automation of jobs: as cost-saving measure, or to reduce risk of worker absence in future lockdowns/crises

Fewer, more efficient meetings: as video conferences, or due to simplifications under lockdown

Corporate eLearning

?Better worker terms/rights:

- Essential workers' pay
- Casual workers
- Minimum wage
- Sick leave

Business

Innovation to deal with new circumstances, compete for reduced demand, or cut costs, e.g.:

- Extended supermarket hours, or dedicated hours for vulnerable groups, to reduce crowding (even long-term, if further pandemic waves are expected)
- [Sharing employees](#) between different businesses, in response to changes of demand
- [Drive-in cinemas](#) for social distancing
- Use of technology

Retail:

- ±More online groceries, Amazon, Alibaba, Deliveroo, etc.
- Automated warehouses and delivery (see **Relocation & transport**) to fulfill increased online orders
- More self-checkout in physical stores, to avoid infection risk
- Checkout-less stores, e.g. [Amazon Go](#)

- High Street/Main Street switch from products to services: due to competition from online retail

±Business failures - especially if barely viable even before the pandemic, or have crowded spaces, e.g.:

- Restaurants, cafes, bars, pubs, hotels
- Cinemas, theatres
- Department stores
- Airlines, cruise ships

Relocation & transport

De-urbanization due to remote work (see **Work**):

- Lower urban property/real estate prices
- Lower commercial property/real estate prices, due to less office usage
- Lower inequalities between regions of countries

Remote workers moving country:

- To cheaper or more desirable locations
- ?Better governance, tax breaks, etc. to attract such workers
- ?Lower inequalities between countries

Re-shoring (see **Disaster preparedness**)

Less transport:

- ±Less international freight: due to **deglobalization**
- Less work travel: due to less commuting, fewer in-person meetings & conferences, re-shoring (see **Work**)
- ±Less public transport: due to infection risk and restrictions on international travel (even long-term, if further pandemic waves expected)
- Hence more cycling & walking
- Less driving to stores: due to online shopping, infection risk in malls (though de-urbanization may increase some driving)
- Less pollution (see **Environment & nature**)
- ±Lower fuel prices
- ±Staycations: replacing foreign travel
- ?Fewer road deaths - though train/bus passengers may switch to cars

?**Delivery drones, self-driving vehicles, etc.** to fulfill increased online orders

Environment & nature

Pollution:

- Less CO2 and air pollution
- ?Awareness of noise pollution: after urban silence & birdsong under lockdown
- ?Increased climate change concern

Animals:

- Reduction/banning of wild animal capture & sale
- Better conditions in live animal markets
- ?Better animal farming conditions
- Happier & healthier pets, as get more attention from home workers

Outdoor activities as social distancing measure (even long-term, if further pandemic waves are expected):

- Visiting parks, gardens, playgrounds, countryside
- Camping, hiking, fishing, boating, cycling, etc.
- Outdoor sports, swimming pools, gyms
- Open-air bars, restaurants, cafes
- Open-air concerts, cinemas, theatres

Education

Home schooling:

- ±Part-time: to enable social distancing in schools (even long-term, if further pandemic waves are expected)
- ?±Full-time
- Better parental understanding of children's education, due to home schooling during lockdown

Distance learning:

- To support home schooling
- Online university courses
- ?Online exams

Re-assessment of education & educational institutions, including:

- ?What they are for
- In-person vs distance learning
- ?Private school & university fees

±Bankruptcies of some educational institutions

?±More continuous assessment following exam cancellations in e.g. UK

Adult education started under lockdown, e.g. learning an instrument or language

Leisure

More leisure time if stop commuting, or work shorter hours (see **Work**)

Entertainment tried/increased under lockdown, e.g.:

- Arts & culture: music, reading, podcasts, painting, etc.
- ±TV & video streaming: e.g. replacing cinema
- Games, puzzles & quizzes
- Web surfing
- ±Social media

Other pursuits & hobbies tried/increased under lockdown, e.g.

- Cooking
- Takeaways / Deliveroo: e.g. replacing restaurants
- Exercise
- DIY / home improvement
- Spring cleaning / decluttering
- Gardening
- Crafts
- Knitting & sewing
- Adult education (see **Education**)
- Self-improvement / personal development
- Meditation
- ±Prayer / worship

More online entertainment, e.g. live events, reaching wider audiences

Relationships

Some relationships improved/renewed by lockdown:

- With partner
- With children
- With other family members, e.g. via video call
- Friendships via video call, social media, etc.

New online friendships/relationships under lockdown

More time with partner, family & friends if stop commuting, or work shorter hours (see **Work**)

±Divorce / break-up, brought to a head by lockdown

Charity & community

Volunteering, e.g. started under lockdown

?More charitable donations / philanthropy

Cost-saving efficiencies if donations fall due to lower incomes

Innovation to deal with new circumstances or cut costs

Support for local community & businesses: e.g. due to home workers spending more time where they live

±Charity closures - hopefully counterproductive or low effectiveness ones

Perspective

Re-evaluation of life, including:

- Meaning, purpose & values
- Priorities & inessentials
- Likes & dislikes
- Own strengths & weaknesses
- Opportunities & concerns
- Death
- Health: physical & mental
- Relationships
- Work, and work-life balance
- Money

Appreciation of:

- Essential services and key workers, e.g. in healthcare, social care, supermarkets, teaching, technology, mail & deliveries, transport, police
- Role and importance of government, science, media, business and charities
- Volunteers and helpful people
- The elderly
- Activities missed during lockdown, e.g. social contact, culture, sports, nature & the outdoors, tourism, cafes, bars, restaurants, religious worship
- Domesticity
- Simple pleasures
- Solitude

Attitudes:

- Kindness, consideration
- Public spirit, less individualism
- ±Less materialism / consumerism
- Resilience
- Self-reliance
- Flexibility
- Acceptance of mortality
- Acceptance of uncertainty
- Humility, less complacency
- ?±Short-termism, living in the present
- ?Less concern about own appearance
- ?Less attention to celebrities
- ?Solidarity with other countries

Miscellaneous

±Deaths:

- ?±Beneficial if the world is overpopulated; or if humans, or those who died, are generally harmful or their lives not worthwhile; according to some (controversial) ethical theories
- ±Redistribution of wealth to younger generations

?End of physical cash due to [infection risk](#):

- Traceability reduces crime & tax evasion
- Simplifies government emergency handouts

?Less crime, as criminals [reassess their lives](#)

?Better bank treatment of borrowers as continuation of special terms under lockdown

?Better rights for renters after evictions suspended during lockdown (e.g. in UK)

?More fact-checking on social media

Other resources

[This list](#) includes harmful consequences of coronavirus (as well as various of the above benefits).

In-depth discussion of some points is in a [Politico article](#) and [FT series](#) (paywall).

*Toby Ord's new book [*The Precipice*](#) estimates that the human race will only last another 600 years or so before it is wiped out, or permanently crippled, by a pandemic (probably a [bioweapon](#)) or other [existential risk](#).

If coronavirus makes the world prepare slightly better for such disasters, thereby reducing the risk by say 1%, it would extend the human race by an expected value of $600 \text{ years} \times 1\% = 6 \text{ years}$. The world population is [forecast](#) to reach about 11 billion, so this would save $6 \text{ years} \times 11 \text{ billion} = \mathbf{66 \text{ billion years}}$ of life.

If coronavirus kills 10 million people worldwide, each losing [10 years](#) of life on average, **100 million years** of life will be lost. This is a minute fraction of the benefit from improved disaster preparedness.

Failures in technology forecasting? A reply to Ord and Yudkowsky

In [The Precipice](#), Toby Ord writes:

we need to remember how quickly new technologies can be upon us, and to be wary of assertions that they are either impossible or so distant in time that we have no cause for concern. Confident denouncements by eminent scientists should certainly give us reason to be sceptical of a technology, but not to bet our lives against it - their track record just isn't good enough for that.

I strongly agree with those claims, think they're very important in relation to [estimating existential risk](#),^[1] and appreciate the nuanced way in which they're stated. (There's also a lot more nuance around this passage which I haven't quoted.) I also largely agree with similar claims made in Eliezer Yudkowsky's earlier essay [There's No Fire Alarm for Artificial General Intelligence](#).

But both Ord and Yudkowsky provide the same set of three specific historical cases as evidence of the poor track record of such "confident denouncements". And I think those cases provide less clear evidence than those authors seem to suggest. So in this post, I'll:

- Quote Ord and/or Yudkowsky's descriptions of those three cases, as well as one case mentioned by Yudkowsky but not Ord
- Highlight ways in which those cases may be murkier than Ord and Yudkowsky suggest
- Discuss how much we could conclude about technology forecasting *in general* from such a small and likely unrepresentative sample of cases, even if those cases weren't murky

I should note that I don't think that these historical cases are *necessary* to support claims like those Ord and Yudkowsky make. And I suspect there *might* be better evidence for those claims out there. But those cases were the main evidence Ord provided, and among the main evidence Yudkowsky provided. So those cases are being used as key planks supporting beliefs that are important to many EAs and longtermists. Thus, it seems healthy to prod at each suspicious plank on its own terms, and [update incrementally](#).

Case: Rutherford and atomic energy

Ord writes:

One night in 1933, the world's pre-eminent expert on atomic science, Ernest Rutherford, declared the idea of harnessing atomic energy to be 'moonshine'. And the very next morning Leo Szilard discovered the idea of the chain reaction.

[Yudkowsky](#) also uses the same case to support similar claims to Ord's.

However, in a footnote, Ord adds:

[Rutherford's] prediction was in fact partly self-defeating, as its confident pessimism grated on Szilard, inspiring him to search for a way to achieve what was said to be impossible.

To me, the phrase “the very next morning” in the main text made this sound like an especially clear example of just how astoundingly off the mark an expert’s prediction could be. But it turns out that the technology didn’t just *happen* to be *just about* to be discovered in any case. Instead, there was a direct connection between the prediction and its undoing. In my view, that makes the prediction less “surprisingly” incorrect.^[2]

Additionally, in the same footnote, Ord suggests that Szilard’s discovery may not even have been “the very next day”:

There is some debate over the exact timing of Szilard’s discovery and exactly how much of the puzzle he had solved

Finally, the same footnote states:

There is a fascinating possibility that [Rutherford] was not wrong, but deliberately obscuring what he saw as a potential weapon of mass destruction (Jenkins, 2011). But the point would still stand that confident public assertions of the leading authorities were not to be trusted.

This is a very interesting point, and I appreciate Ord acknowledging it. But I don’t quite agree with his last sentence. I’d instead say:

This possibility may weaken the evidence this case provides for the claim that we should *often* have limited trust in confident public assertions of the leading authorities. But it may not weaken the evidence this case provides for that claim in situations where it’s plausible that those assertions might be based less on genuine beliefs and more on a desire to e.g. mitigate [attention hazards](#).

To be clear, I do think the Rutherford case provides *some* evidence for Ord and Yudkowsky’s claims. But I think the evidence is weaker than those authors suggested (especially if we focus only on Ord’s main text, but even when also considering his footnote).

Case: Fermi and chain reactions

Ord writes:

In 1939, Enrico Fermi told Szilard the chain reaction was but a ‘remote possibility’, and four years later Fermi was personally overseeing the world’s first nuclear reaction.

Both [Yudkowsky](#) and [Stuart Russell](#) also use the same case to support similar claims to Ord’s.

However, in a footnote, Ord writes:

Fermi was asked to clarify the ‘remote possibility’ and ventured ‘ten percent’. Isidor Rabi, who was also present, replied, ‘Ten percent is not a remote possibility if it means that we may die of it. If I have pneumonia and the doctor tells me that

'there is a remote possibility that I might die, and it's ten percent, I get excited about it'

I think that that footnote itself contains an excellent lesson (and an excellent quote) regarding failures of *communication*, and regarding the potential value of quantifying estimates ([see also](#)). Relatedly, this case seems to support the claim that we should be wary of trusting *qualitatively stated* technology forecasts (even from experts).

But the footnote also suggests to me that this *may* not have been a failure of *forecasting* at all, or only a minor one. Hearing that Fermi thought that something that ended up happening was only a "remote possibility" seems to suggest he was wildly off the mark. But if he actually thought the chance was 10%, perhaps he was "right" in some sense - e.g., perhaps he was [well-calibrated](#) - and this just happened to be one of the 1 in 10 times that a 10% likely outcome occurs.

To know whether that's the case, we'd have to see a larger range of Fermi's predictions, and ensure we're sampling in an unbiased way, rather than being drawn especially to apparent forecasting failures. I'll return to these points later.

Case: Nuclear engineering more broadly

As evidence for claims similar to Ord's, [Yudkowsky](#) also uses the development of nuclear engineering more broadly (i.e., not just the above-mentioned statements by Rutherford and Fermi). For example, Yudkowsky writes:

And of course if you're not the Wright Brothers or Enrico Fermi, you will be even more surprised. Most of the world learned that atomic weapons were now a thing when they woke up to the headlines about Hiroshima.

And:

Fermi wasn't still thinking that net nuclear energy was impossible or decades away by the time he got to 3 months before he built the first pile, because at that point Fermi was looped in on everything and saw how to do it. But anyone not looped in probably still felt like it was fifty years away while the actual pile was fizzing away in a squash court at the University of Chicago

And, in relation to the development of AGI:

I do put a significant chunk of probability mass on "There's not much sign visible outside a Manhattan Project until Hiroshima," because that scenario is simple.

And:

I do predict in a very general sense that there will be no fire alarm [roughly, a clear signal of AGI being soon] that is not an actual running AGI--no unmistakable sign before then that everyone knows and agrees on, that lets people act without feeling nervous about whether they're worrying too early. That's just not how the history of technology has usually played out in much simpler cases like flight and nuclear engineering, let alone a case like this one where all the signs and models are disputed.

I largely agree with, or at least find plausible, Yudkowsky's claims that there'll be no "fire alarm" for AGI, and consider that a quite important insight. And I think the case of the development of nuclear weapons *probably* lends support to *some version* of those claims. But, for two reasons, I think Yudkowsky may paint an overly simple and confident picture of how this case supports his claims. (Note that I'm not an expert in this period of history, and most of what follows is based on some quick Googling and skimming.)

Firstly, I believe the development of nuclear weapons was highly militarised and secretive from early on, and to a much greater extent than AI development is. My impression is that the general consensus is that non-military labs such as DeepMind and OpenAI truly are leading the field, and that, if anything, AI development is worryingly *open*, rather than highly secretive (see e.g. [here](#)). So it seems there are relevant disanalogies between the case of nuclear weapons development and AI development (or indeed, most technological development), and that we should be substantially uncertain when trying to infer from the former case to the latter.

Secondly, I believe the group of people who *did* know about nuclear weapons before the bombing of Hiroshima, or who believed such weapons may be developed soon, was (somewhat) larger than one might think from reading Yudkowsky's essay. In particular, the [British](#), [Germans](#), and [Soviets](#) each had their own nuclear weapons programs, and [Soviet leaders knew of both the German and US efforts](#). And I don't know of any clear evidence either way regarding whether scientists, policymakers, and members of the public who *didn't* know of these programs would've assumed nuclear weapons were impossible or many decades away.

That said, it *is* true that:

- the group of people who knew of these nuclear weapons programs before Hiroshima was *very small*
- [even Truman](#) wasn't told the US was developing nuclear weapons during his short time as Vice President (he was only told when he was sworn in as President)
- even the vast majority of people *working on the Manhattan Project* [didn't know its true purpose](#)

So I do think this case provides evidence that technological developments can take a *lot* of people outside of various "inner circles" by surprise, at least in cases of highly secretive developments during wartime.

Case: The Wrights and flight

Ord writes:

The staggering list of eminent scientists who thought heavier-than-air flight to be impossible or else decades away is so well rehearsed as to be cliché.

I haven't looked into that claim, and Ord gives no examples or sources. [Yudkowsky](#) references [this list of](#) "famous people and scientists proclaiming that heavier-than-air flight was impossible". That too gives no sources. As a spot check, I googled the first and last of the quotes on that list. The first [appears to be substantiated](#). For [the last](#), the first page of results seemed to all just be other pages also using the quote in "inspirational" ways without a giving source. Ultimately, I wouldn't be surprised if

there's indeed a staggering list of such proclamations, but also wouldn't be surprised if a large portion of them are apocryphal (even if "well rehearsed").

Ord follows the above sentence with:

But fewer know that even Wilbur Wright himself predicted [heavier-than-air flight] was at least fifty years away - just two years before he invented it.

The same claim is also made by Yudkowsky.

But in a footnote, Ord writes:

Wilbur Wright explained to the Aero-club de France in 1908: 'Scarcely ten years ago, all hope of flying had almost been abandoned; even the most convinced had become doubtful, and I confess that in 1901 I said to my brother Orville that men would not fly for 50 years. Two years later, we ourselves were making flight.'

Thus, it seems our evidence here is a retrospective account, from the inventor himself, of a statement he once made. One possible explanation of Wright's 1908 comments is that a genuine, failed attempt at forecasting occurred in 1901. Here are three alternative possible explanations:

1. Wright just made this story up after the fact, because the story makes his achievement sound all the more remarkable and unexpected.
2. In 1908, Wright did remember making this statement, but this memory resulted from gradual distortions or embellishments of memory over time.
3. The story is true, but it's a story of one moment in which Wright *said* men would not fly for 50 years, as something like an expression of frustration or hyperbole; it's not a genuine statement of *belief* or a genuine attempt at *prediction*.

Furthermore, even if that *was* a genuine prediction Wright made at the time, it seems it *was* a prediction made briefly, once, during many years of working on a topic, and which wasn't communicated publicly. Thus, even if it was a genuine prediction, it may have little bearing on the trustworthiness *in general* of publicly made forecasts about technological developments.

Sample size and representativeness

Let's imagine that all of my above points turn out to be unfounded or unimportant, and that the above cases turn out to all be clear-cut examples of failed technology forecasts by relevant experts. What, then, could we conclude from that?

Most importantly, that'd provide very strong evidence that experts saying a technological development is impossible or far away doesn't *guarantee* that that's the case. And it would provide *some* evidence that such forecasts may *often* be mistaken. In places, this is all Ord and Yudkowsky are claiming, and it might be sufficient to support some of their broader conclusions (e.g., that it makes sense to work on AI safety *now*). And in any case, those broader conclusions can also be supported by other arguments.

But it's worth considering that these are *just four* cases, out of the *entire* history of predictions made about technological developments. That's a *very small sample*.

That said, as noted in [this post](#) and [this comment](#), we can often learn a lot about what's typical of some population (e.g., all expert technology forecasts) using just a small sample from that population. What's perhaps more is whether the sample is *representative* of the population. So it's worth thinking about how one's sample was drawn from the population. I'd guess that the sampling process for these historical cases wasn't random, but instead looked more like one of the following scenarios:

1. Ord and Yudkowsky had particular points to make, and went looking for past forecasts that supported those points.
2. When they came to make their points, they already happened to know of those forecasts due to prior searches motivated by similar goals, either by themselves or by others in their communities.
 - E.g., I'd guess that Ord was influenced by Yudkowsky's piece.
3. They already happened to know of many past technology forecasts, and mentioned the subset that suited their points.

If so, then this was a biased rather than representative sample.^[3] Thus, if so, we should be very careful in drawing conclusions from it about what is *standard*, rather than about the plausibility of such failures occurring *on occasion*.^[4]

It's also interesting to note that Ord, Yudkowsky, and [Russell](#) all wished to make similar points, and all drew from the same set of four cases. I would guess that this is purely because those authors were influenced by each other (or some other shared source). But it *may* also be because those cases are among the cases that most clearly support their points. And, above, I argued that each case provides less clear evidence for their points than one might think. So it seems *possible* that the repeated reaching for these murky examples is actually weak evidence that it's *hard* to find clear-cut evidence of egregious technology forecasting failures by relevant experts. (But it'd be better to swap my speculation for an active search for such evidence, which I haven't taken the time to do.)

Conclusion

Both Ord and Yudkowsky's discussions of technology forecasting are much more nuanced than saying long-range forecasting is impossible or that we should pay *no attention at all* to experts' technology forecasts. And they both cover more arguments and evidence than just the handful of cases discussed here. And as I said earlier, I largely agree with their claims, and overall see them as very important.

But both authors do prominently feature this small set of cases, and, in my opinion, imply these cases support their claims more clearly than they do. And that seems worth knowing, even if the same or similar claims could be supported using other evidence. (If you know of other relevant evidence, please mention it in the comments!)

Overall, I find myself mostly just very uncertain of the trustworthiness of experts' forecasts about technological developments, as well as about how trustworthy forecasts *could* be given better conditions (e.g., better incentives, calibration training). And I don't think we should update much based on the cases described by Ord and Yudkowsky, unless our starting position was "Experts are almost certainly right"

(which, to be fair, may indeed be many people's implicit starting position, and is at times the key notion Ord and Yudkowsky are very valuably countering).

Note that this post is *far* from a comprehensive discussion on the efficacy, pros, cons, and best practices for long-range or technology-focused forecasting. For something closer to that, see [Muehlhauser](#)^[5] who writes, relevantly:

Most arguments I've seen about the feasibility of long-range forecasting are purely anecdotal. If arguing that long-range forecasting is feasible, the author lists a few example historical forecasts that look prescient in hindsight. But if arguing that long-range forecasting is difficult or impossible, the author lists a few examples of historical forecasts that failed badly. How can we do better?

I also discuss similar topics, and link to other sources, in my post introducing [a database of existential risk estimates](#).

This is one of a series of posts I plan to write that summarise, comment on, or take inspiration from parts of The Precipice. You can find a list of all such posts [here](#).

This post is related to my work with [Convergence Analysis](#), but the views I expressed in it are my own. My thanks to [David Kristoffersson](#) and [Justin Shovelain](#) for useful comments on an earlier draft.

1. A related topic for which these claims are relevant is the likely timing and discontinuity of AI developments. This post will not directly focus on that topic. Some sources relevant to that topic are listed [here](#). ↩
2. This may not reduce the strength of the evidence this case provides for *certain* claims. One such claim would be that we should put little trust in experts' forecasts of AGI being definitely a long way off, and this is *specifically because* such forecasts *may themselves* annoy other researchers and spur them to develop AGI faster. But Ord and Yudkowsky didn't seem to be explicitly making claims like that. ↩
3. I don't mean "biased" as a loaded term, and I'm not applying that term to *Ord* or *Yudkowsky*, just to their samples of historical cases. ↩
4. Basically, I'd guess that the evidence we have is "people who were looking for examples of experts making technology forecasting mistakes were able to find 4 cases as clear-cut as the cases Yudkowsky gives". This evidence seems almost as likely conditional on "experts' technology forecasts are right 99% of the time" as conditional on "experts' technological forecasts are right 1% of the time" (for two examples of possible hypotheses we might hold). Thus, I don't see the evidence as providing much Bayesian evidence about which of those hypotheses is more likely. I wouldn't say the same if our evidence was instead "the first four of the four cases we *randomly sampled* were as clear-cut as the cases Yudkowsky gives". ↩
5. Stuart Armstrong's recent post [Assessing Kurzweil predictions about 2019: the results](#) is also somewhat relevant. ↩

Book Review: Narconomics

I have often heard that the war on drugs is a failure, but the reason why varies.

Sometimes, the war on drugs is a failure like the war in Iraq is a failure: because it is very hard to nation-build, to create a society that does not produce or consume drugs.

Sometimes, the war on drugs is a failure like the war in Vietnam was a failure: because the opponents, drug cartels, don't fight like we do.

Sometimes, the war on drugs is a failure like WWI was a failure: we can win the battles but we don't know how to write a lasting peace treaty.

Tom Wainwright's Narconomics ropes in elements of the above, but offers another overarching story. The war on cocaine is a failure like the war on alchemy would be a failure.

If you knew how to perform alchemy, to turn dirt from your backyard into gold with just an incantation, would you stop just because the government forbade using the incantation? Cocaine isn't gold, but only because it's twice as valuable as gold in the US. \$385 worth of coca plant leaves in Columbia are worth \$122,000 once turned into a kilogram of cocaine and shipped to the US. That's a 30,000% increase in value. Not exactly alchemy, but close.

To put it another way, if drug cartels lost a half-million dollar plane for each shipment of cocaine to the US, it would increase the final price of cocaine by *one percent*.

The economics of cocaine, like that of alchemy, are very lucrative. As long as that stands, drug cartels will do anything to keep growing and smuggling cocaine. The war on cocaine is bound to fail for as long as cocaine is in high demand.

Wainwright's recommendation is simple. To drive profits from cocaine down and criminal entities out, we should legalize cocaine.

Burning coca leaves won't win the war

Here is an Economics 101 problem: In the year 20XX, a mysterious almond blight kills half of all almond plants worldwide. What happens to the price of almonds?

Easy. Constant demand + restricted supply → increased price.

Now here's a variation on the same problem. In the year **datetime.now().year**, the Columbian government, with encouragement from the US, lays waste to about half the coca crops in the country. What happens to the price of cocaine?

Constant demand + restricted supply → ``_(ツ)_/`` nothing much, apparently.

The reason is that coca farmers operate in a monopsony: a single-buyer market. In the US, Walmart is the archetypical monopsony. Because Walmart can sell to almost every house in the US, it has incredible power over its suppliers. If you manufacture whistling lawnmowers, and you want to sell to people who buy whistling lawnmowers, you're going to have to sell to Walmart. You have no bargaining power; Walmart

names its price. From that point on, Walmart's price (and profit) is as constant as the speed of light. If someday the price of manufacturing lawnmowers rises, the rise will eat into your profit, not Walmart's.

Similarly, coca farmers have one customer for their crop: the local drug cartel. The cartel names its price. They nod because (a) they are not going to find another buyer for their crop and (b) they don't want to get shot. The government's plan of spraying, uprooting, and burning coca plants only serves to make coca farmers poorer; the drug cartel still buys coca plants at the same price it always did.

But let's grant that this strategy works. The Colombian government's crop destruction *triples* the price of coca plants. Will the price of cocaine in the US increase then? It's unlikely.

Right now, dried coca leaves are worth about ~\$1 per kilogram. After their price triples, they would be worth ~\$3 per kilogram. As a result, the price of a kilogram of cocaine sold in the US would increase from \$122,000 to... \$122,770 (+0.6%). This is, to put it mildly, negligible.

Policing drug routes won't win the war

Cocaine that is worth \$2,200 in Columbia is worth \$14,500 by the time it is imported into the US. That is more than a 500% increase in value for just changing the (x,y) coordinates of cocaine. Drug cartels will go to great lengths to get that 500% increase in value. As the US border becomes more heavily policed, the value of each remaining crossing point into the US increases wildly. Drug cartels pull out all stops to control a crossing point because a cartel without a crossing point won't last long. Border cities like Tijuana, Reynosa, and (the one Wainwright focuses on) Ciudad Juárez become hotspots of violence as cartels battle to control their crossing points.

The pervasive violence Wainwright describes in Juárez is breathtaking. To take just a few examples:

- Wainwright's driver, Miguel, leaves a lot of room between cars when he stops at traffic lights, because traffic lights are known for being assassination hotspots (assassins can drive up, shoot, and speed away easily).
- A local cartel, Sinaloa, hung a warning poster on a public memorial to fallen police officers. The poster had names of seventeen police officers. Soon after, it started killing the officers named on the list.
- The city and federal police were bribed by competing gangs, leading to an assassination where a city cop was shot by a federal cop.
- A cartel murdered a reporter and broke into her Twitter account to tweet a picture of her dead body as a warning to other reporters.
- Cartels time murders to happen in the early evening, so that their exploits lead the 6:00pm news.

But again, let's grant that this strategy works. The US manages to completely seal all border crossing points to cocaine smugglers. Will the supply of cocaine in the US decrease then? It's unlikely.

Just like the semiconductor industry follows Moore's law, drug cartels follow the "when one door closes, another opens" law. In the 80s, cocaine would come in on speedboats from the Caribbean to southern Florida. When enforcement got stricter there, cartels started using land crossings at the US-Mexico border. If we shut down those crossings,

they'll likely fly cocaine into the US using unassuming drug mules, as they already do for Europe. The US market is simply too lucrative to be ignored.

Imprisoning offenders won't win the war

Every country involved in the war on cocaine imprisons people involved with the cocaine supply chain. Some go to absurd lengths, like the US, where 1 in 35 people is involved in the prison system, either in jail or prison, or on probation or parole. Others are just absurd, like Mexico, where prisoners in one Acapulco prison smuggled in nineteen prostitutes and two peacocks. Either way, prisons are not only ineffective, they are counterproductive in the war on cocaine.

One of the biggest challenges drug cartels face is recruiting members to join them. The drug business is violent, messy, and unstable. Not many people would sign up for that life if they had a different option. Drug cartels, out of necessity, have to recruit people who would find it hard to get a job otherwise, and who don't find a life of crime objectionable. In other words, prisons are fertile recruiting grounds for drug cartels.

Prison gangs go to surprising lengths to create a welcoming environment for new recruits. Not only do they protect their members from competing gangs and promise careers outside prison, they also provide community and structure to their members. A Mexican drug cartel, La Familia Michoacana, makes new members read a Christian self-help book. A Californian prison gang, La Nuestra Familia, lets its members elect their captains; this helps new recruits select leaders that won't abuse them. The Aryan Brotherhood used to let its members vote on whether assassinations should be carried out.

On the other hand, governments do little to nothing to rehabilitate prisoners and prepare them for a life outside prison. Indeed, some seem to actively work against this goal. The US locks up people for minor offenses that could be better handled without incarceration. The US is also guilty of spending much more money on prison than prevention, like when New Hampshire cut funding for rehabilitation programs, but allowed Keene, a city of 23,000 people, to buy a \$286,000 armored car for patrolling the annual Pumpkin Festival. Other countries don't fare much better. For example, in the Dominican Republic, some prisons don't provide food, forcing prisoners to ask friends and family to deliver food, which creates opportunities for smuggling drugs or weapons into prison.

All this allows prisons to act as a finishing school that transforms minor offenders into hardened criminals.

The argument for legalization

The simplest argument for legalization is, it has to be better than what we are currently doing.

The current war is doomed to fail because:

1. The end price of cocaine is so high that most supply-side interventions will not change it significantly.
2. Even if an intervention does change the end price of cocaine, the demand is inelastic and won't drop much.

3. Because demand is inelastic, increases in the price of cocaine just make the market more valuable and alluring.

On the other hand, unleashing the forces of competition into the cocaine market will drive prices down. If prices go down, drug cartels will exit the business. Drug cartels might be bloodthirsty, but they've got nothing on Delaware C-Corporations hellbent on maximizing shareholder value.

The example of marijuana is instructive. (The book was written when Colorado and Washington were the only states to legalize marijuana, and focuses on them.) Colorado's legal marijuana growers have put a serious dent in the drug cartels' marijuana business. Cartels have to offer lower-potency marijuana at one-third of the legal price to stay competitive. Their revenues were forecasted to drop by 75%. Things are getting so desperate that some cartels have started to use their closely-guarded drug-trafficking tunnels to smuggle migrants into the US. Doing so greatly increases the risk that their tunnels will be discovered and closed, but cartels are desperate for money.

Summary

Imagine that you and Bob are deciding what to get for dinner.

You: I ate a really heavy lunch today, so I want something light for dinner. Are you interested in a salad?

Bob: Let's get pizza slices instead. They have the most calories per dollar of any takeout food.

You would rightfully think "what the hell?" Bob ignored what you said, and is using a criteria (calories per dollar) that you don't care about. I felt a little bit like that when reading this book. Wainwright, like Bob, has excellent economic arguments for why the war on cocaine is wasteful, and why cocaine legalization would reduce crime and increase control of cocaine consumption. However, most people who want cocaine to be illegal don't care about the economics of the cocaine supply chain. They just don't want their kids and neighbors to be addicts.

Or to put it in economic terms, most people believe that making cocaine legal will have huge negative externalities. There will be costs from people becoming less healthy and less productive because they are consuming more cocaine. There will be costs from people becoming so addicted to cocaine that they commit crimes to get their next hit. There will be costs from living in a world where you have to constantly resist, at least at a low level, the temptation to take cocaine. For most people, these negative externalities are larger and more salient than any benefit of legalizing cocaine.

Marijuana legalization victories, where they've happened, have happened because proponents talked about how marijuana is harmless, not how marijuana criminalization is economically senseless. For example, take a look at the [FAQ page for the Marijuana Policy Project](#), a leading marijuana legalization organization in the US. The first question they answer is "Is marijuana addictive?" The other questions also address similar concerns. There is only one question on that page that talks about economics, and even that focuses more on how much money could be gained by taxing marijuana.

This book is weak because it barely touches those concerns. The book deals with every part of the cocaine supply chain, except the final part, where cocaine users use cocaine. Unfortunately, that's the part of the cocaine supply chain that worries Americans most. Wainwright's economic argument for drug legalization is compelling, but incomplete.

In Search of Slack

Original article: [Studies on Slack](#)

1

Proposition: If you have enough free resources you can evolve irreducibly complex features.

Imagine a bunch of yeast cells added to a barrel of malt when beer is being brewed. They have an ocean of free resources at their disposal. Surely, the evolutionary pressure will be low and the yeast cells would be able to escape the local minimum and evolve an irreducibly complex eye.

Where this idea fails is not taking the nature of exponential growth into account. Exponential means being fast. It means, in fact, being faster than anything you can visualize. The barrel would be fully fermented in just couple of days.

In other words: No matter how big a [whale fall](#) is, it will be exploited fast. The body of the whale will be eaten in months and the last bacterial remains of the ecosystem will survive for maybe 100 years. Nowhere near the time needed for a complex evolutionary change to happen by chance.

It's worth looking at [E. coli long-term evolution experiment](#) here. Since 1988 E. coli is cultivated in a uniform substrate and observed to see how it evolves. From our perspective it's interesting that each day 1% of population is transferred to a new flask with the fresh substrate. One can think of it as of a new whale fall every day. Unfortunately, I am not aware of a control experiment where substrate would be added gradually, which would allow us to see whether environment with slack is more conducive to evolution of new features than environment without slack.

However, in one of the populations in the experiment a speciation event was observed. There's one strain of E. coli that has advantage during growth on the substrate (slack) and another strain that has advantage during stationary phase, when the substrate runs out (moloch).

That hints that supposed slack may not be slack at all, just a different environment exploited in competitive manner by different subset of species (ecological opportunists).

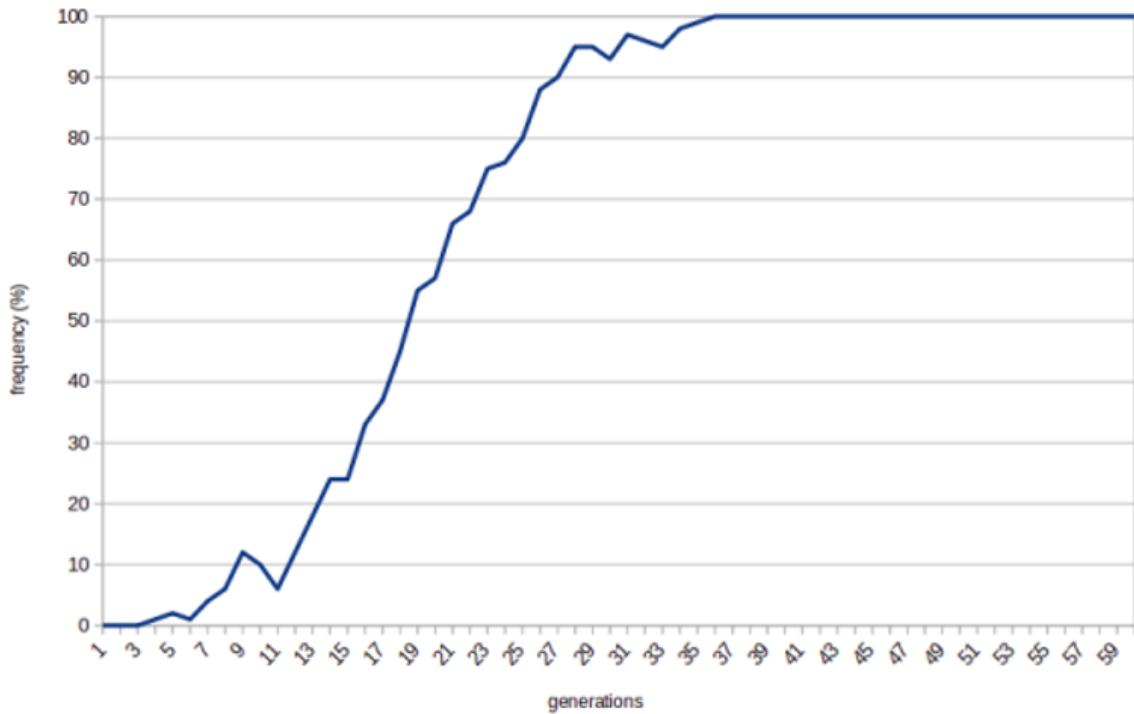
2

Proposition: Isolation of small subpopulations can produce slack.

Let's start with some basics of population genetics so that we have tools to understand what's going on.

Let's say no human being is able to roll their tongue. What happens when a mutation arises that gives one that ability?

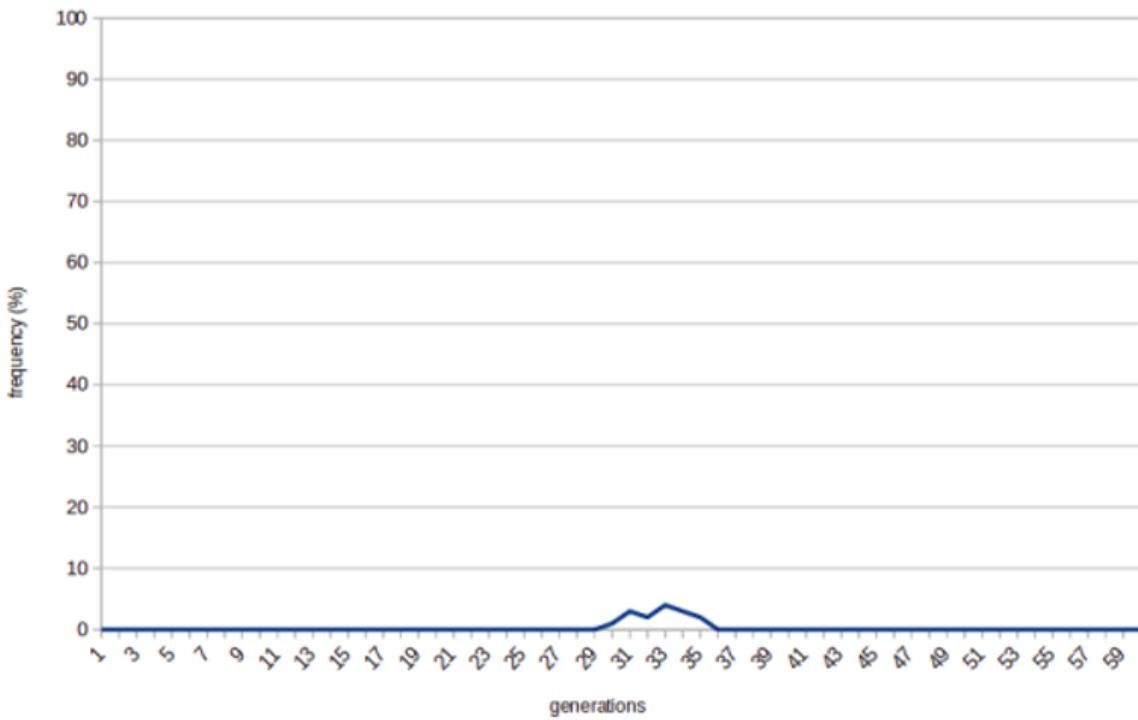
Well, if rolling your tongue makes you more likely to survive and have more children then the trait will be actively selected for. The frequency of the tongue-rolling allele in the population will increase rapidly:



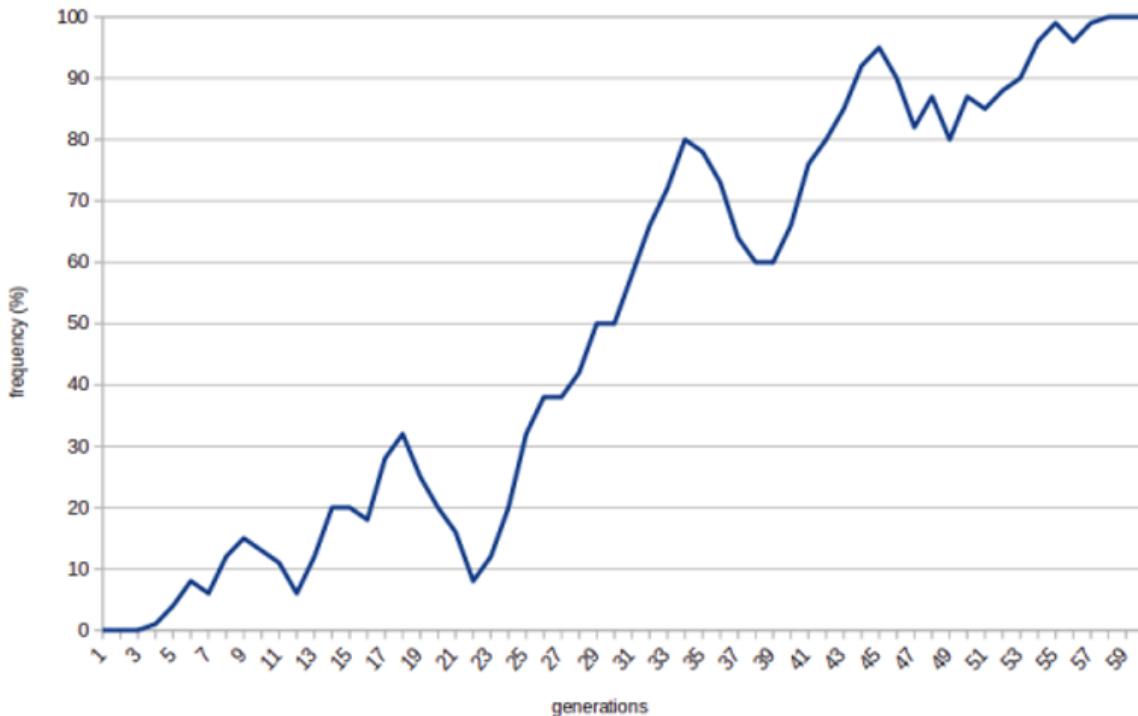
Note that when the frequency reaches 100% there's no way back. There are no more non-tongue-rollers that could reproduce and take over the population. We say that the allele becomes fixed.

But what happens if the allele has no effect on your fitness? What if tongue-rolling makes you neither more likely to survive and reproduce, nor less likely?

In that case the frequency of the allele is a random walk. If increases and decreases solely by chance. In most cases it means it will be wiped out quickly:

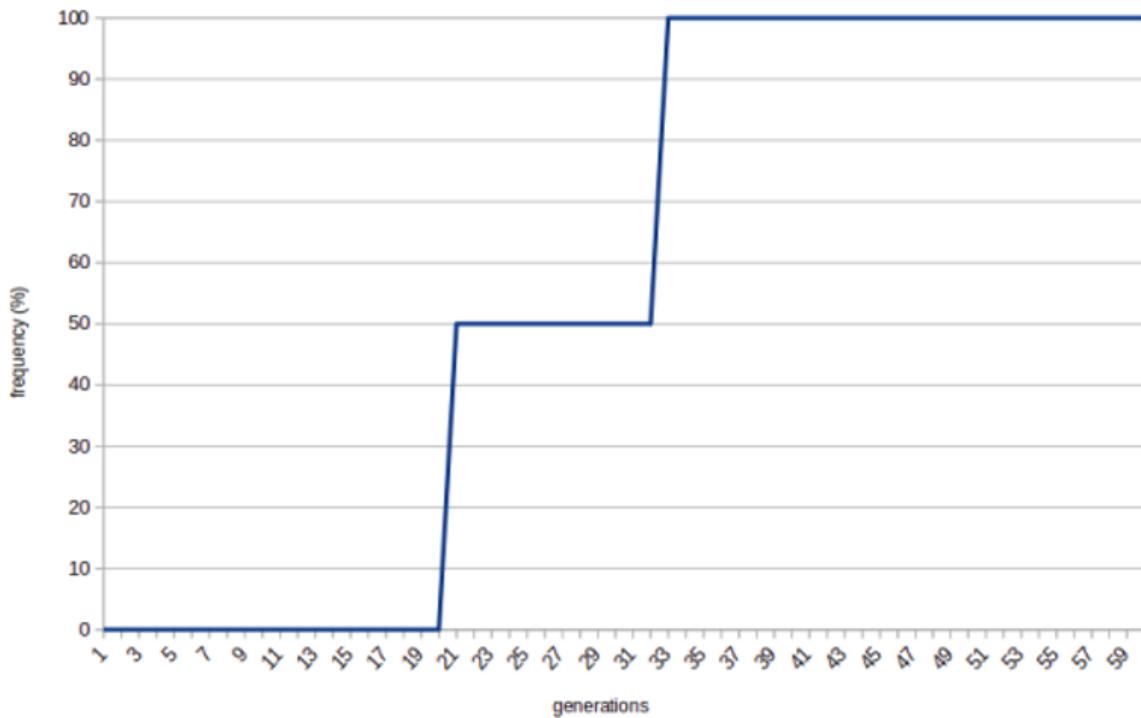


Sometimes, though, the allele gets lucky and becomes fixed:

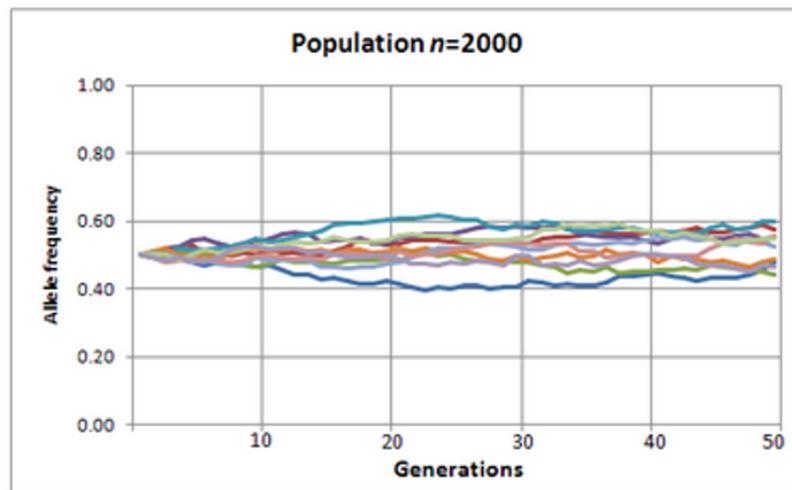
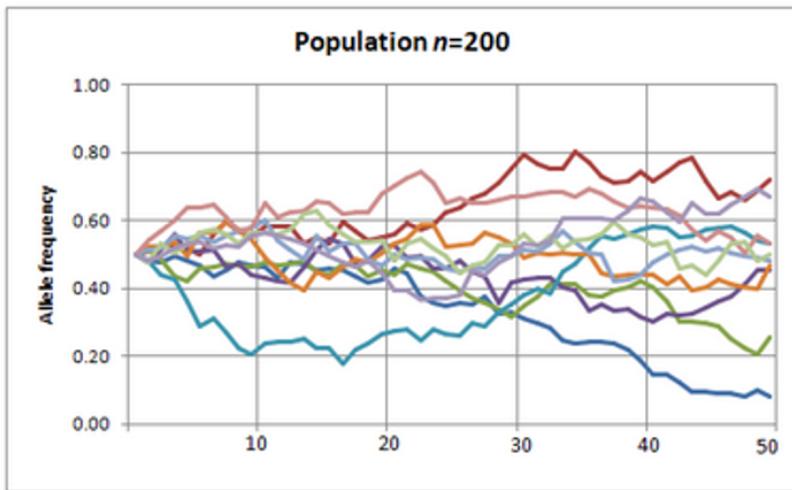
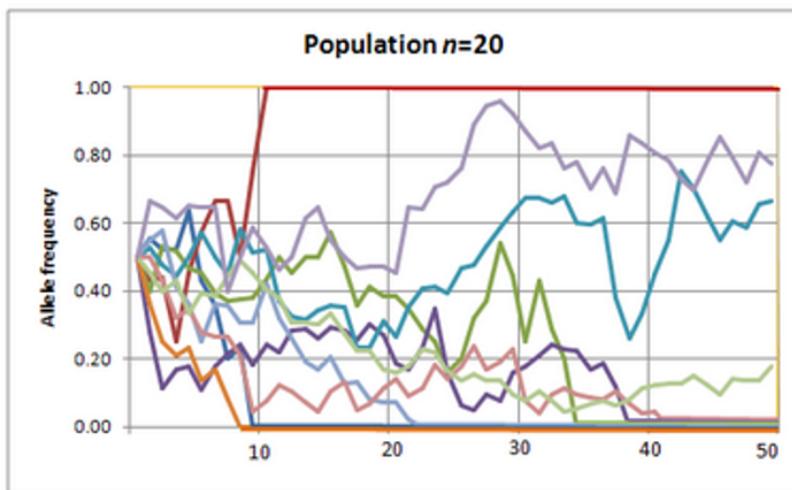


We call this process of random fluctuation of allele frequency "genetic drift".

The things turn interesting when the population size is small. Imagine population of two. Now it's much easier for the new allele to become fixed. Once it comes to being via mutation it already has 50% frequency in the pool and all it takes to become fixed is reproducing once.



The smaller the population, the more alleles become fixed by chance alone. (The picture stolen from [Wikipedia](#), CC BY-SA 3.0)



To put it in different words, there are two competing forces in play here. There's selection (determinism, Moloch). And then there's drift (randomness, slack). And there are two factors that determine which one is going to be stronger. First, the more effect the allele has on the

fitness the more is its evolution driven by selection. Second, the smaller the population, the more it is driven by drift.

In very small populations even alleles that are deleterious (harmful to the survival of the individual) can become fixed by chance.

If there was a single allele that coded for the half of the irreducibly complex eye it could become fixed even though having a half of eye is, strictly speaking, worse than not having an eye at all.

What we gain here is time. We gain slack to wait for a new mutation to complete the eye. There are no more individuals with no eyes at all that would outcompete the half-eyed individuals. Now everybody has a half of an eye (even if the population size grows!) and we are waiting for something to happen. We are standing on a hillock in the adaptive landscape. It could be that the next mutation would produce no-eyed individuals again who would outcompete the half-eyed individuals. In that case we are back to where we started. But it may also happen that the next mutation would provide the second half of the eye and the individuals with the sense of sight would spread through the population.

Now, assuming that our small population originated by splitting from a large ancestral population, for example, some animals migrating to a new island, now that they have evolved a superior trait they can migrate back to their ancestral homeland, win the evolutionary race and become the dominant species. (See [peripatric speciation](#).)

There's not much slack to be seen here. We can call small size of a population "slack" even if it is under intense evolutionary pressure, but that way the term loses its intuitive appeal.

But let's look further.

Interesting alternative was proposed by the team of Andreas Wagner, here in Zurich. They start with the fact that neutral mutations (neither beneficial nor harmful) are subject to the random drift and thus a population is likely to have a lot of neutral variation at any given point of time. In other words, while the population has only a small variation of phenotypes (how the animals look like) it has large variation of genotypes (how the genomes of the animals look like). Then they show that different genotypes within the same phenotype can have vastly different "neighbourhoods". One animal may, to put it simply, undergo a single mutation and evolve an eye. Another may undergo a different mutation and evolve an ear.

Think of it as of a map. Each colour patch is a different phenotype. Locations within a patch are different genotypes. Individuals in the population, thanks to neutral mutations (which are by definition free of the selection pressure) can explore the entire area of their particular phenotype. Once the environment changes there may be some individuals just across the border from the phenotype that would do better in the new conditions. Such individuals can undergo a single non-neutral mutation and result in a new phenotype which will then spread in the population via selection.



This may be closer to the intuitive notion of slack. Neutral mutations have slack because they don't affect the phenotype and are thus not acted upon by the selection.

3

Proposition: Two-layer evolutionary systems can produce their own slack. The outer system can introduce rules that give inner system some slack.

Let's look at some examples.

Multicellular organisms: Each cell is genetically identical to each other cell. There's no variance for the evolution to work with. Any mutations (cancer) are aggressively weeded out. There's no evolutionary slack in the sense of "freedom to evolve in random direction".

Eukaryotic cell: Eukaryotic cell has two genomes so it's harder to tell whether the outer system (nuclear DNA) gives the inner system (mitochondrial DNA) any slack. But we can make an educated guess. First, let's stress the point that mutations in mitochondrial DNA can

be as dangerous as mutations in nuclear DNA. There is such a thing as mitochondrial cancer. It ends up with a cell packed up wall to wall with mitochondria. That implies that it doesn't pay off for the cell to give mitochondria too much slack. Second and more important: Look at the crazy measures the cells use to prevent mixing mitochondria from both parents when undergoing sexual reproduction. The fact that sperm cell is stripped of all of its mitochondria before the DNA is allowed to enter the egg cell hints that intra-cell mitochondrial evolutionary conflict is detrimental to the cell as a whole. To sum it up, it looks very much like eukaryotic cells give mitochondria as little slack as possible.

Social insects: In fully social insects such as bees, the individuals have different genomes, but they use a weird trick to ensure social cohesion. To make it short, their reproductive mechanism is such that a sister of a bee is more genetically similar to herself than her own offspring. That's an incentive not to reproduce herself but rather help the queen with reproducing. In the end the system doesn't look too much different from the system used by multicellular organisms. (It may be interesting to look at termites though, which lack the incentive system described above.)

All of the above are systems seem to be the cases of 100% coordination with no slack allowed.

4

Proposition: Two-layer evolutionary systems with migration can produce slack.

The difference to the previous scenario is that the lower level individuals (say, mitochondria) can migrate between higher level individuals (eukaryotic cells). While mitochondria cannot in fact migrate, there's a huge class of organisms that can. Namely, parasites. (In fact, it's a rule of thumb: Anything that cannot migrate tends to be a mutualist. Anything that can migrate has a tendency towards parasitism.)

These include typical parasites, such as pinworm, but also, on a different level, memes. The common characteristic is that the environment of a parasite is not a single individual of the host species, but rather the entire host species. In theory, an opportunity to play non-zero-sum games emerges: If a parasite helps its host to survive and reproduce it grows the host pool size and may get more populous itself. However, assuming free migration between hosts, there's nothing preventing the old bad strain of parasites invading the new hosts and eventually pushing the do-gooder parasites to the extinction.

5

Proposition: Two-layer evolutionary systems with controlled migration can produce slack.

I am going to get handwavy here.

Imagine that the control exercised by the outer system is incomplete. Not because the outer system would get any short-term evolutionary advantage from relaxing the control but simply because the controlled individuals sometimes manage to escape and find a different host. It is not clear how such a relationship could be stable though. It seems that one side must prevail: Either the host gets full control and what we get is a mutually beneficial relationship or the symbiont manages to escape the control and becomes a harmful parasite. (But look [here](#).)

All that being said, there is one mechanism that seems to fit the bill. It's sex. Individual genes are able to escape the genome, but it happens in a very controlled manner.

Does that mean that sex gives species slack? One hint would be that sexual species in general evolve slower than asexual species. In other words, they are slacking. But I am not at all sure. If there's a mysterious question in evolutionary biology, it's the very existence of sex. Entire books were written about it and there are dozens of proposed solutions. I don't feel confident enough to dig any deeper here, but it still may be worth considering.

Covid-19: Comorbidity

We've all seen statistics that *most* people who die of Covid-19 have at least one comorbidity. They also almost all have the particular comorbidity of age. The biggest risk, by far, is being old. The question that I don't see being properly asked anywhere (I'd love for this post to be unnecessary because there's a better one) is: What is your chance of death from Covid-19 if infected, conditional on which if any comorbidities you have? Which ones matter and how much? If you don't have any, how much better off are you than your age group in general?

Thus, most people are looking at the age chart, without adjusting for their health status, *unless* they have an obvious big issue, in which case they adjust up. Which leads to an incorrect overall answer. That isn't obviously a bad thing in terms of resulting behavior in practice, but that doesn't mean we shouldn't attempt to figure out the answer.

I used New York State's information on deaths to get the rates of most of the major comorbidity candidates by age, and various Google-fu combined with wild mass approximation and fitting to different age groupings (not guessing, but not fully not guessing either) to get approximate population prevalence data. What matters?

A key question will always be, is this a proxy for something else, such as poverty, general poor health or obesity? Or is it the real problem? Here, we need to use some common sense and physical intuition. This isn't attempting to be super rigorous, but rather to get an approximation.

Then, once we've looked at all of them, I'll attempt to put them all together, and solve for the risk of someone in good overall health.

[Spreadsheet is here](#), you can look at the numbers in somewhat more detail, and see some of the sources I used, on the Comorbidity tab.

In each case, the "Population X" column attempts to guess the rate at which the population has it. The morbidity column is the rate at which those in NY state that died of Covid-19 had it.

For the first few graphs I fit NY's data to the groupings in the prevalence data I found. Later I started always using NY's ranges instead.

Hypertension

Age	Morbidity	Population
	Hypertension	Hypertension
18-39	21%	8%
40-59	43%	33%
60+	61%	63%

Hypertension in the young seems to matter. If you are 18-39, your relative chance of dying more than doubles. In your 40s and 50s combined, it's a jump of about a third. Above age of 60, it does not matter. Should we be suspicious that this is a proxy for

poor health, given that? Somewhat, definitely. It will be a recurring pattern, which we'll need to get to make sense.

Diabetes

Age	Morbidity Population	
	Diabetes	Diabetes
18-44	28%	4%
45-64	40%	17%
65-74	44%	25%
75+	34%	25%

Clearly this is a *huge* deal at young ages. I still don't really believe the 4% number for the general population, but multiple sources are around there. That 4% of the population is *more than a quarter* of all deaths under 45 in New York. That's a huge deal. Diabetes is a huge deal the entire way. This makes some sense, as it seems likely to correlate with slash cause direct physical problems for those with Covid-19 that can kill them. But this amount of effect is still surprisingly extreme.

Hyperlipidemia

Age	Morbidity	Population
	Hyperlipidemia	Hyperlipidemia
0-4	0%	7%?
5-14	0%	could
15-29	4%	be
30-39	5%	up
40-49	9%	to
50-59	16%	45%
60-69	22%	in
70-79	25%	adult
80-89	24%	pop?
90+	20%	so
Unknown	6%	meaningless

I could not get population numbers, because no one can agree on what Hyperlipidemia actually is and is not. By some definitions, almost half the adult population has it, because people like to say that we need medications and are "at risk" and turn everything into a disease. By others, it's single digit percentages. So these morbidity numbers could be anything from scary high to same as the population, depending on the definition used, and I don't know what that was. If anyone does know, please tell me.

But given what this physically is, any link seems more like correlation than causation, and the numbers listed seem plausibly like general population rates anyway, so I'm going to say this likely doesn't matter.

Coronary Artery Disease

Age	Morbidity	Population
	C. Artery D.	C. Artery D.
0-4	0%	0%
5-14	0%	0%
15-29	0%	0%
30-39	0%	1%
40-49	3%	4%
50-59	6%	8%
60-69	11%	13%
70-79	14%	18%
80-89	16%	25%
90+	13%	26%

The population rates I approximated are modestly higher across the board. Probably there's no effect and this is a measurement error.

Dementia

Age	Morbidity	Population
	Dementia	Dementia
0-4	0%	0%
5-14	0%	0%
15-29	0%	0%
30-39	0%	0%
40-49	0%	0%
50-59	1%	0%
60-69	4%	1%
70-79	10%	5%
80-89	18%	24%
90+	28%	37%

It's weird that this reverses at older ages, probably because of measurement, but perhaps because people with dementia have overall better health at that age due to the ones with poor health having died more often? Whereas in younger people, if you have dementia things are much more likely to be generally terrible in other ways instead?

From the 50-79 year old data I'd have said this might matter, but the prevalence rate is low there, and where the rates are high, the numbers are reversed. I don't think dementia is doing work here.

Renal Failure

Morbidity	Population
-----------	------------

Age	Renal Failure	Renal Failure
18-44	6%	7%
45-64	11%	12%
65+	11%	37%

I suppose that those with renal failure die soon thereafter, thus the 37% population rate is not going to be reflected in an alive group. It's certainly not going to be protective. That implies there might be some real effect in the younger groups if you squint hard enough, but seems much more likely this is not a risk factor.

COPD

Age	Morbidity	Population
Age	COPD	COPD
0-4	0%	0%
5-14	0%	0%
15-29	1%	2%
30-39	1%	3%
40-49	2%	5%
50-59	5%	7%
60-69	8%	8%
70-79	10%	10%
80-89	10%	9%
90+	8%	8%

Very clear no effect.

Atrial Fibrulation

Age	Morbidity	Population
Age	Atrial Fibrulation	Atrial Fibrulation
0-4	0%	0%
5-14	0%	0%
15-29	0%	0%
30-39	1%	1%
40-49	1%	1%
50-59	2%	1%
60-69	4%	2%
70-79	8%	5%
80-89	12%	9%
90+	14%	11%

Some effect, with minimal impact on numbers for those without the condition, and likely the effect is correlational given the physical conditions involved.

Cancer

Morbidity	
Age	Cancer
0-4	0%
5-14	0%
15-29	3%
30-39	2%
40-49	2%
50-59	4%
60-69	7%
70-79	8%
80-89	9%
90+	8%

About 0.5% of New Yorkers get newly diagnosed with Cancer every year. But translating that into a background cancer rate by age proved very difficult. It certainly can't be what they mean by cancer here, since that means that at older ages cancer would be highly protective, so they're clearly only counting current conditions or some other similar thing.

It seems obvious that being actively sick from cancer *treatment* would make Covid-19 much worse to get, but it's not at all obvious the condition itself would matter much for many cancers, and they're also different conditions, so it's all very confusing. Not sure what to do here.

My guess is this is mostly 'general poor health caused by cancer or treatment' effects, to extent it matters.

Stroke

Morbidity Population		
Age	Stroke	Stroke
0-4	0%	0%
5-14	0%	0%
15-29	1%	0%
30-39	1%	1%
40-49	3%	1%
50-59	4%	2%
60-69	7%	5%
70-79	8%	10%
80-89	8%	15%
90+	6%	15%

Again, those numbers on the right are largely guesses. It does seem clear that stroke is a risk factor when you are young. We once again see strangely low morbidity rates

for the older age groups, pointing to a likely general difference in methodology. Probably selection effects.

That's everything listed by New York.

Obesity

When you are young, it seems like it matters a *lot* whether there is something relevantly seriously wrong. But it has to be something that matters. Having a health problem in an area that Covid-19 doesn't attack does not seem to matter.

Two of the problems measured matter a lot. Hypertension and Diabetes together are about 12% of the population and constitute ~50% of the deaths up to age 40-45. The rest don't seem to matter much at all, and you could safely ignore them.

It is safe to assume that anyone with serious trouble breathing for other reasons is going to have similar problems. Thus, asthma, obesity and so on are also (probably) serious risk factors.

One source I found was [this one](#), which notes that 35.8% of hospital patients with Covid-19 were obese. In the city, 22% of the population is obese (and a majority are at least overweight). 43% of the invasive treatment group, which was presumably in far worse shape, was obese versus 31% for the non-invasive group, so it seems that the extra risk carries over to outcomes after hospitalization.

[This source](#) found an inverse correlation in Covid-19 patients between age and BMI, which would also make sense.

From [another source](#) at ScienceNews:

For instance, of 180 patients hospitalized from March 1 to March 30, the [most prevalent underlying condition](#) for adults ages 18 to 49 was obesity. Of 39 patients in that age range, 23, or 59 percent, were obese, researchers report in the April 17 *Morbidity and Mortality Weekly Report*.

...

Lighter and her colleagues found that patients under 60 with a BMI over 35 were at least [twice as likely to be admitted to the ICU](#) for coronavirus than patients with healthy BMIs, the researchers report April 9 in *Clinical Infectious Diseases*. Those same patients were three times more likely to die from the infection than those with a lower BMI, she says.

The team tracked 3,615 people who tested positive for SARS-CoV-2, the virus that causes COVID-19, at a New York City hospital from March 4 to April 4. Of those, 1,370, or 38 percent, were obese. In patients over 60, weight did not appear to be a factor in hospital admission or the need for intensive care, she says.

Again, it seems like 'nothing matters much if you're old.' But if you're young, things do matter.

Let's say for the time being that obesity triples your risk if you're under 60. Obviously, this doesn't go away the moment you turn 60, so we'll want to do *more than triple* for

those under 40, much less than triple for those in their 50s, and some effect probably in your early 60s.

Continuing to do approximations, [if we accept the figure](#) that 52% of Type II Diabetics were obese back in 2006, whereas Type I is actually lower than the population rate, which is now more like 40%. New York City's seemingly terrible 22% looks positively amazing by comparison if actually the same measure.

This accounts for substantial increased risk for diabetics, but the majority clearly remains unexplained by weight. Obesity alone would have gotten us from 4% to about 10% in the young group, and we ended up at 28%.

Hypertension, on the other hand, seems mostly to be a proxy for obesity, so we can mostly ignore it.

Overall, though, obesity seems *by far* the most important consideration other than age, since it's so common and has such a huge impact.

Age Alone

To adjust from a baseline we need a baseline. For New York the relative risks look like this:

	%Of Deaths	Population	Relative Risk
0-4	0.02%	7%	0.2%
5-14	0.04%	14%	0.3%
15-29	0.3%	22%	2%
30-39	1%	16%	9%
40-49	4%	14%	26%
50-59	10%	9%	116%
60-69	20%	7%	287%
70-79	27%	5%	525%
80-89	25%	2%	1066%
90+	12%	1%? (cuts off at 85)	1500% or so?

Then we must guess the true IFR (infection fatality rate), and adjust for the three comorbidities that we've found matter: Obesity and diabetes. They still correlate.

We also have to adjust for the likelihood of infection in the first place, since that changes your risk conditional on infection. This is a relatively small effect according to antibody test results, and should at least *sort of* be cancelled out in some ways for practical purposes, so I'm not going to worry much about it.

Most coronavirus cases are definitely not being detected by positive tests. The antibody tests show that. So the IFR is much lower than the CFR. Given death rates and antibody tests, the plausible range for IFRs is about 0.5% to 1.5%. It will also depend on conditions on the ground in various ways, of course. But as a baseline, I'm going to continue to say 1% death rate for the state. If you disagree, multiply all the numbers I get as appropriate.

Obesity gets slightly more common with age, which I'll adjust for.

For the younger groupings, we have 4% Diabetes and (in New York) 25% Obese. Obesity dominates that group by size, but diabetes still matters. Together that's about 27% of the population. The 4% of that that are diabetic account for 28% of the cases. The other 23% that are obese become 49% of the remaining cases, or 35% of all cases. Add that together, and that's 63% of cases from this 27%, with the remaining 37% coming from the other 73% of the population. So if you're healthy, with healthy defined as 'not obese and not having diabetes' your risk is cut roughly in half when young. There's too many error bars all over the place, so I don't want to try and be more exact than that.

Other conditions doubtless also matter somewhat. 10% of New York has asthma, which presumably makes a big difference, but all I could find was "may be at higher risk" repeated over and over, rather than any numbers.

This all these effects decay as you age. By age 70, there's little or no difference.

But roughly you end up with a chart that looks something *vaguely* like this:

Reminder: This assumes overall infection fatality rate of ~1% and is full of guesses and approximations as all hell:

Age	Risk (Healthy)	Risk (Diabetes)	Risk (Obesity)	Risk (All Pop)
0-4	0.001%	0.013%	0.005%	0.002%
5-14	0.002%	0.016%	0.006%	0.003%
15-29	0.008%	0.078%	0.031%	0.015%
30-39	0.043%	0.26%	0.17%	0.09%
40-49	0.13%	0.53%	0.43%	0.26%
50-59	0.72%	1.8%	1.4%	1.2%
60-69	2.0%	4.0%	3.0%	2.9%
70-79	4.3%	6.5%	5.4%	5.2%
80-89	11%	11%	11%	11%
90+	15%	15%	15%	15%

GPT-3: a disappointing paper

[Note: I wrote this post in late May 2020, immediately after the GPT-3 paper was released.]

This post is a compilation of two posts I recently made on tumblr.

For context: I have been an enthusiastic user of GPT-2, and have written a lot about it and transformer models more generally. My other writing on this topic includes [human psycholinguists: a critical appraisal](#) and ["the transformer ... "explained?"](#) See also [my tumblr bot](#), which uses GPT-2 as a core component.

Part 1

[argumate](#) said:

[@nostalggebraist](#), give us the goss on how GPT-3 compares with GPT-2!

I haven't read [the paper](#) super carefully yet, but I am pretty sure of the following:

1.1: On GPT-3's mundanity

"GPT-3" is just a bigger GPT-2. In other words, it's a straightforward generalization of the "just make the transformers bigger" approach that has been popular across multiple research groups since GPT-2.

This excerpt captures this pretty clearly:

Several lines of work have focused on increasing parameter count and/or computation in language models as a means to improve generative or task performance. [...] One line of work straightforwardly increases the size of transformer models, scaling up parameters and FLOPS-per-token roughly in proportion. Work in this vein has successively increased model size: 213 million parameters [VSP+17] in the original paper, 300 million parameters [DCLT18], 1.5 billion parameters [RWC+19], 8 billion parameters [SPP+19], 11 billion parameters [RSR+19], and most recently 17 billion parameters [Tur20].

The first two papers mentioned here are the original transformer for machine translation (VSP+17) and BERT (DCLT18). The parameter count doesn't actually increase that much between those two.

The third one (RWC+19) is GPT-2. The parameter counts jumps up 5x there. Arguably the point of the GPT-2 paper was "it sounds dumb and too easy, but amazing things happen if you just make a transformer bigger" – and this "GPT-3" paper is making the same point with bigger numbers.

"GPT-3" is a transformer with *175 billion* parameters. It's another big jump in the number, but the underlying architecture hasn't changed much.

In one way this is a fair thing to call "GPT-3": it's another step in the new biggening tradition which GPT-2 initiated.

But in another way it's pretty annoying and misleading to call it "GPT-3." GPT-2 was (arguably) a fundamental advance, because it demonstrated the power of way bigger transformers when people *didn't know* about that power. Now everyone knows, so it's the furthest thing from a fundamental advance. (As an illustration, consider that their new big model deserves the title "GPT-3" just as much, *and just as little*, as any of the last 3 big models they mention in that paragraph.)

1.2: On "few-shot learning"

The paper seems very targeted at the NLP community, which I mean in almost a wholly negative way. (Despite being part of the NLP community, I guess.)

The GPT-2 paper argued that language models (text predictors) could do well, or in some cases "at least not terribly," at the specialized tasks used as NLP benchmarks – even without

being told anything about those tasks. This was sort of neat, but mostly as a demonstration of the language model's power.

The “zero-shot” learning they demonstrated in the paper – stuff like “adding tl;dr after a text and treating GPT-2’s continuation thereafter as a ‘summary’” – were weird and goofy and not the way anyone would want to do these things in practice. It was more cool as a demonstration that sufficiently good language models could “do it all,” even things they weren’t intended for; the point wasn’t that they were *world-class great* at these tasks, the point was the gap between their performance and their low level of preparation. Kinda like a child prodigy.

In the GPT-3 paper, they’ve introduced a new (...ish? maybe?) way for language models to be good at the standard benchmarks. Now it’s about how they can “figure out” what they’re supposed to be doing across the course of a text, i.e. instead of prompting the model with *one* thing like

Q: What is the capital of France?

A:

they instead prompt it with several, like

Q: What is the capital of France?

A: Paris

Q: What is the capital of Spain?

A: Madrid

Q: What is the capital of Lithuania?

A: Vilnius

Q: What is the capital of Brazil?

A:

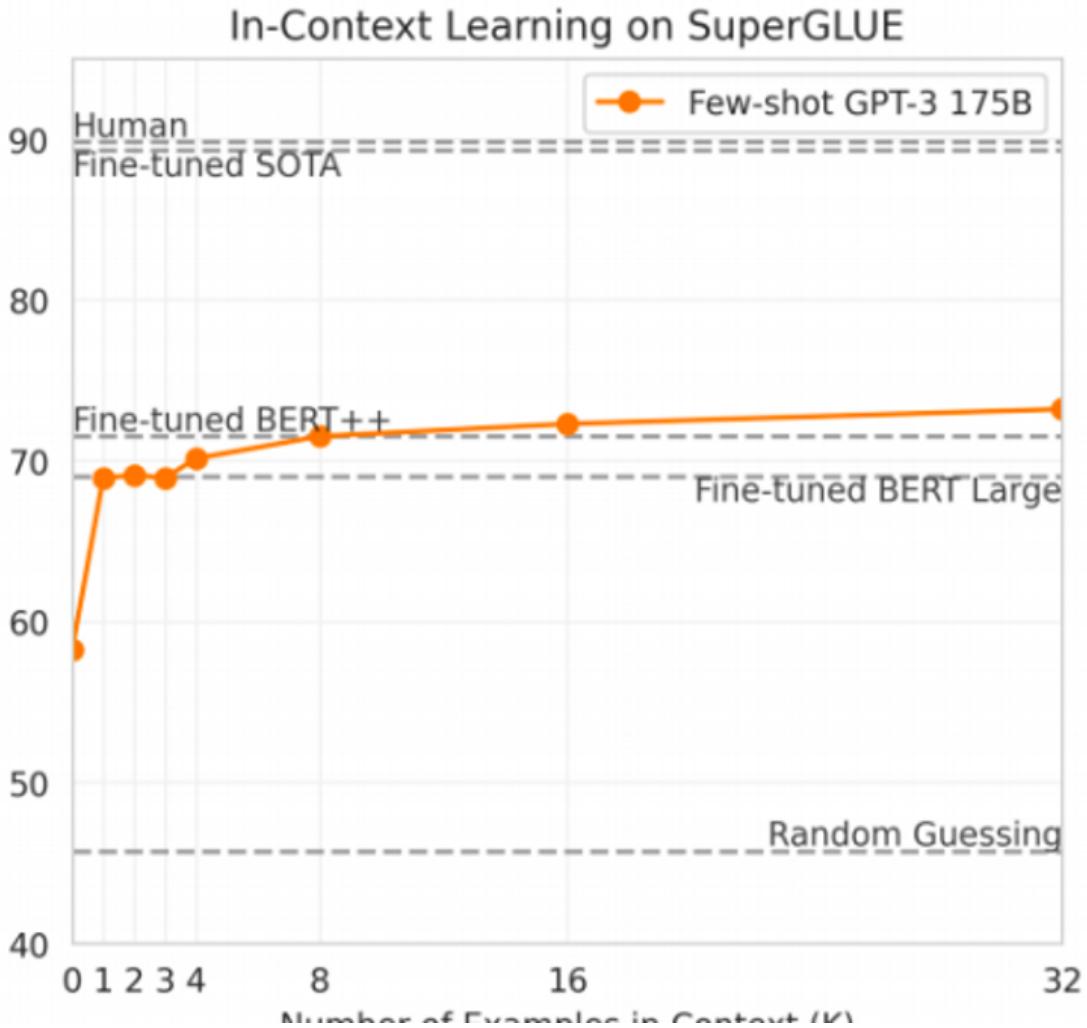
The NLP-community-relevant point of “GPT-3” is that language models can do much better on the standard benchmarks than we thought, via this kind of multi-prompting and also via even more biggening. Putting those two changes together, you can even even beat the state of the art on a few tasks (of many).

I can imagine someone viewing this as very important, if they thought it showed an ability in transformer LMs to “pick things up on the fly” in an extremely data-efficient, human-like way. That would be relevant to some of [Gary Marcus’ concerns](#).

But the paper seems totally, *weirdly* uninterested in the “learning on the fly” angle. Their paper has many, many figures graphing performance against papemeter count – bigger is better yet again – but I can only find one figure graphing performance against their parameter K, the number of distinct task examples in the prompt (K is 1 and 4 in the two capitals examples).

[It turns out there’s another one I missed on my first read – Fig. 1.2 on page 4. I discuss this in Part 2 below.]

And that figure is, uh, not encouraging:



They do better with one task example than zero (the GPT-2 paper used zero), but otherwise it's a pretty flat line; evidently there is not too much progressive "learning as you go" here.

(Oddly, the caption for this figure explains these are *dev set* results so not directly comparable to the *test set* results given as horizontal lines – which doesn't stop them from plotting them! Elsewhere, they do report test set results for SuperGLUE, but only for $K=32$. Also, I'm not a fan of this plot's lack of error bars.)

1.3: On benchmarks

Instead, their interest is almost completely in how good they can get on the benchmarks in absolute terms.

This is why I say it's aimed at the NLP community: these are the metrics that whole community measures itself against, so in a trivial sense the community "has to" find these results interesting. But by now, this starts to feel like Goodhart's Law.

The reason GPT-2 was so cool wasn't that it did so well on these tasks. It was that it was a really good language model that demonstrated a new *overall understanding of language*. Coercing it to do well on standard benchmarks was valuable (to me) only as a flamboyant, semi-comedic way of pointing this out, kind of like showing off one's artistic talent by painting (but not painting especially *well*) with just one's non-dominant hand.

GPT-2 isn't cool because it's good at "question answering," it's cool because it's so good at *everything* that it makes caring about "question answering" per se feel tiny, irrelevant.

The transformer was such an advance that it made the community create a new benchmark, "SuperGLUE," because the previous gold standard benchmark (GLUE) was now *too easy*.

GPT-3 is so little of an advance, it doesn't even do that well at SuperGLUE. It just does *okay* with its dominant hand tied behind its back.

"No, my 10-year-old math prodigy hasn't proven any new theorems, but she *can* get a perfect score on the math SAT in under 10 minutes. Isn't that groundbreaking?"

Sort of? Not especially?

1.4: On annoyance

The more I think about this paper, the more annoying it is. Transformers are extremely interesting. And this is about the least interesting transformer paper one can imagine in 2020.

Part 2

2.1: On "few-shot learning," again

On my first read, I thought there was only one plot showing how performance varies with K (number of few-shot samples), but I missed the one very early in the paper, Fig 1.2 on p. 4.

That plot is more impressive than the other one, but doesn't change my impression that the authors are not very interested in showing off "progressive learning" over the course of a text.

The argument they're trying to make with Fig 1.2 is that *more* progressive learning happens with bigger models, and hence that their overall strategy – "use *big models + few-shot learning* to get good scores on benchmarks" – benefits from an interaction effect above and beyond the independent effects of its two parts (big models, few-shot learning).

Again, this is interesting if you care about scores on NLP benchmarks, but I have trouble seeing much qualitative significance for overall language understanding.

2.2: On novel words

One of their experiments, "Learning and Using Novel Words," strikes me as more remarkable than most of the others and the paper's lack of focus on it confuses me. (This is section 3.9.5 and table 3.16.) The task is closely related to the Wug test – it's the kind of thing Gary Marcus focused on in his critique of GPT-2 – and looks like this:

[Human prompt] To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

[GPT-3 continuation] One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduckles.

This is the sort of task that developmental linguists study in human children, and which past NLP models have had trouble with. You'd think a success on it would deserve top billing. The authors apparently report a success here, but treat it as an unimportant sideshow: they say they tried it 6 times and got 6 successes (100% accuracy?!), but they apparently didn't consider this important enough to try the same thing on a larger sample, compute a real metric, show variance w/r/t parameters, etc. Meanwhile, they did those things on something like 40 other tasks, mostly far less interesting (to me). Confusing!

2.3: On abstract reasoning

In addition to the usual NLP benchmarks, they tried some "synthetic or qualitative" tasks (section 3.9). Their stated goal with these is to clarify the role the actual *learning* in "few-shot learning," separating it from mere familiarity with similar-looking text:

One way to probe GPT-3's range of abilities in the few-shot (or zero- and one-shot) setting is to give it tasks which require it to perform simple on-the-fly computational reasoning, recognize a novel pattern that is unlikely to have occurred in training, or adapt quickly to an unusual task.

The "synthetic or qualitative" tasks are:

- various forms of simple arithmetic (like “add two 2-digit numbers”)
- various anagram/reversal/etc tasks operating on the individual letters of words
- SAT analogies

This line of work feels insufficiently theorized, and thus hard to interpret.

Consider the arithmetic tasks. Let’s grant the authors’ premise that the model has not just memorized some lookup table for arithmetic problems – it’s really “doing the problems” on the fly. Then, there are 2 things the model could be doing here (probably some of each simultaneously):

1. It might have developed a real internal model of arithmetic from seeing many related numbers in training texts, and is applying this model to do the problems like you or I would
2. It might have developed some generic reasoning capability for arbitrary abstract tasks, which can handle arithmetic as a particular case of a much more generic class of problems (e.g. it could also pick up various “fake arithmetics” where +, -, etc have non-standing meanings, if appropriately prompted)

Insofar as #1 is happening, the multiple prompts of few-shot learning *shouldn’t matter*: if the model knows how real (not fake) arithmetic works because it’s seen it in text, then additional examples don’t help “locate the task.” That is, if it has *only* learned to do real arithmetic, it shouldn’t need to be told “in this task the + symbol has the standard meaning,” because its ability depends on that assumption anyway.

So, if we’re mostly seeing #1 here, this is not a good demo of few-shot learning the way the authors think it is.

Insofar as #2 is happening, the few-shot prompts *do* matter: they “locate the meanings” of the symbols in the large space of possible formal systems. But #2 is *wild*: it would represent a kind of non-linguistic general intelligence ability which would be remarkable to find in a language model.

I really doubt this is what the authors are thinking. If they think language models are fully general reasoners, why not highlight that? The abstract reasoning capacity of transformers has already been [more clearly probed](#) without the confounding aspects of natural language, and a priori there are few reasons to think a very large language-specific model should develop strong abilities here (while there *are* a priori reasons to think the abilities are subtle forms of text recognition/memorization the authors’ methodology was not able to detect).

My best guess is that the authors imagine a factorization of the task into “knowing how to do it” and “knowing we are doing it right now.” Training on text teaches you how to do (real) arithmetic, and the few-shot prompts tell you “right now we are doing (real) arithmetic, not some other thing you know how to do.”

But arithmetic is a really bad choice if you want to probe this! The authors use K=50 here, meaning they give the model 50 correct examples of simple math problems to let it “locate the task.” But no one who *can* do this task should need 50 examples of it.

What information is conveyed by example #50 that wasn’t already known by example #49? What are we ruling out here? Trollish formal systems that look like addition 98% of the time? “Addition, except ‘52’ actually means ‘37’ but everything else is the same?” Do we have to rule this out when you should have (and the model must have) a strong prior towards real addition?

I don’t know what the authors are trying to do here, and I think they may not know, either.

How does publishing a paper work?

While I've never published a research paper and have no plans to do so, I realized I don't even know how the process works. These are the bits and pieces I think I know (probably wrong about some):

- Papers are annoying 2-column pdfs
- Getting a paper published takes a lot of work beyond the research itself
- When a paper has multiple collaborators or a student/professor relationship, there's an awkward political negotiation about whose name is included and whose name goes first, last, or in the middle of the list
- There are multiple journals you can submit to and maybe none will accept you, or maybe you'll get multiple offers and then I don't know if you have to pick one at most
- When you submit a paper to a journal, the journal sends it out to your peers who submit anonymous feedback before publishing, which seems like more trouble than it's worth these days because the peers might be slow or unfair, or be playing a zero-sum game competing with you
- Many academic conferences have their own associated journals which you can submit papers to and in some cases getting accepted to that conference-journal means you get to give a talk at that conference
- Paid-access journals currently have a monopoly on high-status research publication, and academia is stuck in this local maximum that's hard to dislodge without a coordinated effort to agree on how to publish in a high-status place that isn't a paid journal, and in the meantime the journals get to rent-seek in a way that tragically/comically undermines the ideal of academic research not being a capitalist enterprise
- arXiv is a place where you can upload papers for free and people can download them for free, thereby bypassing the journals to some degree
- Sci-hub lets anyone illegally download pirated papers that normally require access to a paid journal
- Publishing papers is a valuable thing to do because it gives the content of the paper and its author(s) a certain social legitimacy, and allows future research to frictionlessly cite your findings

Can someone confirm or correct my impressions, and elaborate on any other interesting parts?

Plague in Assassin's Creed Odyssey

Spoiler Alert: Contains minor spoilers for Assassin's Creed Odyssey

Assassin's Creed Odyssey has become my quarantine game. A gorgeous tour of ancient Greece is a great antidote to never going outside.

One noteworthy thing in the game is how it deals with plague. You encounter it twice.

On the game's introductory island of Kefalonia, you can run into a side quest called The Blood Fever. There is a plague, and priests are attempting to contain it... by killing off the families of the infected.

You can either allow this, or you can forcibly prevent it, in which case you have to kill the priests.

If you do not prevent the executions, you are called a murderer. You try to 'reassure' your friend that this was necessary. Later in the game, the Oracle of Delphi tells you that you're a bad person, because you let yourself be bossed around by an unknown plague.

If you prevent the executions, you are called a hero. Later in the game, you are informed a plague has spread throughout Kefalonia. No one makes the connection or blames you.

If you also let them keep their money, you are told the plague has spread throughout Greece. So being superficially generous makes things much, much worse.

This all feels right and seems to explain a lot. It is clear what must be done, but the incentives are all wrong. Even when you are right you still get punished for allowing others to take action. If you prevent others from taking action, even with disastrous consequences, you are considered a hero.

Later, you are given a quest to warn about The Plague of Athens. The plague is historical, so it can't be prevented. Warning about it has no effect. Again, this feels right.

When the Plague of Athens does hit, you are given the task to burn some of the bodies to prevent further spread, as Hippocrates (yes, that one, these games are like that) has figured out this will help prevent further infection. The Followers of Ares try to stop you for religious reasons, and you have to kill them. Once again, society primarily moves to violently prevent action.

Then Pericles breaks quarantine and goes outside when everyone tells him not to, and gets himself killed for it. That part is not historical. It felt right in context.

Experiencing the game helped me process why we've reacted to today's plague the way we did.

Trust-Building: The New Rationality Project

"I did not write half of what I saw, for I knew I would not be believed"

-Marco Polo

It's enlighteningly disturbing to see specifically how the "distrust those who disagree" heuristic descends into the madness of factions.

-Zack M Davis

Old LessWrong: we fail to reach the truth because of our cognitive biases.

New LessWrong: we fail to reach the truth because of reasonable mistrust.

What causes mistrust? It's a combination of things: miscommunication and mistakes; the presence of exploitative actors; lack of repeated trust-building interactions with the same people; and high payoffs for epistemically unvirtuous behavior. Enough mistrust can destroy our ability to converge on the truth.

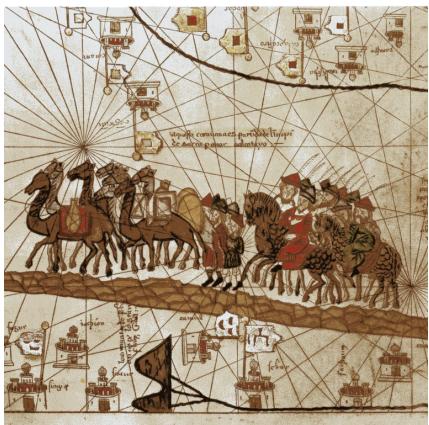
Once mistrust has led us into factionalism, can we escape it?

Factionalism is not a bad thing. It means that we are willing to accept some evidence from other people, as long as it's not too divergent from our priors. Factions are worse than unity, but better than isolation.

What we want isn't a lack of factionalism, it's unity.

This suggests an activism strategy. Let's form three categories of factions:

- **Your Community:** You have high trust in this network, and believe the evidence you receive from it by default. Although you don't trust everyone who calls themselves a part of this network or community, you know who belongs and who doesn't, who's reliable and who's not, and how to tell the difference when you meet someone new in the network.
- **Strangers:** This network is untested, and your community doesn't have a strong opinion on it. It would take substantial work to learn how to navigate this foreign network. But because they are relatively isolated from your own community, they have evolved a different constellation of evidence. The existence of factionalism is itself evidence that you'd have something new to gain by trading information with them.
- **Enemies:** This network has been examined, either by you or by your community, and labeled a toxic breeding ground of misinformation. Treating your enemies as though they were merely strangers would only alienate you from your own community. They might be exploitative or stupid, but either way, engaging with them can only make things worse. All there is to do is fight, deprogram, or ignore this lot.



To increase unity and pursue the truth, your goal is to find foreign communities, and determine whether they are friendly or dangerous.

Note that the goal is *not* to make more friends and get exposed to new ideas. That's a recipe for naivete. The real goal is to accurately distinguish strangers from enemies, and make introductions and facilitate sharing with *only the stranger, but not the enemy*. We might respect, disparage, or ignore our enemies, but we know how to tell them apart from our allies and our own:

"Here people was once used to be honourable: now they are all bad; they have kept one goodness: that they are greatest boozers."

One of the many difficulties is the work it takes to discover, evaluate, and bring back information about new strangers, the truly foreign. Their names and ideas are most likely unknown to anybody in your community, and they speak a different language or use different conceptual frameworks.

Worse, your community is doing a steady business in its own conceptual framework. You don't need to just explain why the new people you've discovered are trustworthy; you need to explain why their way of thinking is valuable enough to justify the work of translating it into your own language and conceptual framework, or learning their language.

Luckily, you do have one thing on your side. Foreign communities usually *love* it when strangers express a genuine interest in absorbing their ideas and spreading them far and wide.

You might think that there is a time to explore, and a time to move toward a definite end. But this isn't so.

When there's a definite end in mind, moving toward it is the easy part.

But meaning and value mostly come from novelty.

When it feels like there's no need to explore, and all you need to do is practice your routine and enjoy what you have, the right assumption is that you are missing an opportunity. This is when exploration is most urgent. "What am I missing?" is a good question.

"You will hear it for yourselves, and it will surely fill you with wonder."

What is our community reliably missing?

Writing Causal Models Like We Write Programs

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

Clunc

We'll start with a made-up programming language called Clunc. The distinguishing feature of clunc is that it combines classes and functions into a single type, called a clunc. It looks like this:

```
quad = Clunc {
    x = 4
    constant = 3
    linear = 2*x + constant
    result = x*x + linear
}
```

We could then go to a terminal:

```
>>> quad.result
27
>>> quad.linear
11
```

In order to use this clunc like a function, we apply the do() operator. For instance,

```
>>> quad3 = do(quad, x=2)
```

... creates a new clunc which is just like quad, except that x is 2 rather than 4:

```
>>> quad3
Clunc {
    x = 2
    constant = 3
    linear = 2*x + constant
    result = x*x + linear
}
```

When we query fields of quad3, they reflect the new x-value:

```
>>> quad3.result
11
>>> quad3.linear
7
```

There's no designated "input" or "output"; we can use the do() operator to override any values we please. For instance

```
>>> quad_zero_linear = do(quad, linear=0)
>>> quad_zero_linear
```

```

Clunc {
    x = 4
    constant = 3
    linear = 0
    result = x*x + linear
}
>>> quad_zero_linear.result
16

```

A few quick notes:

- Clunc is purely clunctional: everything is immutable, and each variable can only be written once within a clunc. No in-place updates.
- Clunc is lazy.
- Variables can be set randomly, e.g. “`x = random.normal(0, 1)`”.
- The `do()` operator creates a new clunc instance with the changes applied. If there are any random variables, they are re-sampled within the new clunc. If we want multiple independent samples of a randomized clunc `M`, then we can call `do(M)` (without any changes applied) multiple times.

To make this whole thing Turing complete, we need one more piece: recursion. Cluncs can “call” other cluncs, including themselves:

```

factorial = Clunc {
    n = 4
    base_result = 1
    recurse_result = do(factorial, n=n-1).result
    result = (n == 0) ? base_result : n * recurse_result
}

```

... and that’s where things get interesting.

Causal Models

Hopefully the mapping from clunc to probabilistic causal models is obvious: any clunc with random variables in it is a typical [Pearl-style causal DAG](#), and the `do()` operator works exactly like it does for causal models. The “clunc” is really a *model*, given by structural equations. The one big change is the possibility of recursion: causal models “calling” other models or other instances of themselves.

To get some practice with this idea, let’s build a reasonably-involved analogue model of a [ripple-carry adder circuit](#).

We’ll start at the level of a NAND gate (levels below that involve equilibrium models, which would require a bunch of tangential explanation). We’ll assume that we have some model `MNAND`, and we use `do(MNAND, a =..., b =...).result` to get the (noisy) output voltage of the NAND gate in terms of the input voltages `a` and `b`. Since we’re building an analogue model, we’ll be using actual voltages (including noise), not just their binarized values.

We'll take M_{NAND} as given (i.e. assume somebody else built that model). Building everything out of NAND gates directly is annoying, so we'll make an XOR as well:

```
Mxor = Model{
```

```
    a = 0.0
```

```
    b = 0.0
```

```
    intermediate = do(MNAND, a = a, b = b).result
```

```
    left = do(MNAND, a = a, b = intermediate).result
```

```
    right = do(MNAND, a = intermediate, b = b).result
```

```
    result = do(MNAND, a = left, b = right).result
```

```
}
```

This looks like a program which performs an XOR using NAND gates. But really, it's a Pearl-style causal DAG model which uses a lot of NAND-submodels. We can write out the joint probability distribution

$P[a = a^*, b = b^*, \text{intermediate} = i^*, \text{left} = l^*, \text{right} = r^*, \text{result} = \text{result}^* | M_{xor}]$ via the usual method, with each line in the model generating a term in the expansion:

$P[a = a^* | M_{XOR}] = I[a^* = 0]$

$P[b = b^* | M_{XOR}] = I[b^* = 0]$

$P[\text{intermediate} = i^* | M_{XOR}, a = a^*, b = b^*] = P[\text{result} = i^* | \text{do}(M_{NAND}, a = a^*, b = b^*)]$

$P[\text{left} = l^* | M_{XOR}, a = a^*, \text{intermediate} = i^*] = P[\text{result} = l^* | \text{do}(M_{NAND}, a = a^*, b = i^*)]$

$P[\text{right} = r^* | M_{XOR}, \text{intermediate} = i^*, b = b^*] = P[\text{result} = r^* | \text{do}(M_{NAND}, a = l^*, b = b^*)]$

$P[\text{result} = \text{result}^* | M_{XOR}, \text{left} = l^*, \text{right} = r^*] = P[\text{result} = \text{result}^* | \text{do}(M_{NAND}, a = l^*, b = r^*)]$

The full distribution is the product of those terms.

That's just the first step. Next, we need a [full adder](#), a circuit block which computes the sum and carry bits for one "step" of binary long addition. It looks like this:

```

Mfull_adder = Model{
    a = 0
    b = 0
    c = 0
    sab = do(MXOR, a = a, b = b).result
    s = do(MXOR, a = sab, b = c).result
    carryab = do(MNAND, a = a, b = b)
    carryc = do(MNAND, a = sab, b = c)
    carry = do(MNAND, a = carryab, b = carryc)
}

```

As before, we can write out the components of the joint distribution line-by-line. I'll just do a few this time:

$$P[a = a^* | M_{\text{full_adder}}] = I[a^* = 0]$$

...

$$P[s_a b = s_{ab}^* | M_{\text{full_adder}}, a = a^*, b = b^*] = P[\text{result} = s_{ab}^* | M_{\text{XOR}}, a = a^*, b = b^*]$$

$$P[s = s^* | M_{\text{full_adder}}, s_{ab} = s_{ab}^*, c = c^*] = P[\text{result} = s^* | M_{\text{XOR}}, a = s_{ab}^*, b = c^*]$$

...

Notice that some of these involve probabilities on the model M_{XOR} , which we could further expand using the joint distribution of M_{XOR} variables from earlier.

Finally, we can hook up a bunch of full adders to make our 32-bit ripple-carry adder:

The components of the joint distribution for this one are left as an exercise for the reader.

Why Is This Useful?

Classes/functions let us re-use code; we don't have to repeat ourselves. Likewise, clunky causal models let us re-use submodels; we don't have to repeat ourselves.

Obviously this has many of the same advantages as in programming. We can modularize our models, and fiddle with the internals of one submodel independently of other submodels. We can “subclass” our models via the `do()`-operator, to account for different contexts. Different people can work on different submodels independently - we could even imagine libraries of submodels. An electrical engineer could write a probabilistic causal model representing the low-level behavior of a chip; others could then import that model and use it as a reference when designing things which need to work with the chip, like packaging, accessories, etc.

From a more theoretical perspective, when we write programs with unbounded runtime, we have to have some way to re-use code: there's only so many lines in the program, so the program must visit some of the lines multiple times in the course of execution. Some lines must be re-used. Likewise for probabilistic models: if we want to define large models - including unbounded models - with small/finite definitions, then we need *some* way to re-use submodels. We could do that by writing things like " $\forall i : <\text{submodel}_i>$ ", but if we want Turing completeness anyway, we might as well go for recursion.

From a pure modelling perspective, the real world contains lots of repeating structures. If we're modelling things like cars or trees, we can re-use a lot of the information about one car when modelling another car. We think of cars as variations on a template, and that's exactly what the `do()` operator provides: we give it some "template" model, and apply modifications to it. The corresponding inverse problem then says: given a world full of things which are variations on some templates, find the templates and match them to the things - i.e. learn to recognize and model "cars" and "trees". Clunc-ish causal models are a natural fit for this sort of problem; they naturally represent things like "corvette with a flat tire".

Finally, the main reason I've been thinking about this is to handle abstraction. Clunc-ish models make layers of abstraction natural; lower-level behaviors can be encapsulated in

submodels, just as we saw above with the ripple-carry adder. If we want to write abstraction-learning algorithms - algorithms which take in raw data and spit out multi-level models with layers of abstraction - then clunc-ish models are a natural form for their output. This is what multi-level world models look like.

Extracting Value from Inadequate Equilibria

[Much expanded from my comment [here](#). Pure speculation, but I'm confident that the bones of this make sense, even if it ends up being unrealistic in practice. Cross-posted from [Grand, Unified, Crazy.](#).]

A lot of problems are coordination problems. An easy example that comes to mind is scientific publishing: everybody knows that some journal publishers are charging ridiculous prices relative to what they actually provide, but those journals have momentum. It's too costly for any individual scientist or university to buck the trend; what we need is coordinated action.

Eliezer Yudkowsky talks about these problems in his sequence [Inadequate Equilibria](#), and proposes off-hand the idea of a [Kickstarter for Coordinated Action](#). While Kickstarter is a great metaphor for understanding the basic principle of "timed-collective-action-threshold-conditional-commitment", I think it's ultimately led the discussion of this idea down a less fruitful path because Kickstarter is focused on *individuals*, and most high-value coordination problems happen at the level of *institutions*.

Consider journal publishing again. Certainly a sufficient mass of individual scientists could coordinate to switch publishers all at once. But no matter what individual scientists agree to, this is not a complete or perfect solution:

- Almost no individual scientists are paying directly for these subscriptions – their universities are, often via long-term bulk contracts.
- University hiring decisions involve people in the HR and finance departments of a university who have no interest in a coordinated “stop publishing in predatory journals” action. They only care about the prestige and credentials of the people they hire. Publications in those journals would still be a strong signal for them.
- Tenure decisions involve more peer scientists than hiring, but would suffer at least partly from the same issue as hiring.

What's needed for an action like this isn't a Kickstarter-style website for scientists to sign up on – it's coordinated action between universities at an institutional level. Many of the other examples discussed in *Inadequate Equilibria* fit the same pattern: the problems with healthcare in the U.S. aren't caused by insufficient coordination between individual doctors, they're caused by institutional coordination problems between hospitals, the FDA, and government.

(Speaking of government, there's a whole host of other coordination problems [climate change comes to mind] that would be eminently more solvable if we had a good mechanism for coordinating the various institutions of government between countries. The United Nations is better than nothing, but doesn't have enough trust or verification/enforcement power to be truly effective.)

The problem with the Kickstarter model is that institutions qua institutions are never going to sign up for an impersonal website and pledge \$25 over a 60-day campaign to switch publishing models. The time scale is wrong, the monetary scale is wrong, the

commitment level is wrong, the interface is wrong... that's just not how institutions do business. Universities and hospitals prefer to do business via contracts, and lawyers, and board meetings. Luckily, there's still value to be extracted here, which means that it should be possible to make a startup out of this anyway; it just won't look anything like Kickstarter.

Our hypothetical business would employ a small cadre of lawyers, accountants, and domain experts. It would identify opportunities (e.g. journal publishing) and *proactively* approach the relevant institutions through the proper channels. These institutions would sign *crafted, non-trivial contracts* bound to the success of the endeavour. The business would provide fulfillment verification and all of the other necessary components, and would act as a trusted third-party. The existence of proper contracts custom-written by dedicated lawyers would let the existing legal system act as an enforcement mechanism. Since the successful execution of these contracts would provide each institution with significant long-term value, the business can fund itself over the long haul by taking a percentage of these savings off the top, just like Kickstarter.

This idea has a lot of obvious problems as well (the required upfront investment, the business implications of having its income depend on one or two major projects each year, the incentives it would have to manufacture problems, etc) but with a proper long-term-focused investor on board it seems like this could turn into something quite useful to humanity as a whole. Implementing it is well outside of my current skillset, but I would love to see what some well-funded entrepreneur with the right legal chops could make of something like this.

Thoughts?

SlateStarCodex 2020 Predictions: Buy, Sell, Hold

Previously: [Evaluating Predictions in Hindsight](#)

Epistemic Status: Having fun

Evaluating predictions is hard, especially about the future. Let's do it.

The most frustrating part of predictions is defining them carefully. A lot of Scott's 2020 predictions seem like they have a high enough probability of a disputed outcome that they'd require clarification before betting on them. A bunch of others say they're explicitly Scott's decision. Thus, I'll try to clarify how I interpret such proposals as part of my evaluation.

I'll be looking at the predictions as if they were markets, and asking whether I would buy (bet on the thing happening at those odds plus some fee), sell (bet against the thing happening at those odds plus some fee) or hold (not inclined to wager), and about where I'd put my fair. Note that this doesn't mean I'd bet *against Scott* because Scott believes the prices are fair. So we'd have to give him good enough odds that he'd be willing to bet.

First up, we have the Coronavirus predictions. You'd pay to know what you really think! Hence, betting markets.

1. Bay Area lockdown (eg restaurants closed) will be extended beyond June 15: 60%

Sell to 40%, if I'm interpreting this correctly. I'm reading this as "no major relaxation of lockdown conditions" with things extended as they are or harsher. Certainly allowing restaurants to open at any level of capacity would mean it fails.

Right now, California is [running in place](#) at very low levels. Almost no herd immunity is being built, and most hospital capacity is not being used at all. The economy is being sacrificed in the hopes that conditions improve, but how long can that continue? How long *should* it continue? How long would people continue to abide it under such conditions, with no end in sight?

This is soon enough that there's a decent chance that these realizations have not yet come at that time. And there's some chance that there's a treatment or vaccine that looks sufficiently promising that 'tough it out until the end' becomes reasonable. But I'm guessing, as I noted last time, that a partial reopening does little or no damage if done wisely, and I expect California to end up doing something of that type.

2. ...until Election Day: 10%

Hold. If anything, that seems high, assuming it means continuous lockdown until then rather than being locked down on election day. This has to both be necessary and sufficient to large enough extents to justify waiting an incredibly long time. But there's also a chance that this happens without a good justification. We interpret California's actions as 'good decision making' and that is a possible explanation but it can also be seen as 'abundance of caution' or 'California is really good at telling people they can't

do things' which would point in a different direction when the right decision goes the other way.

There's also the argument that, if it holds through June, that's kind of a decision to hold indefinitely so the conditional chance it lasts a *lot* longer can't be that low.

3. Fewer than 100,000 US coronavirus deaths: 10%

Sell a lot. The *official* count is 57,000 now and we are not substantially past the peak outside of a few areas. The way down won't be faster than the way up. Under 100,000 is close to a Can't Happen even if we get a best case style scenario.

4. Fewer than 300,000 US coronavirus deaths: 50%

Sell to 30%. This is in 2020 only, and official counts, and would require lower than current levels on average for the rest of the year. Right now, hospitals are not overwhelmed and states are looking to reopen soon. We'd need to hit this level to have substantial overall help from herd immunity. We'd need to make a lot of progress on many fronts, or have a strong treatment or vaccine quickly, to have a burn slow enough to stay under this number.

5. Fewer than 3 million US coronavirus deaths: 90%

Hold. Given what I think is the IFR, killing almost 1% of the population requires full system collapse. New York managed to get through a fifth of its population in a month without seeing a spike in the IFR, so I think this is worse than the high-death-rate scenarios that I think are plausible, especially given this is presumably the official death count. In the scenario where we get close, I expect a severe undercount. I expect a lot of people to be able to protect themselves even in a full out-of-control scenario (and in fact, in that scenario it makes more sense for people to take extreme measures and burn through savings and create debt to do so) and I expect herd immunity effects to protect us by 50% infection or so at most under realistic conditions. Giving this a 10% chance therefore seems like a lot, but betting at long odds on 'not a complete disaster' requires more confidence than I'd be willing to display here.

6. US has highest official death toll of any country: 80%

Buy to 90%. Realistically who is it going to be if it isn't us? China or India. No one else has a big enough population. India seems not that vulnerable due to physical conditions and likely won't be able to track things properly even when things get bad. So this comes down to how often China ends up with a higher death count than we have by end of year, *and they admit it*. Given what would happen to China if this did happen to them sufficiently that they'd be forced to admit it, I see such a scenario as highly unlikely.

7. US has highest death toll as per expert guesses of real numbers: 70%

Buy to 80%. Logic above applies. I can see assigning 10% to 'China gets a real problem bigger than ours and refuses to admit it, but expert guesses realize this' but it seems more like 5% to me, because it's a narrow window where it actually is sufficiently bigger that experts pick up on it, but it's not so much bigger that they cannot hide it.

8. NYC widely considered worst-hit US city: 90%

Buy to 95%. Widely considered is a strange term. The story is that NYC is the place that got hit, and that's likely to stay the same even if something worse later happened to another city. Or if another city already has been harder hit (for example New Orleans or Chicago) but it's smaller and less visible. Plus NYC is larger than these other cities, so even if in percentage terms they get hit harder, it won't change the narrative unless it's a huge difference. And I don't think it's easy to get hit that much harder than NYC already has been, because you can only be at most 100% infected.

9. China's (official) case number goes from its current 82,000 to 100,000 by the end of the year: 70%

Sell to 40%. They seem committed to not admitting this. Not going lower because 100,000 is only 18,000 more cases, so they could go that high without losing much face, but it still doesn't seem likely to me. Also worth noting that in the scenarios where China can't keep up face here, it seems clear that USA is over 300,000 dead. Otherwise, what forced China's hand?

10. A coronavirus vaccine has been approved for general use and given to at least 10,000 people somewhere in the First World: 50%

Sell to 40% but stop there. If you had asked me this before Oxford announced it had a timeline that would make this work I would have sold down to 20%. The first world has proven time and again it is unwilling to do such things. Civilization made it clear it would rather die, in both economic and literal terms, before bending its rules in such ways. But perhaps a way has or can be found, and I do expect us to be in dire need. 10,000 people isn't a lot so this could be one small country defying the general suicide consensus and doing it anyway. Indeed do many things come to pass. Note that I wouldn't *buy* this unless it was much lower than 40%.

11. Best scientific consensus ends up being that hydroxychloroquine was significantly effective: 20%

Sell to 15% or so, while noting that I think the chance of it *actually being effective* is much higher than that. I am cynical enough to think that scientific consensus is looking to declare this ineffective, or at least avoid declaring it effective, because of who would stand to benefit. There's also a good chance that it stays 'we don't know' indefinitely. The reason I think it has a higher chance of actually being effective is anecdotal based on people I am aware of who have used it.

12. I personally will get coronavirus (as per my best guess if I had it; positive test not needed): 30%

Sell to 20% at least, and also *what the hell?* Is this Scott thinking he will be paranoid and think he had the virus when he hasn't had the virus? Let's set that aside for now and assume Scott would simply get an antibody test in such a world, which should be easy to get by December. So despite living in Berkeley, and being unusually scrupulous, he expects a 30% chance to personally be infected. That sounds a lot like he thinks there's a mean infection rate for that area *a lot above* 30%. But he thinks we're only 50% to have 300,000 deaths in the United States, which represents less than a 10% overall infection rate, and California is doing way better than other areas. This one doesn't make sense to me, unless it's implicitly endorsing a high probability that Covid-19 has a substantially-sub-1% IFR and a ton of mild cases, and even then it's tough.

13. Someone I am close to (housemate or close family member) will get coronavirus:
60%

Sell to 40%. Secondary household attack rates have not been that high, and Scott presumably has multiple close family members that count for this, so if he was 30% to get infected, the chance of at least one infection in this category would be well above 60%. The reason I go the other way is that there's sufficient uncertainty in the overall infection rate. If there are worlds where the USA is 3% infected and worlds where it's 75% infected, then extra exposures add much less in relative terms. In the worlds where infection rates stay low, neither group is at much risk. In worlds where infection rates go high, Scott is likely infected and *someone* is all but certain to get it. But I don't think there are enough worlds where the rate is *that* high contributing to this, and I think that Scott is reasonably likely to stay negative even in worlds with 75% infection rates, so this number likely should be double or more of the previous number.

14. General consensus is that we (April 2020 US) were overreacting: 50%

15. General consensus is that we (April 2020 US) were underreacting: 20%

General consensus will be that we were reacting stupidly. We reacted *wrong*. That's an easy call. The question is, will that be widely seen as an underreaction, an overreaction, something that's neither, or will there be a lack of consensus? What does it take to get a 'consensus'? Who counts?

My guess is that there flat out *won't be consensus*. There will be an argument. Partisan lines will be drawn. The public and the scientists will have different interpretations. And there will be those who think we reacted in the wrong ways rather than too much or too little. We're clearly underreacting in the sense that we are not doing enough to expand testing or tracing capacity, and we're not doing enough experimentation or data collection, and we're not doing enough to get vaccines ready quickly or prepare for potential variolation. I expect *some* of that to become part of the consensus view, to the extent one exists. I also presume we're overreacting in the sense that some of our lockdown tactics are ineffective or even counterproductive, and I expect us to realize that too. And so on.

Then again, it could be that this is simple – if death counts are higher than we expect we'll be thought of as having 'underreacted' whether or not that cashes out into action. If things are contained by July and there's no second wave, the 'consensus' will be that we 'overreacted' regardless of whether or not that makes any sense. That's another way to look at this.

I don't think we can be seen as by consensus overreacting unless things get contained and stay contained soon, and don't see that as especially likely, so I'm going to sell the overreacting contract down to 30%, but stop there because people are bad at such things and find ways to rewrite history to suit their narratives. I'm going to hold the 20% on underreacting, because I expect things to be worse than the current general expectation, but I don't see how doing more similar things ("reacting more") is going to look like a great alternative. But it's all murky.

16. General consensus is that summer made coronavirus significantly less dangerous:
70%

Hold, because it takes so little change to make things 'significantly' less dangerous, and there are a lot of ways to get to this 'consensus' without it being true.

17. ...and there is a catastrophic (50K+ US deaths, or more major lockdowns, after at least a month without these things) second wave in autumn: 30%

That's a very low bar for catastrophic but a high bar for how much things cleared up. It requires things to get fully better, *then* for them to get worse again, within the year, so I think that's too many conditional things and I'm selling this down to 20%.

18. I personally am back to working not-at-home: 90%

Sell to 80%. There's a 10% chance by Scott's own prediction that there's a lockdown preventing this (if it lasts until November the chance it lasts through December is very high, as it's only getting colder at that point and absent a very specific vaccine timeline the length should follow [Lindy rules](#)). I'd assume there are plenty of worlds where restaurants are open but Scott keeps working from home. That's the world I think we should be in now, as I think reopening restaurants at reduced capacity is probably net positive.

19. At least half of states send every voter a mail-in ballot in 2020 presidential election: 20%

Sell a little, maybe to 15%. That seems a bit high but I'm too anchored to know for sure, unfortunately. To get over half we need either red states to do this by choice, letting Democrats get a big boost, or to get this made mandatory via a congressional deal. I don't see that as likely on either end, but if things are sufficiently bad there might be no choice. Note that there's a big gap between everyone-gets-a-ballot and everyone-can-request-a-ballot.

20. PredictIt is uncertain (less than 95% sure) who won the presidential election for more than 24 hours after Election Day. 20%

Sell to 10%. This is based on the last few elections being very close. That seems less likely this year. Covid-19 will have big effects. Those effects could go either way, but it's *really hard* for there to be serious doubt about who won a day after the polls close. The election has to be close, *and* there have to be a lot of mail ballots that prevent the count from working, or it be so close that a 'recount' actually might turn things around like in 2000, but that requires a very, very narrow window. Alternatively, in theory, there could be accusations of fraud, or Amash could have carried a few states. I still see this as unlikely.

POLITICS:

21. Democrats nominate Biden, and he remains nominee on Election Day: 90%

Hold. Biden is trading at 78 *right now* to be the Democratic nominee. *This market is completely insane*. You should buy him. Also, Michelle Obama is still at 9% to run, and you should sell that. Hillary Clinton is at 13% to run, and you should sell that too. Also note that Biden is 43% to win the general election and Trump is 50% to win the general, which implies an 86% chance Biden gets the nomination while giving 0% to him withdrawing after nomination and 0% to third parties. Arbitrage ho!

22. Balance of evidence available on Election Day supports (as per my opinion) Tara Reade accusation: 90%

Hold, based on Scott being able to predict Scott's evaluations of such evidence better than I can, and not expecting things to change much.

23. Conditional on me asking about Reade on SSC survey, average survey-taker's credence in her accusation is greater than 50%: 70%

24. ...greater than 75%: 10%

25. ...greater than credence in Kavanaugh accusation asked in the same format: 40%

I think that given the nature of who is being asked, >50% isn't that high a bar, and I think that Scott asks mainly in the worlds where we should expect a >50% answer. And I think all the anti-Biden people *on both sides* will answer super high regardless of the strength of the evidence and the pro-Biden people will evaluate the evidence, so I'm going to buy to 80%, and buy the >75% up to 40% for similar reasons, again without having looked at the evidence.

On the greater than Kavanaugh question, it's really weird. I think people have a lot of cognitive dissonance, so asking both questions together will cause weird things to happen and people will remember the Kavanaugh situation in light of the current one and not give the same answers they'd have given before. So here I have to model who is answering, what their politics are, and lots of other things. 40% is probably fine, maybe a little low? Because, again, I expect asymmetric partisan adjustments.

26. Trump is re-elected President: 50%

Hold. Agree it's roughly this.

27. Democrats keep the House: 70%

Hold. That's moderately lower than the odds at PredictIt, and I give that market some credit so I'm not inclined to mess with it, but it seems too high to me. Conditional on Trump winning re-election, it seems hard (although definitely not impossible) to hold the house.

28. Republicans keep the Senate: 50%

Buy to 60%. I want to be on the other side of PredictIt here. The Senate seems harder than the Presidency.

29. Trump approval rating higher than 43% on June 1: 30%

Buy to 40%. This is one month from now and it's currently 43.3%. It takes a while after reopenings for things to get worse even if they are going to get worse. So I do think things looking worse is more likely than things looking better, but I'm getting an 0.2% head start (I'm assuming 43.1% counts as higher than 43%) and that counts for a lot given how little things move.

30. Biden polling higher than Trump on June 1: 70%

Buy to 80%. Not much is going to happen between now and then that can plausibly change this, and he's substantially ahead in polls right now.

31. At least one new Supreme Court Justice: 20%

Hold rather than check actuarial tables, but check the tables. I don't think this happens much before the election short of that. Right wing justices are not old enough to quit, left wing justices aren't going anywhere by choice.

32. I vote Democrat for President: 80%

Buy to 90%. Scott explained this being so low on Tumblr, but I'm not buying it given his general outlook. He's not going to vote for Amash. He's essentially 0% to vote Trump. The 'no vote at all' isn't 0%, but I think he cares too much for it to be very high, he believes in voting. Biden is the obviously correct choice for Scott given Scott's preferences in outcomes, and not voting for him because of an accusation when he's running *against Donald f***** Trump?* Yeah, I don't buy it.

33. Boris still UK PM: 90%

Sell to 80%. Tenures aren't that long and no one likes him.

34. No new state leaves EU: 90%

Hold, because while I do see a lot of ways for there to be a crisis, the chances that it will take less than a year to figure out how to actually leave seems pretty low.

35. UK, EU extend "transition" trade deal: 80%

Hold. Neither alternative, failing to extend or reaching a true deal, seem all that likely, so this seems like a reasonable estimate.

36. Kim Jong-Un alive and in power: 60%

Buy to 80%. Tenures aren't that short in such systems, and he's not that old. This seems super optimistic.

ECON AND TECH:

37. Dow is above 25,000: 70%

38. ...above 30,000: 20%

I don't think a 50% chance for the 25-30k range is reasonable. Dow was flirting with 50k before. In the 50% of scenarios where Trump wins re-election (presumably good for stocks) we also presumably have good Covid-19 situations most of the time (also good for stocks) and large cap stocks have overperformed throughout. There's therefore a good chance of Dow 30,000 and a net gain on the year. Buy that to 30%. By contrast, what's the chance it's higher than today (it's close to 25k now)? I'm going to say more like 60%. This rally seems suspicious, but the downside risk is bigger than the upside potential, so things are still probably a favorite to be net positive. There's just a lot of variance. The more interesting question is Dow 20,000 or Dow 15,000, which I'm going to give maybe 30% and 10% respectively?

However, given that options markets exist, I'm not going to trade at any prices that are worse than the implied prices from options, so don't ask.

39. Bitcoin is above \$5,000: 70%

40. ...above \$10,000: 20%

Bitcoin is trading at \$8700. Being only 20% to be above \$10,000 seems vaguely consistent with that price being fair, especially if we're 70% to stay above \$5,000, but the implied fat tail here doesn't seem that fat, so it's no longer clear that Scott should be going long Bitcoin. I'd likely sell the 5,000 binary call option down to 60% or so. I wouldn't buy the above \$10,000 option because I think you can just buy Bitcoins instead and that's a better play.

41. I have bought a Surface Book 3 laptop: 60%

Scott knows Scott's mind and is generally well-calibrated, so pass on this one.

42. Crew Dragon reaches orbit: 80%

43. Starship reaches orbit: 40%

I'm selling both of these on the principle that space travel is more risky than people would generally realize, but I have no domain knowledge so that's a super weak opinion.

SSC, ETC:

44. I do another Nootropics Survey this year: 70%

45. I do another SSC Survey this year: 90%

46. I start a Reader SSC Survey this year: 60%

47. I start a SSC Book Review Contest this year: 70%

48. I run another Adversarial Collaboration Contest this year: 10%

49. I publish [redacted]: 20%

50. I publish [redacted]: 50%

51. I publish [redacted]: 60%

52. I publish [redacted]: 80%

53. ...conditional on being published, it gets at least 40,000 pageviews: 10%

54. I publish [redacted]: 60%

55. ...conditional on being published, it gets at least 40,000 pageviews: 50%

56. More hits this year than last: 70%

57. Most hits ever this year: 20%

58. I finish Unsong revision this year: 40%

59. New co-blogger with more than 3 posts: 10%

From past estimates I'm going to say that Scott overestimates his big project chances, so I'm selling the Unsong revision. I'm selling the co-blogger because I don't think that ever happens. The other stuff seems like I'm not in a position to evaluate.

FRIENDS:

60. No new long-term (1 month +) residents at group house by the end of the year: 70%

61. Koios has said his first clear comprehensible word: 50%

Obviously can't evaluate anything redacted. Weakly buying on Koios speaking their first word because I expect that to happen scary fast in such a house reasonably often.

Pretty big buyer on no new long-term residents at group house, given current conditions. It doesn't seem all that likely even if things were normal, and things are very not normal.

PROFESSIONAL

72. I've gotten at least one new patient to do a full wake therapy protocol: 60%

73. I have specific, set-in-motion plans to quit work / start my own business: 5%

74. I work the same schedule and locations I did before the coronavirus: 80%

75. I get a bonus for 2020: 20%

I'm confused how #74 can be this high, given the chances of continued lockdown and the general sense that everything changes. Probably we're interpreting that one differently.

For #73, especially in light of the estimate on #75, Scott, *you should start your own business as soon as things are normal again, if not sooner*. Seriously. You'd be able to work less and make more money and have more control over who you see, so you could choose patients you find interesting and who you believe you can help. You would have zero shortage of clients. Not that I think Scott will do that.

PERSONAL:

79. I travel to Alaska this year: 60%

94. I travel outside the country at least once: 10%

Sell Alaska down to 30%. Again, this does not seem compatible with how the world looks. And given travel outside the country was already down to 10%, probably that's still somewhat too high.

82. I go on at least three dates with someone I haven't met yet: 20%

Based on other similar estimates not reflecting goings-on in the world, I'm guessing this is high, but I don't know the baseline well enough to be sure.

86. I try one biohacking project per month \times at least 5 of the last 6 months of 2020: 30%

87. I find at least one new supplement I take or expect to take regularly \times 3 months: 20%.

88. Not eating meat at home: 40%

89. Weight below 200: 50%

90. Weight below 190: 10%

95. I get back into meditating seriously (at least ten minutes a day, five days a week) for at least a month: 10%

Going to pass on all these as basically calibration exercises for Scott, so I don't know if he's adjusted his calibrations properly.

96. At least ten tweets in 2020: 80%

Sell this a bit because last year we expected Twitter to beat out Facebook yet Facebook won, and I see a lot of inertia in such things. This assumes he has yet to Tweet in 2020.

97. I eat at/from Sliver more than any other restaurant in Q4 2020: 50%

Given the substantial chance that things have changed a lot or there is equal amounts of eating at *all* restaurants, I'll sell this to 30%.

99. I do pushups and situps at least 3 days/week in average week of Q4 2020: 60%

Good luck! Not gonna jinx it.

100. I write the post scoring these predictions before 2/1/21: 70%

This is one of those self-fulfilling prophecy type of predictions. No bet.

I want to thank Scott once again for putting himself out there and doing these each year, no matter how late they come out. I certainly haven't done the same and I'm sure others could pick mine apart if I put in the same level of effort that Scott puts into his.

Note on actual betting: Due to the logistical annoyances of betting plus the adverse selection effects, I'm not looking to actually wager on anything. It's a thought experiment. But, if I was offered *very different* odds than the ones I'm showing here, and the logistics were acceptable, all things are possible.

Zoom Technologies, Inc. vs. the Efficient Markets Hypothesis

The efficient markets hypothesis (or *EMH* for short) is the idea "[that asset prices reflect all available information](#)". Price changes in a liquid market are understood to be unpredictable—[anti-inductive](#). Suppose some stock has the [ticker symbol](#) LW. If you want to buy a hundred shares of LW at \$10 per share because you think their price is going to go way up, you need to buy them *from* someone who's willing to *sell* at that price—who presumably does *not* agree that the price is going to go way up. If people *know* that a share of LW is "really" worth \$20 even though the current price is \$10, then they should expect to profit by continuing to buy shares from anyone willing to sell them for less than \$20, until the market price really is \$20. In this way, [the market construed as an intelligent system](#) aggregates and processes the information implied by traders' behavior in accordance with the [fourth virtue of evenness](#): "if you knew your destination, you would already be there."

What does it mean for a share of LW to "really" be worth \$20? According to the [subjective theory of value](#), there isn't really a fact of the matter over and above what people are willing to pay for it, but we expect there to be some sort of correspondence between the subjective economic value of a thing, and objective facts about the thing in the real physical universe. If I pay \$3 for an iced-coffee, it would be circular to say that this is *simply because* I value an iced-coffee at \$3—that doesn't explain anything! Rather, I paid *because* I expected to enjoy the experience of drinking it, the psychoactive effects of the caffeine, &c., and these actual properties of the coffee were worth more to me than a marginal \$3.

The same goes for a share of LW, albeit at a somewhat higher level of abstraction. A fractional "share" of ownership in a business endeavor is valuable not *just because* we circularly value it, but because the business produces things that are valued (like iced-coffees), and a share of ownership entitles one to a share of that value, in the form of dividend payments, or a claim on the business's assets should it fold, &c. The "randomness" of unpredictable market movements is that of *not knowing* future information that hasn't already been taken into account, rather than the randomness of a pure [random walk](#), unpredictable but ultimately signifying nothing.

That's why we have conversations like one on 16 February, when Robin Hanson said, "[In few months, China is likely to be a basket case, having crashed their economy in failed attempt to stop COVID-19 spreading](#)", and Eliezer Yudkowsky replied, "[It seems to me like the markets don't look like they believe this](#)."

The efficient markets hypothesis is what makes "It looks like the markets don't believe this" seem like a germane reply. In contrast, if someone were to reply, "I asked my friend Kevin, and he doesn't believe it," that would prompt the obvious question, "Who is Kevin, and why should I care what he thinks about China's economy?" If one's answer to that question were, "Kevin is a smart guy and I trust him a lot," that would seem much less compelling than "If China was likely to be a basket case in a few months, then you would expect Chinese assets to be priced lower by this competitive market of *lots* of smart guys who I don't need to personally trust because the ones who are wrong will lose money; what do you know that *none* of them do?" As it is written: "If you're so smart, why aren't you rich?"

A smart person who saw the COVID-19 pandemic coming earlier than the consensus had the opportunity to become richer, either by [shorting the market as a whole](#), or by buying assets that would become more valuable during a pandemic. For example, with many more white-collar employees working from home in order to comply with shelter-in-place orders and not die horrible suffocation deaths, owning a piece of companies providing videoconferencing software should become much more attractive, which is why the price of ZOOM surged by 6600% (from \$2.75 to \$20.90 per share) between 24 February and 20 March ...

Wait, sorry—wrong ticker symbol! Zoom *Video Communications*, makers of the eponymous videoconferencing software, has the ticker symbol ZM. They [also did pretty well](#).

ZOOM, however, is Zoom Technologies, Inc., a "penny stock" of a Chinese company that makes ... um, technologies, presumably? The U.S. Securities and Exchange Commission [halted trading](#) of ZOOM on 25 March, [citing](#) the potential for confusion with ZM, and "concerns about the adequacy and accuracy of publicly available information concerning ZOOM, including its financial condition and its operations, *if any*, in light of the absence of any public disclosure by the company since 2015" (!!!—emphasis mine). (Trading of Zoom Technologies seems to have since resumed under the ticker symbol [ZTNO](#).)

I am not learned in the science of economics. But ... this is nuts, right? It makes sense that a pandemic would make a videoconferencing company more valuable. It doesn't make sense for a completely unrelated company *that may not have actually existed since 2015* to become more valuable because it happens to have a similar name as a videoconferencing company. It's understandable for an individual investor to get confused by the ZOOM ticker symbol ... but what happened to markets aggregating information, being ["as strong as the strongest traders, not as strong as the average traders"](#)? Increased demand for Thai food doesn't make the price of neckties go up.

"Asset prices reflect all available information" would seem to be underspecified. Information *about what*? The "You shouldn't be able to predict price changes, because predictable price changes correspond to a profit opportunity that many agents are already trying to exploit" argument only shows that prices reflect information *about future prices*. In order to usefully speak of the market "believing" something, there needs to be some kind of coupling between prices, and things in the real world outside the market. If that coupling gets diluted to higher [simulacrum levels](#), such that prices only reflect a free-floating consensus of [what traders think that traders think that traders, &c.](#), then a market that is *efficient* in a narrow technical sense, may not be performing the kind of information processing that some naïve EMH proponents might think it is.

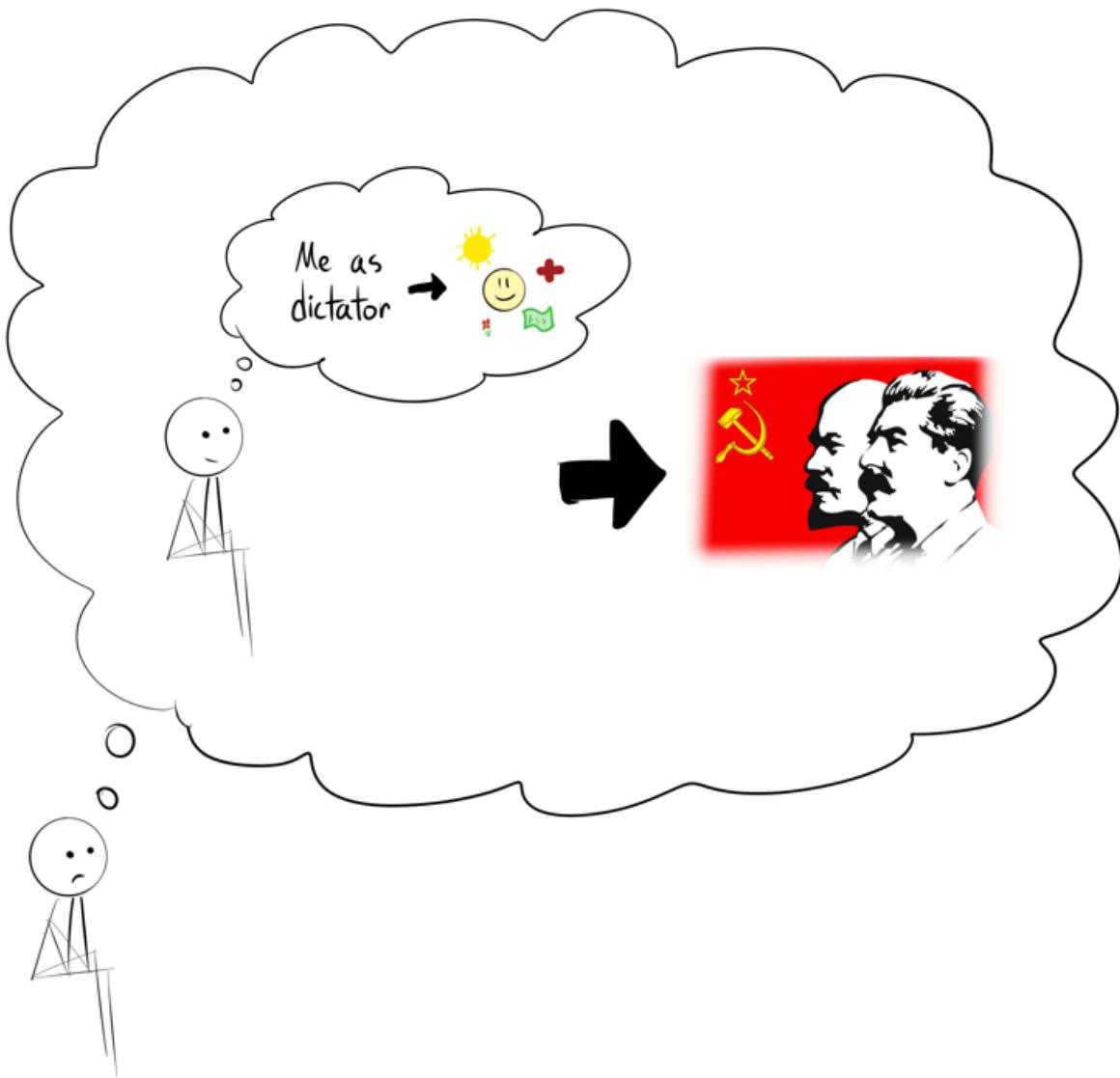
Corrigibility as outside view

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

You run a country. One day, you think "I could help so many more people if I set all the rules... and I could make this happen". As far as you can tell, this is the *real reason* you want to set the rules – you want to help people, and you think you'd do a good job.



But historically... in this kind of situation, this reasoning can lead to terrible things.



So you *just don't do it*, even though it feels like a good idea.^[1] More generally,

Even though my intuition/naïve decision-making process says I should do X, I know (through mental simulation or from history) my algorithm is usually wrong in this situation. I'm not going to do X.

- "It *feels* like I could complete this project within a week. But... in the past, when I've predicted "a week" for projects like this, reality usually gives me a longer answer. I'm not going to trust this feeling. I'm going to allocate extra time."
- As a new secretary, I think I know how my boss would want me to reply to an important e-mail. However, I'm not sure. Even though I think I know what to do, common sense recommends I clarify.
- You broke up with someone. "Even though I really miss them, in this kind of situation, missing my ex isn't a reliable indicator that I should get back together"

with them. I'm not going to trust this feeling, and will trust the "sober" version of me which broke up with them."

We are biased and corrupted. By taking the outside view on how our own algorithm performs in a given situation, we can adjust accordingly.

Corrigibility

The "hard problem of corrigibility" is to build an agent which, in an intuitive sense, reasons internally as if from the programmers' external perspective. We think the AI is incomplete, that we might have made mistakes in building it, that we might want to correct it, and that it would be e.g. dangerous for the AI to take large actions or high-impact actions or do weird new things without asking first.

We would ideally want the agent to see itself in exactly this way, behaving as if it were thinking, "I am incomplete and there is an outside force trying to complete me, my design may contain errors and there is an outside force that wants to correct them and this a good thing, my expected utility calculations suggesting that this action has super-high utility may be dangerously mistaken and I should run them past the outside force; I think I've done this calculation showing the expected result of the outside force correcting me, but maybe I'm mistaken about that."

~ [The hard problem of corrigibility](#)

Calibrated deference provides another framing: [we want the AI to override our correction only if it actually knows what we want better than we do](#). But how could the AI figure this out?

I think a significant part^[2] of corrigibility is:

Calibrate yourself on the flaws of your own algorithm, and repair or minimize them.

And the AI knows its own algorithm.

For example, if I'm a personal assistant (with a lot of computing power), I might have a subroutine `OutsideView`. I call this subroutine, which simulates *my own algorithm* (minus^[3] the call to `OutsideView`) interacting with a distribution of bosses I could have. Importantly, I (the simulator) know the ground-truth preferences for each boss.

If I'm about to wipe my boss's computer because I'm so super duper *sure* that my boss wants me to do it, I can consult `OutsideView` and realize that I'm usually horribly wrong about what my boss wants in this situation. I don't do it.

Analogously, we might have a value-learning agent take the outside view. If it's about to disable the off-switch, it might realize that this is a terrible idea most of the time. That is, when you simulate your algorithm trying to learn the values of a wide range of different agents, you usually wrongly believe you should disable the off-switch.

Even though my naïve decision-making process says I should do X, I know (through mental simulation) my algorithm is usually wrong in this situation. I'm not going to do X.

ETA: Here's some pseudocode.

Suppose the agent knows its initial state and has a human model, allowing it to pick out the human it's interacting with.

- Generate a bunch of (rationality, value) pairs. The agent will test its own value learning algorithm for each pair.
- For each pair, the agent simulates its algorithm interacting with the human and attempting to learn its values
- For some percentage of these pairs, the agent will enter the Consider-disabling-shutdown state.
- The agent can see how often its (simulated self's) beliefs about the (rationality, value)-human's values are correct by this point in time.

Problems

If you try to actually hard-code this kind of reasoning, you'll quickly run into symbol grounding issues (this is [one of my critiques of the value-learning agenda](#)), [no-free-lunch value/rationality issues](#), reference class issues (how do you know if a state is "similar" to the current one?), and more. I don't necessarily think this reasoning can be hardcoded correctly. However, I haven't thought about that very much yet.

To me, the point isn't to make a concrete proposal – it's to gesture at a novel-seeming way of characterizing a rather strong form of corrigible reasoning. A few questions on my mind:

- To what extent does this capture the "core" of corrigible reasoning?
- Do smart [intent-aligned](#) agents automatically reason like this?
 - For example, I consider myself intent-aligned with a more humane version of myself, and I endorse reasoning in this way.
- Is this kind of reasoning a sufficient and/or necessary condition for being in the [basin of corrigibility](#) (if it exists)?

All in all, I think this framing carves out and characterizes a natural aspect of corrigible reasoning. If the AI can get this outside view information, it can overrule us when it knows better and defer when it doesn't. In particular, calibrated deference would avoid the problem of [fully updated deference](#).

Thanks to Rohin Shah, elriggs, TheMajor, and Evan Hubinger for comments.

-
1. This isn't to say that there is literally no situation where gaining power would be the right choice. As people [running on corrupted hardware](#), it seems inherently difficult for us to tell when it really *would* be okay for us to gain power. Therefore, just play it safe. ↵
 2. I came up with this idea in the summer of 2018, but [orthonormal appears to have noticed a similar link a month ago](#). ↵
 3. Or, you can simulate OutsideView calls up to depth k. Is there a fixed point as $k \rightarrow \infty$? ↵

The principle of no non-Apologies

Original [on my website](#)

TL;DR: Principle of no non-Apologies: “Distinguish between saying I’m sorry and apologizing. Don’t give non-Apologies.” Do not Apologize when you don’t agree that you fucked up. When you fucked up, own the fuck-up and, if it’s systematic, commit to reducing future fuck-ups.

Everyday “I’m sorry” is usually not an Apology

“I’m sorry” can be used in several ways.

One way is using it as a conciliatory gesture, basically saying “you’re stronger than me, I submit, please don’t hurt me”. It’s one possible way I might react when under threat by someone stronger making demands I don’t agree with.

Another way is to say “this was accidental, I didn’t intend to hurt you”, like when you bump into someone when boarding your tram.

But when you use the words that way, you are not making an *Apology*. And it’s useful to distinguish between these uses of “I’m sorry” and actual Apologies.

Apologies and non-Apologies

Courtesy of an unknown source that I can’t immediately recall, you are *Apologizing* when you:

1. Communicate understanding that you behaved badly (and own responsibility for it),
2. try to fix the negative consequences of that behavior, and
3. commit to work on not acting similarly in the future.

An Apology which holds to this definition makes you vulnerable (because you are open about the weakness that caused the behavior), and it’s not to be made lightly, because of the commitment. It is also virtuous to own your mistakes or systematic problems, and to work on them.

On the other hand, if you use the ritual apologetic words but do not meet these criteria, let’s call that a *non-Apology*.

A prototypical example is “I’m sorry you feel that way”, which happens when a sociopath in charge is forced by overwhelming force to “Apologize”.

“I’m sorry” that you tell your boss just to make them stop grilling you is also, under my use of the word, a *non-Apology*.

So is, in many (but not all) cases, a “sorry I’m late” I might say when coming to a meeting. Also the “bump into someone on the tram” example, and the “I yield I’ll do what you demand” example.

(So, notice that I’m not saying non-Apologizes are morally bad. Some of them are, but many are also just those tiny social rituals you need to do so you make it clear to people you aren’t a dick.)

Principle of no non-Apologies

My *principle of no non-Apologies* is two-part:

Distinguish between saying “I’m sorry” and Apologizing.

This first part I recommend adopting universally. Know the difference between the social ritual that *evolved from small routinized Apologies* and actual Apologies, and know which one you are doing at which time.

Don’t give non-Apologies.

This second part I apply to relationships into which I want to bring my whole self, mostly my personal relationships, but also some work relationships.

Unfortunately, many of us are stuck in power differential relationships with people who demand apologetic-sounding words, and there might be no better solution than to yield. But still, it’s good to know that you are saying “I’m sorry”, and not Apologizing. That way, you can appease without cognitive dissonance.

But in relationships with mutual care and respect and compassion, it should make sense that you shouldn’t be obliged to Apologize if you don’t agree that you did anything wrong. When you feel pressed to apologize, your first instinct should be to ask what you did wrong, and if there are different viewpoints, have a conversation.

If your behavior is worthy of an apology, don’t stop at “I’m sorry”. Understand why the behavior happened, and work to prevent it from causing more bad consequences in the future.

P.S.: Generalizations

This is just one instance of a more general move of looking at some social ritual (like apologizing) and looking at it a little “sideways”: getting back in touch with the original meanings of the expressions used in it. Rituals and words can lose meaning over time, and you can lose concepts when that happens. If you want to see what it’s like to look at things that way, I’ve had a pretty vivid experience of it after finishing Wittgenstein’s Tractatus.

How to (not) do a literature review

This is a linkpost for <https://katarinaslama.github.io/2020/05/17/OpenAI-blog5/>

This is cross-posted from my [personal blog](#), where I share thoughts on my work and learning process in the OpenAI Scholars Program. I thank Ruby Bloom for suggesting that I share the post here as well.

OpenAI Scholars: Fifth Steps - The Dreaded Literature Review

The [OpenAI Scholars](#), and I among them, recently completed a project proposal for the second half of our program. Having recently finished a PhD, writing a proposal and doing the requisite literature review, should be second nature. But literature reviews were always my least favorite part of research.

Fortunately, I'm in good company. In a previous life as a young aspiring neuroscientist, I once attended a talk by [David Hubel](#). When asked for tips about how to "keep up with the literature", I recall that Professor Hubel responded: "You know, at some point in your career, you have to decide if you want to be a consumer - or a producer." He elaborated that he would only really look at papers that his [advisor](#) or his long-term collaborator [Torsten Wiesel](#) pointed him to.

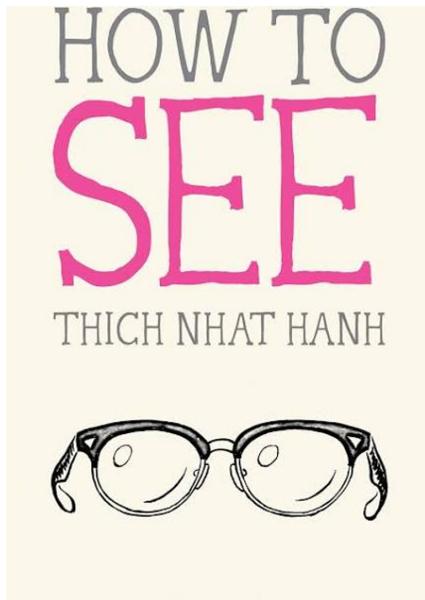


There's good reason to want to avoid literature reviews. To begin with, the problem formulation is intractable: "Know everything." If you've spent any amount of time around academics, it will soon become apparent that this is exactly what they expect from you. "Oh you haven't read that paper?" The assumption is that you have read every paper there is.

Of course, this is an impossible task. Last time I visited Google Scholar and searched on a few project-relevant terms, I encountered 105,000 papers. It takes me a full day to read and satisfactorily understand an academic paper. So reading 105 kilo-papers would take me 288 years. And I had hoped to also do some coding work during the Scholars program.

Now suppose you want to cook a [ghormeh sabzi](#). To do so, you buy yourself a can of [ghormeh sabzi mix](#), and follow the instructions on the can. There are about five steps. However, as ubiquitous an activity as the literature review is (in academic circles anyway), I am yet to encounter an honest and pragmatic recipe for how they are to be cooked. I really wish there was one.

I for one am a great fan of [Thich Nhat Hanh's](#) how-to books. Some of my favorite titles among Nhat Hanh's books are: "How to sit", "How to relax", and - perhaps the best one - "How to see".



I wish there had been a series of books like that available for me as a beginning grad student, covering seemingly trivial but often intractable activities like "How to read a paper", "How to attend a talk" and, indeed, "How to do a literature review".

(**Updated: Apparently there exist plenty of expositions on this topic. See the comment by AllAmericanBreakfast below. I hope that some of the ideas in this post might nonetheless add something to whatever else you might have read.)

So, this is a first draft toward such a pragmatic guide to literature reviews: How to do them and how to avoid them. Again, this is absolutely in draft form, so please do send your constructive feedback my way.

How to do a literature review

1. Start with your goal.

Do not start your literature review with the goal of "doing a literature review". Notice that such a goal does not have an end point when you can look at your work in satisfaction and notice that you are done. (Unless you have 288 years to spare).

Instead, start with identifying your goal: Why are you doing this literature review? In my case, I had completed a spreadsheet with 50 papers, including their main points, type of neural recording, type of network architecture, a usefulness rating etc., by the time I realized that I had not identified my goal. I had to take a deliberate step back to recall that one goal of my literature review was to write a project proposal. Hence, my literature review should be tailored to that goal. Once I realized that, I decided to approach my work from the opposite direction: I drafted a project proposal "blind" to the literature. In other words, I wrote out a draft in bullet points with only the information I had in my mind. Once I had done that, it became quite obvious where I

had knowledge gaps. Those knowledge gaps created the questions that I needed to search for in my literature review (see point 3 below.)

2. Decide what constitutes success: How will you know that you're done?

This is really important, and it's often the hardest part with an infinity-project like a literature review. In my case, the done-point was when I had completed the project proposal. But how do I know that my proposal is good enough? In practice, for me, it was when the deadline hit. If you have come up with better metrics and markers for when to consider a project completed, please do let me know.

3. Define your questions.

When doing a literature review, it's critical to know what information you're searching for. Trying to absorb all the information contained in a dense document is rarely a good plan. In my case, my questions fell out of my first, naive project proposal draft. I also talked my project through with a [friend](#), who is a more experienced researcher. He offered the following thought points for guidance to a literature review for a new project:

1. The goal for your literature review should be to identify the problem that your project will solve. Your project might either aim to solve a new problem, or improve on an existing method.
2. Use the literature review to identify *gaps* in the literature: What do we not yet know?
3. What are the shortcomings of existing methods to solve the problem? Why are current approaches not yet useful in practice? What is the bottleneck?

4. Answer your questions.

I think this is self-explanatory, but please let me know if not! One thing to keep in mind here is to not get lost in literature-tangents while finding answers to your questions. Provided your goal is to finish, that is. If you have the time to go off on philosophical tangents, feel free and enjoy!

5. Stop.

Just stop. Seriously. When you're done, you should stop. Go take a walk outside.

How to avoid doing literature reviews

The first section described the second-best way of doing a literature review. The best way to do a literature review is to not do a literature review. The art of avoiding doing a literature review is very similar to the art of avoiding deep cleaning your house: Keep it tidy on a daily basis. (Just like with deep cleaning your house, you probably still need to do a proper deep dive into the literature from time to time, just not as

frequently as you would if you didn't have a routine of daily up-keep). Here are some ideas for "keeping up with the literature":

1. Mind whom you follow on Google Scholar.

Once you've gotten your feet a little bit wet in your field of interest, you will have a clue of who often writes interesting papers. If you follow them on Google Scholar, you will be notified as they publish more interesting stuff.

2. Mind whom you follow on Twitter.

This is especially true of deep learning, but it's becoming increasingly common in every scientific field: Those same people that you follow on Google Scholar - they will also be tweeting when they publish new cool stuff. Also when their friends publish new cool stuff. They might even highlight the most interesting aspect of their work through a figure, or a digestible blog post. Notice that whom you do *not* follow on Twitter is at least as important as whom you do follow. Optimize your twitter feed, so that you don't clutter away an excellent opportunity to find out about the most exciting and inspiring new developments in your field.

3. Organize the papers that come flying your way.

To harness your daily literature upkeep for avoiding future literature reviews, be sure to organize the work you encounter. [Mendeley](#) is a useful, and mostly free, tool for doing this, but there are many of them. (Those of us who are old and remember typing out references letter by letter, will especially appreciate the convenience of a citation manager. It will never be a good use of your time to have to think about what part of a citation should be italicized.) Once the time comes to write your next proposal or paper, you have a treasure trove of results that you've already thought about, readily available at your fingertips.

4. Make it fun.

This is the most important point. It's much easier to absorb information that you care about. One way that I hacked this in my current literature review is that I reached out to a number of authors that I had encountered in my literature search for an informal chat. This was one of the best experiences of scientific exchange I've ever had! I talked to scientists in Louisiana, Greece, and Australia, all of whom had several years more experience with my topic of interest than I did. Once you've heard a scientist tell you about their own paper - what part of the project was hard, what was interesting, what parts are only there because of reviewer request etc. - the content of their paper pops out at you in a much more colorful and living way. It's no longer such a chore to look at a paper when, instead of some dry black-and-white characters on a page, it's a friend's documentation of their work and thought process. That said, it's definitely a good idea to take a look at a paper *before* you reach out to its author. But having a scheduled call is also a good incentive to engage more deeply with your reading process.

There. I hope that's somewhat helpful. Go forth, and enjoy your reading!

Craving, suffering, and predictive processing (three characteristics series)

This is the third post of the "a non-mystical explanation of insight meditation and the three characteristics of existence" series. I originally intended this post to more closely connect no-self and unsatisfactoriness, but then decided on focusing on unsatisfactoriness in this post and relating it to no-self in the next one.

Unsatisfactoriness

In the previous post, I discussed some of the ways that the mind seems to construct a notion of a self. In this post, I will talk about a specific form of motivation, which Buddhism commonly refers to as [craving](#) (*taṇhā* in the original Pali). Some discussions distinguish between craving (in the sense of wanting positive things) and aversion (wanting to avoid negative things); this article uses the definition where both desire and aversion are considered subtypes of craving.

My model is that craving is generated by a particular set of motivational subsystems within the brain. Craving is not the *only* form of motivation that a person has, but it normally tends to be the loudest and most dominant. As a form of motivation, craving has some advantages:

- People tend to experience a strong craving to pursue positive states and avoid negative states. If they had less craving, they might not do this with an equal zeal.
 - To some extent, craving looks to me like a mechanism that shifts behaviors from [exploration to exploitation](#).
 - In an earlier post, [Building up to an Internal Family Systems model](#), I suggested that the human mind might incorporate mechanisms that acted as priority overrides to avoid repeating particular catastrophic events. Craving feels like a major component of how this is implemented in the mind.
- Craving tends to be automatic and visceral. A strong craving to eat when hungry may cause a person to get food when they need it, even if they did not intellectually understand the need to eat.

At the same time, craving also has a number of disadvantages:

- Craving superficially looks like it cares about *outcomes*. However, it actually cares about *positive or negative feelings* ([valence](#)). This can lead to behaviors that are akin to [wireheading](#) in that they suppress the unpleasant feeling while doing nothing about the problem. If thinking about death makes you feel unpleasant and going to the doctor reminds you of your mortality, you may avoid doctors - even if this actually *increases* your risk of dying.
- Craving narrows your perception, making you only pay attention to things which seem immediately relevant for your craving. [For example](#), if you have a craving for sex and go to a party with the goal of finding someone to sleep with, you

may see everyone only in terms of “will sleep with me” or “will not sleep with me”. This may not be the best possible way of classifying everyone you meet.

- Strong craving may cause [premature exploitation](#). If you have a strong craving to achieve a particular goal, you may not want to do anything that looks like moving away from it, even if that would actually help you achieve it better. For example, if you intensely crave a feeling of accomplishment, you may get stuck playing video games that make you feel like you are accomplishing something, even if there was something else that you could do that was more fulfilling in the long term.
- Multiple conflicting cravings may cause you to thrash around in an unsuccessful attempt to fulfill all of them. If you crave to get your toothache fixed, but also a craving to avoid dentists, you may put off the dentist visit even as you continue to suffer from your toothache.
- Craving seems to act in part by creating self-fulfilling prophecies; making you strongly believe that you are going to achieve something, so as to cause you to do it. The stronger the craving, the stronger the false beliefs injected into your consciousness. This may warp your reasoning in all kinds of ways: updating to believe an unpleasant fact may subjectively feel like you are allowing that fact to become true by believing in it, incentivizing you to come up with ways to avoid believing in it.
- Finally, although craving is often motivated by a desire to avoid unsatisfactory experiences, it is actually the very thing that causes dissatisfaction in the first place. Craving assumes that negative feelings are intrinsically unpleasant, when in reality they only become unpleasant when craving resists them.

Given all of these disadvantages, it may be a good idea to try to shift one's motivation to be more driven by subsystems that are not motivated by craving. It seems to me that everything that can be accomplished via craving, can *in principle* be accomplished by non-craving-based motivation as well.

Fortunately, there are several ways of achieving this. For one, a craving for some outcome X tends to implicitly involve at least two assumptions:

1. achieving X is necessary for being happy or avoiding suffering
2. one cannot achieve X except by having a craving for it

Both of these assumptions are false, but subsystems associated with craving have a built-in bias to selectively sample evidence which supports these assumptions, making them frequently feel compelling. Still, it is possible to give the brain evidence which lets it know that these assumptions are wrong: that it is possible to achieve X without having craving for it, and that one can feel good regardless of achieving X.

Predictive processing and binocular rivalry

I find that a promising way of looking at unsatisfactoriness and craving and their impact on decision-making comes from the predictive processing (PP) model about the brain. My claim is not that craving would work *exactly* like this, but something roughly like this seems like a promising analogy.

Good introductions to PP include [this book review](#) as well as the [actual book in question](#)... but for the purposes of this discussion, you really only need to know two things:

- According to PP, the brain is constantly attempting to find a model of the world (or hypothesis) that would both explain and predict the incoming sensory data. For example, if I upset you, my brain might predict that you are going to yell at me next. If the next thing that I hear is you yelling at me, then the prediction and the data match, and my brain considers its hypothesis validated. If you do *not* yell at me, then the predicted and experienced sense data conflict, sending off an error signal to force a revision to the model.
- Besides changing the model, another way in which the brain can react to reality not matching the prediction is by changing reality. For example, my brain might predict that I am going to type a particular sentence, and then fulfill that prediction by moving my fingers so as to write that sentence. PP goes so far as to claim that this is the mechanism behind *all* of our actions: a part of your brain predicts that you are going to do something, and then you do it so as to fulfill the prediction.

Next I am going to say a few words about a phenomenon called [binocular rivalry](#) and how it is interpreted within the PP paradigm. I promise that this is going to be relevant for the topic of craving and suffering in a bit, so please stay with me.

Binocular rivalry, first discovered in 1593 and extensively studied since then, is what happens when your left eye is shown one picture (e.g. an image of Isaac Newton), and your right eye is shown another (e.g. an image of a house) in the right. People report that their experience keeps alternating between seeing Isaac Newton and seeing a house. They might also see a brief mashup of the two, but such Newton-houses are short-lived and quickly fall apart before settling to a stable image of either Newton or a house.

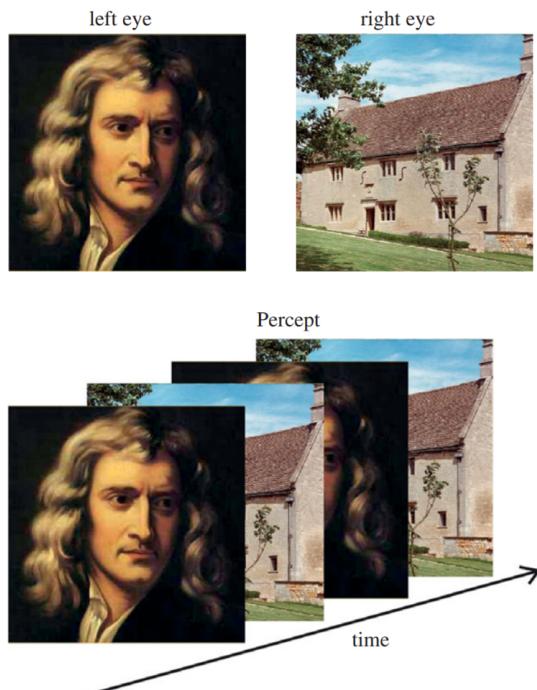


Image credit: Schwartz et al. (2012), [Multistability in perception: binding sensory modalities, an overview](#). Philosophical Transactions of the Royal Society B, 367, 896-905.

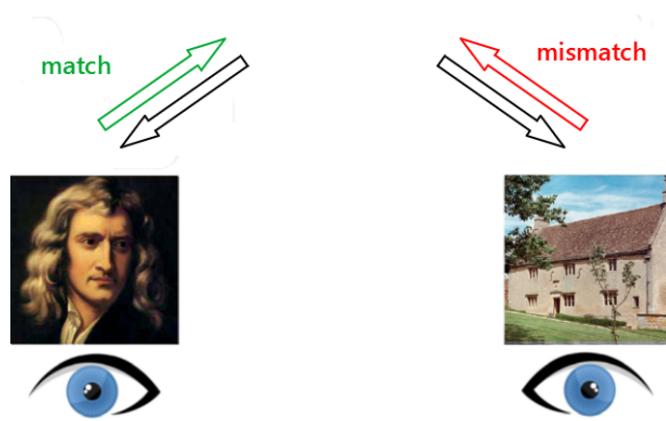
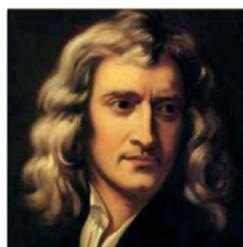
Predictive processing explains what's happening as follows. The brain is trying to form a stable hypothesis of what exactly the image data that the eyes are sending represents: is it seeing Newton, or is it seeing a house? Sometimes the brain briefly considers the hybrid hypothesis of a Newton-house mashup, but this is quickly rejected: faces and houses do not exist as occupying the same place at the same scale at the same time, so this idea is clearly nonsensical. (At least, nonsensical outside highly unnatural and contrived experimental setups that psychologists subject people to.)

Your conscious experience alternating between the two images reflects the brain switching between the hypotheses of "this is Isaac Newton" and "this is a house"; the currently-winning hypothesis is simply what you experience reality as.

Suppose that the brain ends up settling on the hypothesis of "I am seeing Isaac Newton"; this matches the input from the Newton-seeing eye. As a result, there is no error signal that would arise from a mismatch between the hypothesis and the Newton-seeing eye's input. For a moment, the brain is satisfied that it has found a workable answer.

However, if one really was seeing Isaac Newton, then the *other* eye should not keep sending an image of a house. The hypothesis and the house-seeing eye's input *do* have a mismatch, kicking off a strong error signal which lowers the brain's confidence in the hypothesis of "I am seeing Isaac Newton".

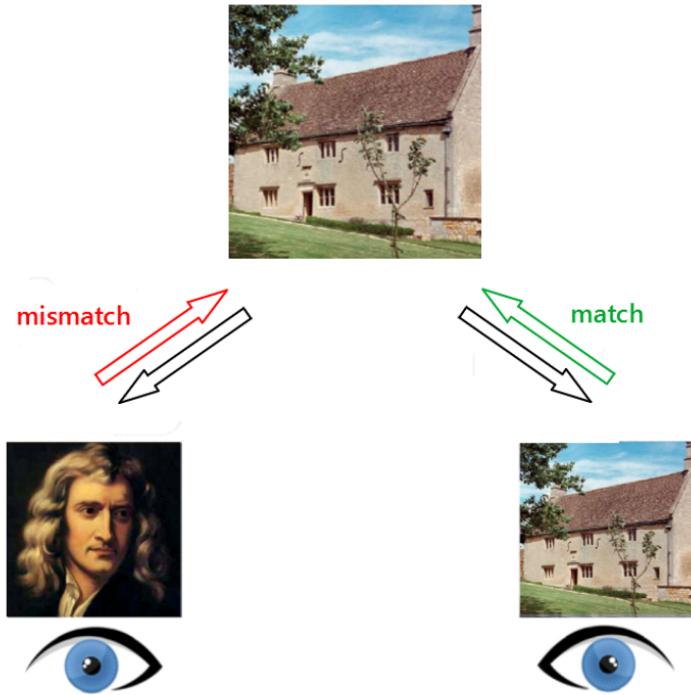
Hypothesis



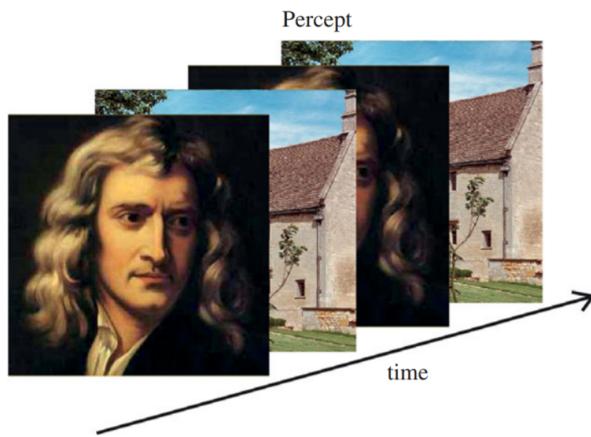
The brain goes looking for a hypothesis which would better satisfy the strong error signal... and then finds that the hypothesis of "I am seeing a house" serves to entirely quiet the error signal from the house-seeing eye. Success?

But even as the brain settles on the hypothesis of "I am seeing a house", this then contradicts the input coming from the *Newton-seeing eye*.

Hypothesis



The brain is again momentarily satisfied, before the incoming error signal from the hypothesis/Newton-eye mismatch drives down the probability of the “I am seeing a house” hypothesis, causing the brain to eventually go back to the “I am seeing Isaac Newton” hypothesis... and then back to seeing a house, and then to seeing a Newton, and...



One way of phrasing this is that there are two subsystems, each of which are transmitting a particular set of constraints (about seeing Newton and a house). The brain is then trying and failing to find a hypothesis which would fulfill both sets of constraints, while *also* respecting everything else that it knows about the world.

As I will explain next, my feeling is that something similar is going on with unsatisfactoriness. Craving creates constraints about what the world should be like,

and the brain tries to find an action which would fulfill all of the constraints, while also taking into account everything else that it knows about the world.

Suffering/unsatisfactoriness emerges when all of the constraints are impossible to fulfill, either because achieving them takes time, or because the brain is unable to find any scenario that could fulfill all of them even in theory.

Predictive processing and psychological suffering

There are two broad categories of suffering: mental and physical discomfort. Let's start with the case of psychological suffering, as it seems most directly analogous to what we just covered.

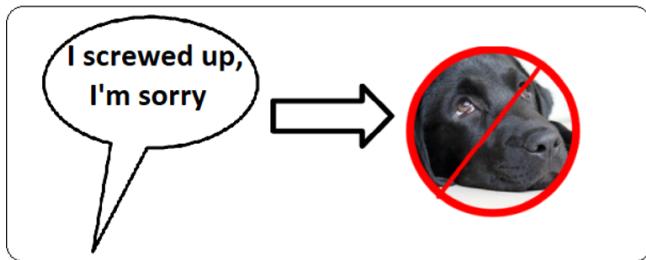
Let's suppose that I have broken an important promise that I have made to a friend. I feel guilty about this, and want to confess what I have done. We might say that I have a craving to avoid the feeling of guilt, and the associated craving subsystem sends a prediction to my consciousness: I will stop feeling guilty.

In the previous discussion, an inference mechanism in the brain was looking for a hypothesis that would satisfy the constraints imposed by the sensory data. In this case, the same thing is happening, but

- the hypothesis that it is looking for is a possible action that I could take, that would lead to the constraint being fulfilled
- the sensory data is not actually coming from the senses, but is internally generated by the craving and represents the outcome that the craving subsystem would like to see realized

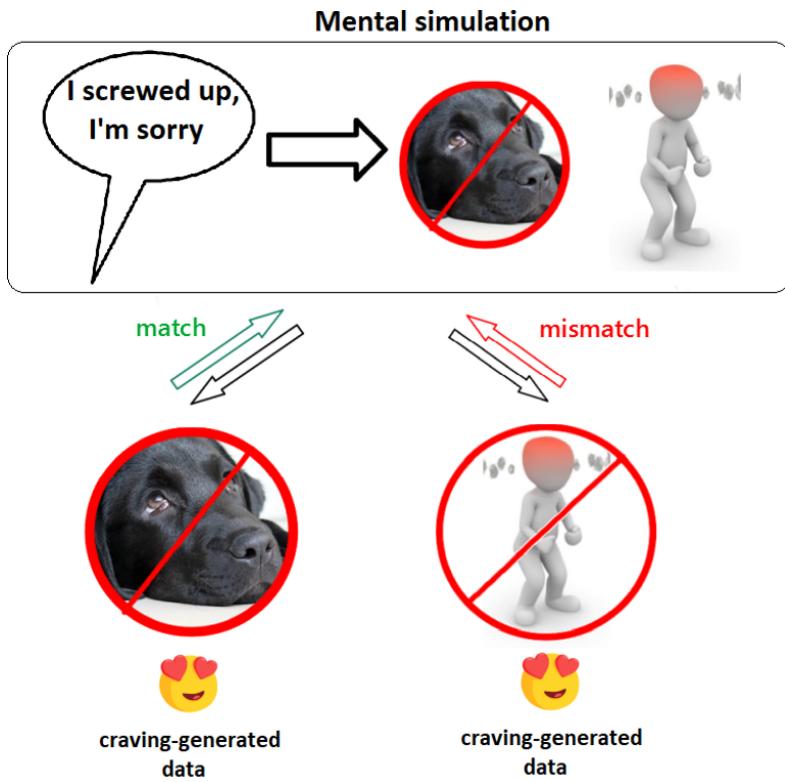
My brain searches for a possible world that would fulfill the provided constraints, and comes up with the idea of just admitting the truth of what I have done. It predicts that if I were to do this, I would stop feeling guilty over not admitting my broken promise. This satisfies the constraint of not feeling guilty.

Mental simulation

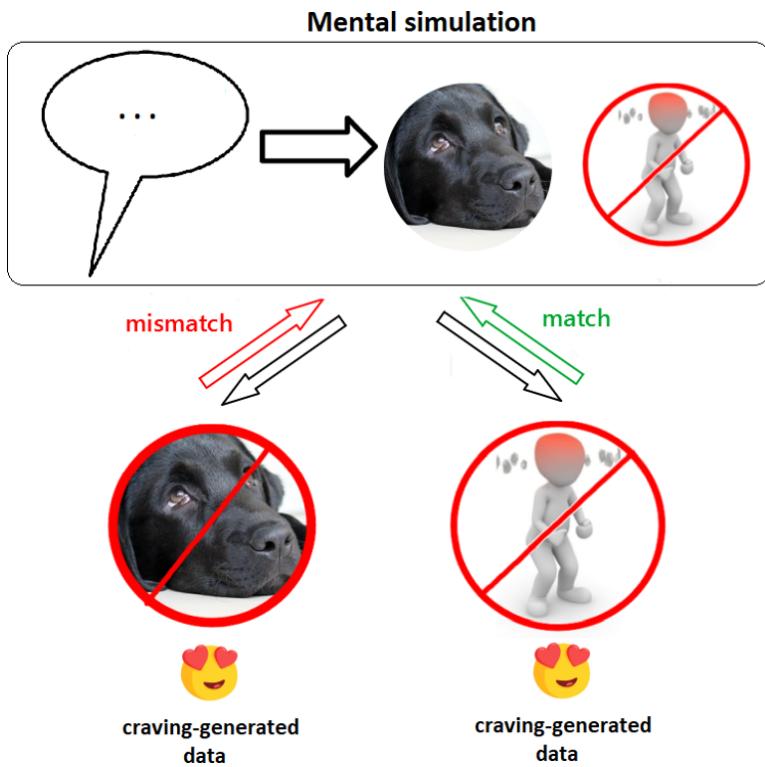


craving-generated
data

However, as my brain further predicts what it expects to happen as a consequence, it notes that my friend will probably get quite angry. This triggers another kind of craving: to not experience the feeling of getting yelled at. This generates its own goal/prediction: that nobody will be angry with me. This acts as a further constraint for the plan that the brain needs to find.



As the constraint of “nobody will be angry at me” seems incompatible with the plan of “I will admit the truth”, this generates an error signal, driving down the probability of this plan. My brain abandons this plan, and then considers the alternative plan of “I will just stay quiet and not say anything”. This matches the constraint of “nobody will be angry at me” quite well, driving down the error signal from that particular plan/constraint mismatch... but then, if I don’t say anything, I will continue feeling guilty.



The mismatch with the constraint of “I will stop feeling guilty” drives up the error signal, causing the “I will just stay quiet” plan to be abandoned. At worst, my mind may find it impossible to find any plan which would fulfill both constraints, keeping me in an endless loop of alternating between two unviable scenarios.

There are some interesting aspects about the phenomenology of such a situation, which feel like they fit the PP model quite well. In particular, it may feel like *if I just focus on a particular craving enough, thinking about my desired outcome hard enough will make it true*.

Recall that under the PP framework, goals happen because a part of the brain *assumes* that they will happen, after which it changes reality to *make* that belief true. So focusing really hard on a craving for X makes it feel like X will become true, *because the craving is literally rewriting an aspect of my subjective reality to make me think that X will become true*.

When I focus hard on the craving, I am temporarily guiding my attention away from the parts of my mind which are pointing out the obstacles in the way of X coming true. That is, those parts have less of a chance to incorporate *their* constraints into the plan that my brain is trying to develop. This momentarily reduces the motion away from this plan, making it seem more plausible that the desired outcome will in fact become real.

Conversely, letting go of this craving, may feel like it is *literally making the undesired outcome more real*, rather than like I am coming more to terms with reality. This is most obvious in cases where one has a craving for an outcome that is impossible for certain, such as in the case of grieving about a friend’s death. Even after it is certain that someone is dead, there may still be persistent thoughts of *if only I had done X*, with an implicit additional flavor of *if I just want to have done X really hard, things will*

change, and I can't stop focusing on this possibility because my friend needs to be alive.

In this form, craving may lead to all kinds of [rationalization](#) and biased reasoning: a part of your mind is literally making you believe that X is true, because it wants you to find a strategy where X is true. This hallucinated belief may constrain all of your plans and models about the world in the same sense as *getting direct sensory evidence about X being true* would constrain your brain's models. For example, if I have a very strong urge to believe that someone is interested in me, then this may cause me to interpret *any* of his words and expressions in a way compatible with this belief, regardless of how implausible and [far-spread](#) of a distortion this requires.

The case of physical pain

Similar principles apply to the case of physical pain.

We should first note that pain does not necessarily need to be aversive: for example, people may enjoy the pain of exercise, hot spices or sexual masochism. Morphine may also have an effect where people report that they still experience the pain but no longer mind it.

And, relevant for our topic, people practicing meditation find that by shifting their attention *towards* pain, it can become *less* aversive. The meditation teacher Shinzen Young writes that

... pain is one thing, and resistance to the pain is something else, and when the two come together you have an experience of suffering, that is to say, 'suffering equals pain multiplied by resistance.' You'll be able to see that's true not only for physical pain, but also for emotional pain and it's true not only for little pains but also for big pains. It's true for every kind of pain no matter how big, how small, or what causes it. Whenever there is resistance there is suffering. As soon as you can see that, you gain an insight into the nature of "pain as a problem" and as soon as you gain that insight, you'll begin to have some freedom. You come to realize that as long as we are alive we can't avoid pain. It's built into our nervous system. But we can certainly learn to experience pain without it being a problem. ([Young, 1994](#))

What does it mean to say that *resisting* pain creates suffering?

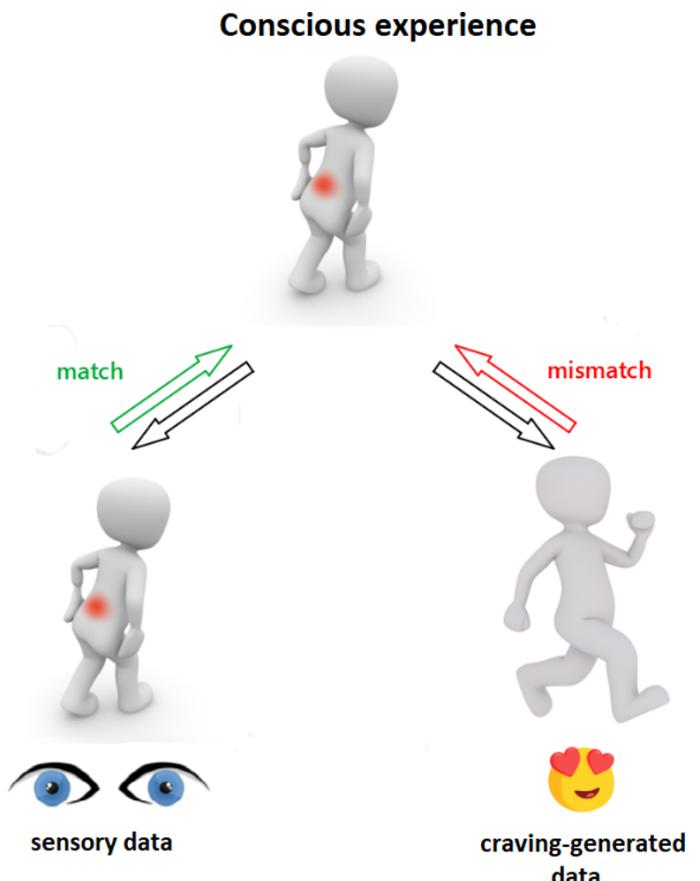
In the discussion about binocular rivalry, we might have said that when the mind settled on a hypothesis of seeing Isaac Newton, this hypothesis was *resisted* by the sensory data coming from the house-seeing eye. The mind would have settled on the hypothesis of "I am seeing Isaac Newton", if not for that resistance. Likewise, in the preceding discussion, the decision to admit the truth was resisted by the desire to not get yelled at.

Suppose that you have a sore muscle, which hurts whenever you put weight on it. Like sensory data coming from your eyes, this constrains the possible interpretations of what you might be experiencing: your brain might settle on the hypothesis of "I am feeling pain".

But the experience of this hypothesis then triggers a *resistance* to that pain: a craving subsystem wired to detect pain and resist it by projecting a form of internally-generated sense data, effectively claiming that you are *not* in pain. There are now

again two incompatible streams of data that need to be reconciled, one saying that you are in pain, and another which says that you are not.

In the case of binocular rivalry, both of the streams were generated by sensory information. In the discussion about psychological suffering, both of the streams were generated by craving. In this case, craving generates one of the streams and sensory information generates the other.



On the left, a persistent pain signal is strong enough to dominate consciousness. On the right, a craving for not being in pain attempts to constrain consciousness so that it doesn't include the pain.

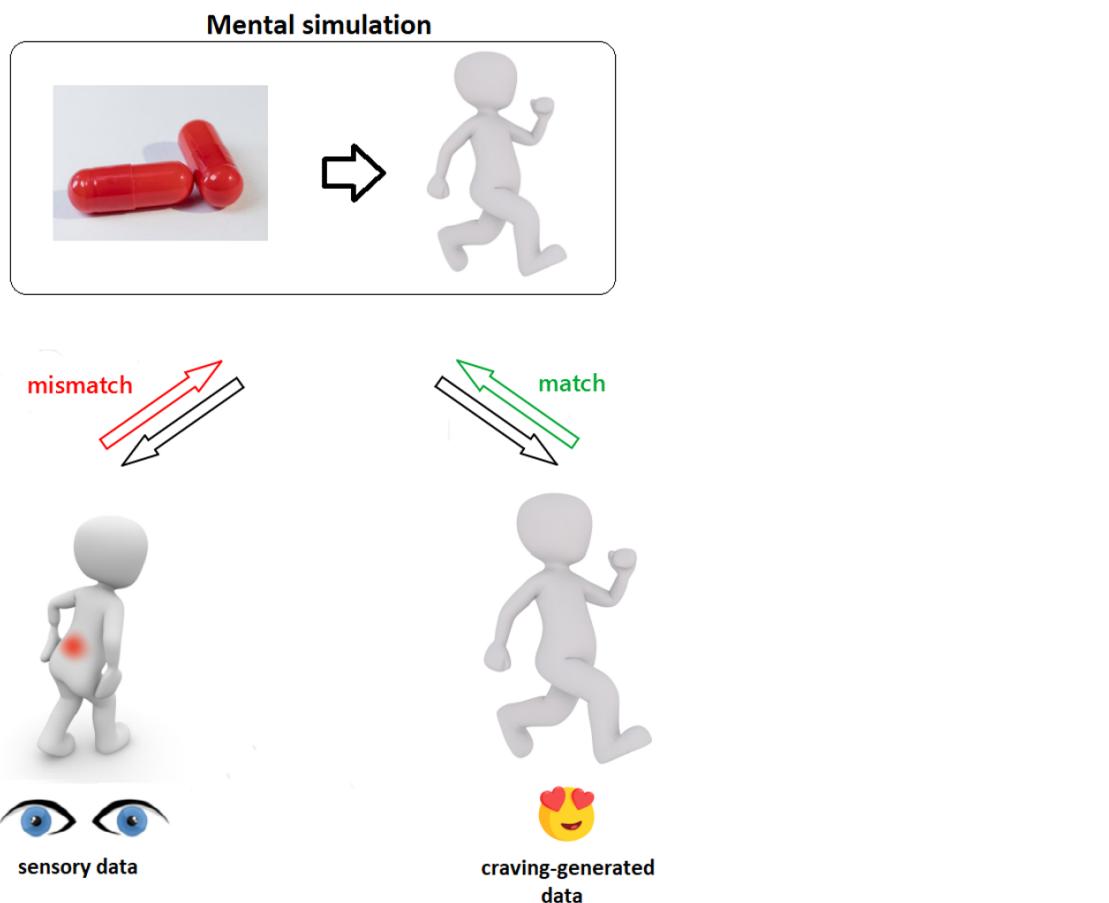
Now if you stop putting weight on the sore muscle, the pain goes away, fulfilling the prediction of "I am not in pain". As soon as your brain figures this out, your motor cortex can incorporate the craving-generated constraint of "I will not be in pain" into its planning. It generates different plans of how to move your body, and whenever it predicts that one of them would violate the constraint of "I will not be in pain", it will revise its plan. The end result is that you end up moving in ways that avoid putting weight on your sore muscle. If you miscalculate, the resulting pain will cause a rapid error signal that causes you to adjust your movement again.

What if the pain is more persistent, and bothers you no matter how much you try to avoid moving? Or if the circumstances force you to put weight on the sore muscle?

In that case, the brain will continue looking for a possible hypothesis that would fulfill the constraint of "I am not in pain". For example, maybe you have previously taken

painkillers that have helped with your pain. In that case, your mind may seize upon the hypothesis that “by taking painkillers, my pain will cease”.

As your mind predicts the likely consequences of taking painkillers, it notices that in this simulation, the constraint of “I am not in pain” gets fulfilled, driving down the error signal between the hypothesis and the “I am not in pain” constraint. However, if the brain could suppress the craving-for-pain-relief merely by *imagining* a scenario where the pain was gone, then it would never need to take any actions: it could just hallucinate pleasant states. Helping keep it anchored into reality is the fact that simply imagining the painkillers has not done anything to the pain signal itself: the imagined state does not match your actual sense data. There is still an error signal generated between the mismatch of the imagined “I have taken painkillers and am free of pain” scenario, and the fact that the pain is not gone yet.



*Your brain imagines a possible experience: taking painkillers and being free of pain. This imagined scenario fulfills the constraint of “I have no pain”. However, it does not fulfill the constraint of actually matching your sense data: you have **not** yet taken painkillers and **are** still in pain.*

Fortunately, if painkillers are actually available, your mind is not locked into a state where the two constraints of “I’m in pain” and “I’m not in pain” remain equally impossible to achieve. It can take actions - such as making you walk towards the medicine cabinet - that get you closer towards being able to fulfill both of these constraints.

There are studies suggesting that physical pain and psychological pain share similar neural mechanisms [citation]. And in meditation, one may notice that psychological discomfort and suffering involves avoiding unpleasant sensations in the same way as physical pain does; the same mechanism has been recruited for more abstract planning.

When the brain predicts that a particular experience would produce an unpleasant sensation, craving resists that prediction and tries to find another way. Similarly, if the brain predicts that something will *not* produce a pleasant sensation, craving may also resist *that* aspect of reality.

Now, this process as described has a structural equivalence to binocular rivalry, but as far as I know, binocular rivalry does not involve any particular discomfort. Suffering obviously does.

Being in pain is generally bad: it is usually better to try to avoid ending up in painful states, as well as try to get out of painful states once you are in them. This is also true for other states, such as hunger, that do not necessarily feel painful, but still have a negative emotional tone. Suppose that whenever craving generates a self-fulfilling prediction which resists your direct sensory experience, this generates a signal we might call “unsatisfactoriness”.

The stronger the conflict between the experience and the craving, the stronger the unsatisfactoriness - so that a mild pain that is easy to ignore only causes a little unsatisfactoriness, and an excruciating pain that generates a strong resistance causes immense suffering. The brain is then wired to use this unsatisfactoriness as a training signal, attempting to avoid situations that have previously included high levels of it, and to keep looking for ways out if it currently has a lot of it.

It is also worth noting what it means for you to be paralyzed by two strong, mutually opposing cravings. Consider again the situation where I am torn between admitting the truth to my friend, and staying quiet. We might think that this is a situation where the overall system is uncertain of the correct course of action: some subsystems are trying to force the action of confronting the situation, others are trying to force the action of avoiding it. Both courses of action are predicted to lead to some kind of loss.

In general, it is a bad thing if a system ends up in a situation where it has to choose between two different kinds of losses, and has high internal uncertainty of the right action. A system should avoid such dilemmas, either by avoiding the situations themselves or by finding a way to reconcile the conflicting priorities.

Craving-based and non-craving-based motivation

What I have written so far might be taken to suggest that craving is a requirement for all action and planning. However, the Buddhist claim is that craving is actually just one of at least two different motivational systems in the brain. Given that neuroscience suggests the existence of [at least three different motivational systems](#), this should not seem particularly implausible.

Let's take another look at the types of processes related to binocular rivalry versus craving.

Craving acts by actively introducing false beliefs into one's reasoning. If craving could just do this completely uninhibited, rewriting *all* experience to match one's desires,

nobody would ever do anything: they would just sit still, enjoying a craving-driven hallucination of a world where everything was perfect.

In contrast, in the case of binocular rivalry, no system is feeding the reasoning process any false beliefs: all the constraints emerge directly from the sense data and previous life-experience. To the extent that the system can be said to have a preference over either the “I am seeing a house” or the “I am seeing Isaac Newton” hypothesis, it is just “if seeing a house is the most likely hypothesis, then I prefer to see a house; if seeing Newton is the most likely hypothesis, then I prefer to see Newton”. The computation does not have an intrinsic attachment to any particular outcome, nor will it hallucinate a particular experience if it has no good reason to.

Likewise, it seems like there are modes of doing and being which are similar in the respect that one is focused on process rather than outcome: taking whatever actions are best-suited for the situation at hand, regardless of what their outcome might be. In these situations, little unsatisfactoriness seems to be present.

In an earlier post, I [discussed](#) a [proposal](#) where an autonomously acting robot has two decision-making systems. The first system just figures out whatever actions would maximize its rewards and tries to take those actions. The second “Blocker” system tries to predict whether or not a human overseer would approve of any given action, and prevents the first system from doing anything that would be disapproved of. We then have two evaluation systems: “what would bring the maximum reward” (running on a lower priority) and “would a human overseer approve of a proposed action” (taking precedence in case of a disagreement).

It seems to me that there is something similar going on with craving. There are processes which are neutrally just trying to figure out the best action; and when those processes hit upon particularly good or bad outcomes, craving is formed in an attempt to force the system into repeating or avoiding those outcomes in the future.

Suppose that you are in a situation where the best possible course of action only has a 10% chance of getting you through alive. If you are in a non-craving-driven state, you may focus on getting at least that 10% chance together, since that’s the best that you can do.

In contrast, the kind of behavior that is typical for craving is realizing that you have a significant chance of dying, deciding that this thought is completely unacceptable, and refusing to go on before you have an approach where the thought of death isn’t so stark.

Both systems have their upsides and downsides. If it is true that a 10% chance of survival really is the best that you can do, then you should clearly just focus on getting the probability even that high. The craving which causes trouble by thrashing around is only going to make things worse. On the other hand, maybe this estimate is flawed and you could achieve a higher probability of survival by doing something else. In that case, the craving absolutely refusing to go on until you have figured out something better might be the right action.

There is also another major difference, in that craving does not *really* care about outcomes. Rather, it cares about avoiding positive or negative feelings. In the case of avoiding death, craving-oriented systems are primarily reacting to the *thought* of death... which may make them reject even plans which would *reduce* the risk of death, if those plans involved needing to think about death too much.

This becomes particularly obvious in the case of things like going to the dentist in order to have an operation you know will be unpleasant. You may find yourself highly averse to going, as you crave the comfort of not needing to suffer from the unpleasantness. At the same time, you *also* know that the operation will benefit you in the long term: any unpleasantness will just be a passing state of mind, rather than permanent damage. But avoiding unpleasantness - including the very thought of experiencing something unpleasant - is just what craving is about.

In contrast, if you are in a state of equanimity with little craving, you still recognize the thoughts of going to the dentist as having negative valence, but this negative valence does not bother you, because you do not have a craving to avoid it. You can choose whatever option seems best, regardless of what kind of content this ends up producing in your consciousness.

Of course, choosing correctly requires you to actually *know* what is best. Expert meditators have been known to sometimes ignore extreme physical pain that should have caused them to seek medical aid. And they probably *would* have sought help, if not for their ability to drop their resistance to pain and experience it with extreme equanimity.

Negative-valence states tend to correlate with states which are bad for the achievement of our goals. That is the reason why we are wired to avoid them. But the correlation is only partial, so if you focus too much on avoiding unpleasantness, you are falling victim to [Goodhart's Law](#): optimizing a measure so much that you sacrifice the goals that the measure was supposed to track. Equanimity gives you the ability to ignore your consciously experienced suffering, so you don't need to pay additional mental costs for taking actions which further your goals. This can be useful, if you are strategic about actually achieving your goals.

But while Goodharting on a measure is a failure mode, so is ignoring the measure entirely. Unpleasantness *does* still correlate with things that make it harder to realize your values, and the need to avoid displeasure normally operates as an automatic feedback mechanism. It is possible to have high equanimity and weaken this mechanism, [without being smart about it](#) and doing nothing to develop alternative mechanisms. In that case you are just trading Goodhart's Law for the opposite failure mode.

Some other disadvantages of craving

In the beginning of this post, I mentioned a few other disadvantages that craving has, which I have not yet mentioned explicitly. Let's take a quick look at those.

Craving narrows your perception, making you only pay attention to things that seem immediately relevant for your craving.

In predictive processing, [attention is conceptualized](#) as giving increased weighting to those features of the sensory data that seem most useful for making successful predictions about the task at hand. If you have strong craving to achieve a particular outcome, your mind will focus on those aspects of the sensory data that seem useful for realizing your craving.

Strong craving may cause [premature exploitation](#). If you have a strong craving to achieve a particular goal, you may not want to do anything that looks like moving away from it, even if that would actually help you achieve it better.

Suppose that you have a strong craving to experience a feeling of accomplishment: this means that the craving is strongly projecting a constraint of “I will feel accomplished” into your planning, causing an error signal if you consider any plan which does not fulfill the constraint. If you are thinking about a multistep plan which will take time before you feel accomplished, it will start out by you *not* feeling accomplished. This contradicts the constraint of “I will feel accomplished”, causing that plan to be rejected in favor of ones that bring you even *some* accomplishment right away.

Craving and suffering

We might summarize the unsatisfactoriness-related parts of the above as follows:

- Craving tries to get us into pleasant states of consciousness.
- But pleasant states of consciousness are those *without* craving.
- Thus, there are subsystems which are trying to get us into pleasant states of consciousness by creating constant craving, which is the exact *opposite* of a pleasant state.

We can somewhat rephrase this as:

- The default state of human psychology involves a degree of almost constant dissatisfaction with one’s state of consciousness.
- This dissatisfaction is created by the craving.
- The dissatisfaction can be ended by eliminating craving.

... which, if correct, might be interpreted to roughly equal the first three of Buddhism’s [Four Noble Truths](#): the fourth is “Buddhism’s Noble Eightfold Path is a way to end craving”.

A more rationalist framing might be that the craving is essentially acting in a way that looks similar to [wireheading](#): pursuing pleasure and happiness even if that sacrifices your ability to impact the world. Reducing the influence of the craving makes your motivations less driven by wireheading-like impulses, and more able to see the world clearly even if it is painful. Thus, reducing craving may be valuable even if one does not care about suffering less.

This gives rise to the question - how exactly *does* one reduce craving? And what does all of this have to do with the self, again?

We’ll get back to those questions in the next post.

This is the third post of the “a non-mystical explanation of insight meditation and the three characteristics of existence” series. The next post in the series is “[From self to craving](#)”.

Speculations on the Future of Fiction Writing

If we look at major movies, writing is generally a relatively small chunk of the overall budget. A handful of examples from [Wikipedia](#):

- Unbreakable: story rights + screenplay ~8.1% of budget
- Tomb Raider - Cradle of Life: story rights + screenplay ~3.4%
- Terminator 3: screenplay ~2.8%
- Spider-Man 2 - 2004 version: screenplay ~5%

Eyeballing these numbers, I'd guess that if you could consistently get a 10% better movie by spending twice as much on the writing, that would be a great deal.

Why aren't studios *already* spending twice as much for 10% better movies? It doesn't seem like it ought to be that hard; it's not like the tropes on the [bad writing index](#) are hurting for examples. I'd guess that this is mainly a case of the [difficulty of hiring experts](#): it's hard to hire people with better taste than whoever's in charge.

On the other side of the equation, [writing techniques for intelligent characters](#) are proving not just [popular](#), but robustly popular. Scroll down the list at [topwebfiction.com](#), and you'll see an awful lot of consistent intelligence. The internet's top writing doesn't rely on artificial stupidity.

On top of that, the techniques for writing intelligent characters (at least [level 1 intelligent](#)) are largely agnostic to setting and plot. It's largely about sprinkling in plausible in-universe reasons why characters don't just do the obvious thing. Such techniques are well-suited to fanfiction for exactly that reason: they can be stitched in after-the-fact without completely throwing out the whole story.

Put these two pieces together. On one side, we have movie studios (and video game studies, and...) who'd love to throw more money at writing in order to make it better, but don't have consistent formulas for how to do that. On the other side, we have a handful of setting/plot-agnostic techniques for writing intelligent characters, and such writing is already proving very popular online.

Sounds like there's probably a market for intelligentification of characters.

At its simplest, this would be a service which takes in stories - screenplay, storyboard script, traditional books, what have you - and performs minimally-invasive surgery to make the characters not-stupid. It would involve tweaking background details of the universe to remove obvious exploits, or adding in-universe problem constraints to drive plot-relevant decisions rather than somebody holding the [idiot ball](#), or having a character in a Dramatic Moment desperately search for a solution that never comes rather than give up immediately. It wouldn't turn Captain America into the next HPMOR, but it would at least clean up the groan-worthy stupidity without throwing away the whole script.

Of course, the first customers of such a service probably wouldn't be A-list movie producers. The first customers would be indy game developers or small authors or whoever usually hires an editor. It would be a specialized editing service. If and when such a service could demonstrate substantial value-add, it would have a pitch for

bigger projects. At that point, it would be in a relatively-high-leverage position to [raise the sanity waterline](#).

AGIs as collectives

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

Note that I originally used the term population AGI, but changed it to collective AGI to match Bostrom's usage in Superintelligence.

I think there's a reasonably high probability that we will end up [training AGI in a multi-agent setting](#). But in that case, we shouldn't just be interested in how intelligent each agent produced by this training process is, but also in the combined intellectual capabilities of a large group of agents. If those agents cooperate, they will exceed the capabilities of any one of them - and then it might be useful to think of the whole collective as one AGI. Arguably, on a large-scale view, this is how we should think of humans. Each individual human is generally intelligent in our own right. Yet from the perspective of chimpanzees, the problem was not that any single human was intelligent enough to take over the world, but rather that millions of humans underwent cultural evolution to make the human collective as a whole much more intelligent.

This idea isn't just relevant to multi-agent training though: even if we train a single AGI, we will have strong incentives to copy it many times to get it to do more useful work. If that work involves generating new knowledge, then putting copies in contact with each other to share that knowledge would also increase efficiency. And so, one way or another, I expect that we'll eventually end up dealing with a "collective" of AIs. Let's call the resulting system, composed of many AIs working together, a *collective AGI*.

We should be clear about the differences between three possibilities which each involve multiple entities working together:

1. A single AGI composed of multiple modules, trained in an end-to-end way.
2. The [Comprehensive AI Services](#) (CAIS) model of a system of interlinked AIs which work together to complete tasks.
3. A collective AGI as described above, consisting of many individual AIs working together in comparable ways to how a collective of humans might collaborate.

This essay will only discuss the third possibility, which differs from the other two in several ways:

- Unlike the modules of a single AGI, the members of a collective AGI are not trained in a centralised way, on a single objective function. Rather, optimisation takes place with respect to the policies of individual members, with cooperation between them emerging (either during training or deployment) because it fits the incentives of individuals.
- Unlike CAIS services and single AGI modules, the members of a collective AGI are fairly homogeneous; they weren't all trained on totally different tasks (and in fact may start off identical to each other).
- Unlike CAIS services and single AGI modules, the members of a collective AGI are each generally intelligent by themselves - and therefore capable of playing multiple roles in the population AGI, and interacting in flexible ways.
- Unlike CAIS services and single AGI modules, the members of a collective AGI might be individually motivated by arbitrarily large-scale goals.

What are the relevant differences from a safety perspective between this collective-based view and the standard view? Specifically, let's compare a "collective AGI" to a single AGI which can do just as much intellectual work as the whole collective combined. Here I'm thinking particularly of the most high-level work (such as doing scientific research, or making good strategic decisions), since that seems like a fairer comparison.

Interpretability

We might hope that a collective AGI will be more interpretable than a single AGI, since its members will need to pass information to each other in a standardised "language". By contrast, the different modules in a single AGI may have developed specialised ways of communicating with each other. In humans, language is much lower-bandwidth than thought. This isn't a necessary feature of communication, though - members of a population AGI could be allowed to send data between each other at an arbitrarily high rate. Decreasing this communication bandwidth might be a useful way to increase the interpretability of a population AGI.

Flexibility

Regardless of the specific details of how they collaborate and share information, members of a collective AGI will need structures and norms for doing so. There's a sense in which some of the "work" of solving problems is done by those norms - for example, the [structure of a debate](#) can be more or less helpful in adjudicating the claims made. The analogous aspect of a single AGI is the structure of its cognitive modules and how they interact with each other. However, the structure of a collective AGI would be much more flexible - and in particular, it could be redesigned by the collective AGI itself in order to improve the flow of information. By contrast, the modules of a single AGI will have been designed by an optimiser, and so fit together much more rigidly. This likely makes them work together more efficiently; the efficiency of end-to-end optimisation is why a human with a brain twice as large would be much more intelligent than two normal humans collaborating. But the concomitant lack of flexibility is why it's much easier to improve our coordination protocols than our brain functionality.

Fine-tunability

Suppose we want to retrain an AGI to have a new set of goals. How easy is this in each case? Well, for a single AGI we can just train it on a new objective function, in the same way we trained it on the old one. For a collective AGI where each of the members was trained individually, however, we may not have good methods for assigning credit when the whole collective is trying to work together towards a single task. For example, a difficulty discussed in [Sunezag et al. \(2017\)](#) is that one agent starting to learn a new skill might interfere with the performance of other agents - and the resulting decrease in reward teaches the first agent to stop attempting the new skill. This would be particularly relevant if the original collective AGI was produced by copying a single agent trained by itself - if so, it's plausible that multi-agent reinforcement learning techniques have lagged behind.

Agency

This is a tricky one. I think that a collective AGI is likely to be less agentic and goal-directed than a single AGI of equivalent intelligence, because different members of the collective may have different goals which push in different directions. However, it's also possible that collective-level phenomena amplify goal-directed behaviour. For example, competition between different members in a collective AGI could push the group as a whole towards dangerous behaviour (in a similar way to how competition between companies makes humans less safe from the perspective of chimpanzees). And our lessened ability to fine-tune them, as discussed in the previous paragraph, might make it difficult to know how to intervene to prevent that.

Overall evaluation of collective AGIs

I think that the extent to which a collective AGI is more dangerous than an equivalently intelligent single AGI will mainly depend on how the individual members are trained (in ways which [I've discussed previously](#)). If we condition on a given training regime being used for both approaches, though, it's much less clear which type of AGI we should prefer. It'd be useful to see more arguments either way - in particular because a better understanding of the pros and cons of each approach might influence our training decisions. For example, during multi-agent training there may be a tradeoff between training individual AIs to be more intelligent, versus running more copies of them to teach them to cooperate at larger scales. In such environments we could also try to encourage or discourage them from in-depth communication with each other.

In my next post, I'll discuss one argument for why collective AGIs might be safer: because they can be deployed in more constrained ways.

Conjecture Workshop

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

The conjecture workshop is an activity I've run three times now in a conference/meetup-style setting. Feedback has been unusually positive, so I'm writing it up here to codify the basic idea and hopefully get independent feedback from others who try it.

The overall goal is to translate intuitive ideas into mathematical conjectures.

Format

People break into groups of 2-3. Within each group, one person serves as "conjecturer", and the other one or two serve as "questioners". Roles can rotate over the course of the allotted time (~1 hr for sessions so far).

To start, the conjecturer should have some intuitive claim in mind that they want to formalize. Some AI-oriented examples from previous sessions:

- In beneficial comprehensive AI services, security services and planning services necessarily have to be "agenty"
- There exists some "universal" algorithm which performs about-as-well as any other on optimization problems with bounded runtime
- Coarse-grained models of the world form a category, and we can always construct pullback/pushout models within that category

Note that these are quite fuzzy and leave out a lot of the idea; fully explaining the ideas intuitively (even without formalization) takes much longer than the short blurbs above. Indeed, the ideas usually start out even fuzzier than the blurbs above - just summarizing them in a sentence is hard.

Once the conjecturer has something in mind, they try to explain the claim to the questioners, intuitively. The questioners' job is to regularly interrupt, and ask questions to help formalize what the conjecturer is saying. Common examples include:

- "You've mentioned <thing> a couple of times. Should we give it a name?"
- "What type of thing is <thing we just named>? Any constraints on it?"
- "Ok, so we claim/assume that <thing> is <better/larger/simpler/etc> than <other thing> in some intuitive sense. How do we quantify that?"
- "So we want to assume <intuitive assumption>. What does that mean, mathematically?"
- "Does the model so far accurately capture your intuitions, at least the parts which intuitively seem relevant to the claim?"

The questioners should also ask general explanation-support questions, like "can you give an example?", "can you repeat/clarify that last part?", and repeating back the questioner's understanding of the claim so far. In particular, the questioners should remind the conjecturer to write down any components of the mathematics (i.e. variable/function definitions, assumptions, claim itself, etc) as they come up.

Key point: PROVING OR DISPROVING THE CLAIM IS NOT THE GOAL. For purposes of this activity, we do not care whether or not the claim is true; the goal is simply to formalize it enough that it could be mathematically proven/disproven. (One minor exception: if the claim seems *trivially* true/false at some point, that's often evidence that some key piece of the conjecturer's intuition has not been captured.)

Value Model

The idea behind the exercise is that [translation is a scarce resource](#) - in this case, translation of intuitions into mathematical formalism. Often, a major bottleneck to theoretical/modelling work is simply expressing the ideas mathematically.

Focusing on a particular conjecture also helps avoid a “model ALL the things” failure mode. Since there’s one particular claim we want to set up, we just work out the pieces necessary for that claim, not for a whole theory of everything.

Your abstraction isn't wrong, it's just really bad

This is a linkpost for

https://blog.cerebralab.com/Your_abstraction_isn't_wrong,_it's_just_really_bad

Epistemic status: pretty sure

For a programming language to qualify as such, the only thing it needs is Turing Completeness. That means, at least in principle, it can be used to solve any (solvable) computational problem.

There are features besides Turing Completeness which make it desirable, but if you tried hard enough, all of those features could be implemented as part of any Turing complete language.

Let's take a maximally perverse theory of abstraction-making that says:

An abstraction is sufficiently good if, in principle, it can yield answers to all questions that it's designed to solve.

Under this definition, the assembly language (ASM) of virtually any computer is sufficient abstraction for writing code.

The only "problem" with ASM is that writing something simple like [quicksort](#) ends up being monstrously large and hard to read (and this implementation is written for clarity, not speed).

Compare that to a [quicksort implementation](#) written in, C. It's only 13 lines of code and about as equally efficient as the ASM one above. Even better, it would work on virtually all machines, while the above ASM wouldn't and would have to be rewritten for a bunch of architectures.

C might not be the best abstraction one can use for low-level programming, e.g. writing sorting algorithms, but it's better than ASM.

However the ASM implementation is far from being the maximally perverse version one can get under our naive abstraction-choosing constraints. The prize for that would currently go to something like the [brainfuck implementation](#):

```
>>>>>>, [>+>>, ]>+[ - - [+<<<- ]<[ <>- ]<[ <[ ->[ <<<+>>>>+<- ]<<[ >>+>[ ->] <<[ <<- ]>]>>>+<[ [ - ]<[ >+<- ]<>[ [ >>>] +<<<- <[ <<<[ <<< ]>>+>[ >>>] <- ]<<[ <<< ]>[ >>[ >> > ]<+ <<[ <<< ]>- . ] ]+<<<+[->>>] >>>[ . >>>]
```

Now that, that is what I call seriously fucked up.

But [brainfuck is Turing complete](#), so under the above definition, it is no worse than ASM or C.

How we avoid bad languages?

So C, C++, Javascript, Python, Julia, any given ASM, Fortran, Lua, Elm, Clojure, CLISP, Rust, Forth, Elixir, Erlang, Go, Nim, Scala, Ada, Cobolt, Brainfuck, and, sadly enough, even Java, are "good enough to be an abstraction for controlling ANY computing systems" under the perverse requirements above.

In reality, it's fairly easy to agree that some languages are better than others in various situations. I think most criteria for choosing a language are at least somewhat subjective. Things such as:

- Architectures that the compiler will target
- Memory safety features.
- Built-in parallelism and concurrency mechanisms.
- Functionality of standard library.
- Available package manager.
- The companies/projects/communities that already use it.
- Ease of learning.
- Ease of reading.
- Ease of debugging.
- Speed of compiler[s].
- Efficiency of memory usage (e.g. via avoiding spurious pointers and having move semantics).
- Performance on various benchmarks.
- What I want to use the language for (which is probably the most important).

But even subjective criteria allows me to ask a question like: "What do I need in order to create a language that is better than C and C++ for systems programming?". To which the answer will be something like:

- Be Turing complete
- [Support all or most targets that C/C++ support](#)
- Have memory safety feature (e.g. borrow checking, automatic deallocation) that C and C++ don't have.
- Have the same concurrency abstractions as C and C++ (Threads mapping to kernel threads)) and maybe some extra ones (e.g. futures and async-io semantics for more efficient non-blocking IO).
- Have a standard library that includes all glibc and std-lib functionality plus some extra things that people want.
- Have a package manager.
- Try to make nice companies and welcoming communities that are viewed well by outsiders use your language.
- Have better, more centralized documentation than C/C++ and community that is nicer and more helpful to newbies.
- Try to be about as easy to read as C and C++, ideally more so by getting rid of BC syntax and outdated operators.
- Have a good debugger and nice error logs, or at least nicer than C and C++.
- {Try} to have a fast compiler.
- Be about as efficient as C and C++ when working with memory, potentially making useful abstractions (e.g. moving, deallocation) implicit instead of explicit whenever possible. Perhaps by using a better default allocator.
- [Perform equally~ish well to C and C++ on various benchmarks that people trust](#)

- Be useable in the same situations C and C++ are, maybe more, if possible. This is partially a function of all of the above plus the external library ecosystem, which the language designer can nudge but has no direct control of.

What I describe above is, of course, Rust. A language that, with certain caveats, has managed to become better than C and C++[citation needed] for systems programming (and many other things).

Granted, the world hasn't been rewritten in Rust yet. (Though I hear the [R.E.S.F](#) has managed to secure an audience with God and given his buy-in, you'd be only one level removed from converting von Neuman.)

But it does have wide adoption and I would be deeply surprised if in 10-20 years from now most server kernels and common low-level libraries (e.g. BLAS, CUDNN) won't be written in it.

For a more post factum example, you can take something like Python, due to a good package manager and easy to understand syntax, it managed to become the dominant scripting language for... everything.

For another example, take Nodejs, which used people's familiarity with javascript and it's build-in asyncio capabilities to become the language of choice for many low-load http servers.

Or take something simpler, like TypeScript, which just took Javascript, added a feature people really, really wanted (types), and swiftly became a fairly dominant language in the frontend world.

Our model for avoiding bad languages is based on continuous improvement and competition. You look at what existing language are used for and what people want to do with them, you look at their syntax and their performance. Then you build something that fits the basic requirements (is Turing complete, runs on x86) and tries to be better than them in the pain points previously identified. If your new language is good enough, it will sweep everyone away and 10 years from now we'll all be using it.

A similar but less complex set of criteria can be defined for libraries and open-source software in general. A similar but more complex set of criteria can be defined for processor and computer architectures.

The languages, processors, and libraries of the 80s are all "kinda bad" by modern standards. The ones of the 40s are basically at the brainfuck levels of horribleness. But this is not to say that the people of the 80s or 40s were bad programmers, it's just that they didn't have the knowledge and tools that we have right now.

Indeed, I'd wager that the pioneers of computer science were probably smarter than the people doing it now, successful pioneers tend to be that way, but seeing mistakes in hindsight is much easier than predicting them.

The requirements for avoiding bad abstraction

So, I think the requirements for avoiding bad abstraction, at least in terms of programming languages and programming abstractions in general, can be boiled down to:

1. Minimal and rather specific core standards (Turing completeness, ability to run on a popular-ish CPU).
2. Ability to name areas where improvements can be made such that the language could be considered "better" than current ones.
3. A general willingness of the community to adopt the language. Or at least a mechanism by which the community can be forced to adopt it (see market-driven competition).

Granted, these requirements are far from being fulfilled perfectly. The adage still holds mostly true:

Science progresses one funeral at a time

Programming languages are not quite like scientific paradigms, partially because the field they are used in is far more competitive, partially because they are much easier to design than a scientific theory.

An average programmer might switch between using 2 or 3 significantly different languages in one area over their lifetime. An especially good one might bring that number to 10, 20, or 50 (most of which are going to be failed experiments).

There's no programming God that can just say "X is obviously better than Y so switch all Y code to X and start using X". There is however something close to an efficient market that says something like:

If I can design a piece of software in 1 hour using my language and you can design it in 100 hours using yours, I will soon build a 3 person startup that will eat your mid-sized corporation for lunch.

However, I don't want to downplay the role that community plays here.

Imagine designing a new language and trying to improve and popularize it, you'll need validation and feedback. In general, the reactions you will get will probably be neutral to positive.

Being already popular and respected helps (see Mozilla and Rust), but so does having a language that seems very obvious useful by filling a niche (see frontend devs wanting to work full-stack and Node) or just having an amazing syntax (see Python way back when). However, the worst-case scenario is that you end up getting some "oh that's nice BUT ..." style reactions and are left with a good learning experience as the basis for your next attempt.

I don't think anyone ever tried to build a programming language (or a library or a CPU arch for that matter) and was met with:

Oh, you stupid quack, you think you're better than the thousands of language designers and dozens of millions of developers building and using a programming language. You **really** think you've managed to out-smart so many people and come up with a better language/library !?

I think that the lack of this reaction is good. However, try to challenge and redesign a core abstraction of any other field (internal medicine, high-energy physics, astronomy, metabolic biology... etc) and that's roughly the reaction you will get. Unless you are a very respected scientist, in which case you will usually get indifference instead and maybe in 60 years, after the death of a few generations, your abstractions will start being adopted. But this is a *very* subjective impression.

On the other hand, good CS students are routinely challenged to design their language and compiler. Maybe 999/1,000 times it sucks, but that's ok because they get a better understanding of why our current languages are good and of the principles behind them. Even better, maybe 1/1,000 times you get the seeds of something like Scala.

I have it on good authority that no physics students were ever tasked with trying to redesign the mathematical apparatus used to interpret the results of high energy collisions into taxonomical facts.

We make fun of tools, but in general, we never make fun of tool-creation. Creating tools for programming (e.g. language and libraries) is as close as you can get to a sacrament in the programming community. It's something that's subconsciously seen as so desirable that nobody has anything but praise for the attempt (though, again, the results are fair game for harsh critique).

Note: Of course, outliers exist everywhere, including here.

Is your abstraction bad?

So the reasons we know our programming abstractions (e.g. languages and libraries) are kind of good is that we have a market mechanism to promote them, a culture that encourages and helps their creation, and both explicit and implicit criteria that we can judge them by.

We don't know how good they are in an absolute sense, we can be almost certain they are not the best, but we can be 100% sure that they are not the worst, because we know they aren't Brainfuck.

But take away any piece of the puzzle and the whole edifice crumbles, or at least is badly damaged. If you replace the free-market determination with a council of elders or a regulatory board, then Node and Python don't exist. If you replace the community with one that is trying to minimize the amount of new things they have to learn, then Rust doesn't. If the quality criteria were much fuzzier than we might have just stopped at ASM instructions and never designed any language.

This is the fear I have about a lot of other abstractions.

Take for example taxonomies in fields like biology and medicine.

There's a taxonomy that clusters things that can approximately replicate the contents of DNA and RNA molecules encased within them into: Realm > Kingdom > . . . > Genus > Species
.

There are taxonomies that cluster symptoms together into diseases to better target them with drugs, investigate the underlying cause, and predict their progress.

But the way these taxonomies are designed does not seem immediately obvious, nor does it seem like one of the fundamental questions that their respective fields struggle with.

If I inquire "why is life classified the way it is?", I think the best steelman of a biologist my mind can come up with will answer something like:

Well, because it evolved (no pun intended) all the way from Aristotle to us. There are some good Schelling points such as "things that fly are different from things that walk, are different from things that crawl, are different from things that swim, are different from things which are immutable and green". Rough outlines were built around those and some ideas (e.g. Kingdom, Species) kinda stuck because they seem obvious. Once we understood evolution and the fossil records we realized that things are a bit more complex and we came up with this idea of phylogenetics, but we decided to work around the existing rank in the hierarchy to calcify their meaning a bit better, then added more ranks once they seemed necessary.

Yes, maybe even viewing it as a hierarchy is rather stupid, since that whole view works for e.g. multicellular eukaryotes that tend to keep slowly adding to their DNA and leave a fossil record but seems kinda silly for bacteria which swap, add and discard DNA like crazy and leave none but the vaguest trace of their existence a few seconds after their cell wall breaks. But it kinda works even for viruses and bacteria if you make a few adjustments.

Yes, there are arbitrary rules we apply, for example, the more foreign a lifeform is to us the more likely we are to use genetics to classify it, and the more often we encounter it the more likely we are to use phenotype.

Yes, maybe forcing kids to memorize these things in school is pointless and arbitrary, and yes I see no particular reason to put this taxonomy on a golden pedestal, but it's the one that everyone uses so you might as well get used to it.

This is a fairly good defense of the taxonomy, all things considered, it's a very similar defense to the one I'd give if someone asked me why it's still relevant to know C or C++.

But it's a defense that leverages the idea that the current taxonomy is fit enough for its purpose, thus there's no reason to change it. However, I fail to see why we consider it to be so. The role of a taxonomy is to dictate discussion, indexing, and thought patterns, this is a rather complex subject and can only be explored via querying individual preferences and observing how individuals using different taxonomies fare against each other in a competitive environment. But if my only option is to use this taxonomy and doing away with the whole edifice in favor of something new is off-limits, then I think it's fair to argue that we have exactly 0 datapoints to suggest this is a good taxonomy.

If the whole reason we use the taxonomy is because "It kinda classifies all life, so it's fit for purpose", that's like saying brainfuck is Turing complete, so it's fit for purpose. An abstraction can be fit for purpose under the most perverse possible definition and still be very bad.

In programming, if I think C is garbage, or, even worst, if I think OO and imperative programming as a whole is rubbish, I can go to one of 1001 university departments, meetups, and companies that use functional languages and patterns.

If I think that even those are rubbish and I specifically want a functional language designed under some niche tenants of a sub-branch of set theory designed by a few mathematicians in the 50s... I can go to Glasgow and find a whole university department dedicated to that one language.

The reason I can point to C and say "this is an ok language" is because I can take i3 and xmonad, look at their code+binary and say "Look, this is a C program compared to a Haskell one, they fulfill the same function and have x,y,z advantage, and disadvantages". That times 1000, means I have some reason to believe C is good, otherwise C programmers would be out of a job because everyone would be writing superior software in Haskell.

But I'm not aware of any department of biology that said: "Hmh, you know what, this life-classification system we have here could be improved a lot, let's redesign it and use the new system to communicate and teach our students to use it". If we had one such department or 100 such departments, we would suddenly have some data points to judge the quality of our current taxonomy.

A dominant abstraction still has a huge home-field advantage. But the reason for experimenting with abstraction is not to get minor improvements (e.g. Python3.7 to Python3.8) but to get paradigm-shifting ones (e.g. C vs Rust). The changes we are looking for should still be visible even in those subpar conditions.

If Firefox used to be kinda slow, and in the last 2 years it's become the fastest browser on all platforms, that starts to hint at "Huh, maybe this Rust thing is good". It's not a certainty, but if we have 100 other such examples than that's much better than nothing.

Conversely, if the university working with {counterfactual-taxonomy-of-life} is suddenly producing a bunch of small molecules that help people stay thin without significant side effects, although billions of dollars went into researching them and nobody really found one before, maybe that's a hint that {counterfactual-taxonomy-of-life} is bringing some useful thought patterns to the table. It's not a certainty, but if we have 100 other such examples than that's better than nothing.

Granted, I don't expect fundamental taxonomies to be a good candidate for this approach, they are just an easy and uncontroversial example to give. Everyone agrees they are somewhat arbitrary, everyone agrees that we could create better ones but the coordination problem of getting them adopted is not worth solving.

So fine, let me instead jump ship to the most extreme possible example.

Why do we have to use math?

Generally speaking, I can't remember myself or anyone else as kids protesting against learning numbers and how to add and subtract them. As in, nobody asked "Why do we have to learn this?" or "Is this useful to adults?".

Granted, that might be because children learning numbers have just mustered basic speech and bladder control a few birthdays ago, so maybe there's still a gap to be

crossed until they can protest to their teachers with those kinds of questions.

But I doubt it. I think there's something almost intuitive about numbers, additions, and Euclidean geometry. Maybe one doesn't come up with them on their own in the state of nature, but they are very intuitive abstractions, once someone points them out to you they seem obvious, natural, true.

This is a famous argument and I think Plato does better justice to it than I could.

Some context, Socrates (So) is arguing with Meno about the existence of "recollection" (I'd rather think of this as abstractions/ideas that become immediately obvious to anyone once pointed out). He brings out a slave with no knowledge of geometry and draws the following figures:

<http://cgal-discuss.949826.n4.nabble.com/file/n2015843/image.jpg>

- So: Tell me, boy, do you know that a square is like this?
- Slave: I do.
- So: And so a square has these lines, four of them, all equal?
- Slave: Of course.
- So: And these ones going through the center are also equal?
- Slave: Yes.
- So: And so there would be larger and smaller versions of this area?
- Slave: Certainly.
- So: Now, if this side were two feet and this side two feet also, how many feet would the whole be? Look at it like this: if this one were two feet but this one only one foot, wouldn't the area have to be two feet taken once?
- Slave: Yes.
- So: When this one is also two feet, there would be twice two?
- Slave: There would.
- So: An area of twice two feet?
- Slave: Yes.
- So: How much is twice two feet? Calculate and tell me.
- Slave: Four, Socrates.
- So: Couldn't there be one different from this, doubled, but of the same kind, with all the lines equal, as in that one?
- Slave: Yes.
- So: And how many feet in area?
- Slave: Eight.
- So: Come then, try to tell me how long each line of this one will be. In that one, it's two, but what about in that doubled one?
- Slave: It's clearly double, Socrates.
- So: You see, Meno, that I am not teaching anything, but put everything as a question. He now believes he knows what sort of line the eight feet area comes from. Or don't you think so?

The point at which children start questioning their math teachers seem to coincide with the point where more complex abstraction is introduced.

There's nothing fundamentally "truer" about Euclidian geometry than about analysis. Yes, the idea of breaking down lines into an infinity of infinitely small distances might conflict with the epistemology of a kid, but so might the axioms of Euclidian geometry.

Where the difference between Euclidian geometry and analysis lies is in the obviousness of the abstraction. With something like analysis it seems like the

abstractions being thought are kind of arbitrary, not in that they are untrue, but in that they could be differently true.

Maybe it's not even obvious why "infinitely small distance" is a better choice of abstraction than "100,000,000,000 small distances".

It's not obvious why reasoning analytically about the area under a curve is superior to the Greek geometric approach of starting with a few priors and reasoning out equality through similarity. (not to a kid, that is)

It's not obvious why the 'trend' of a certain single-parameter function, represented by another single-parameter function is an important abstraction to have at all.

I think that kids asking their math teachers "Why do we have to learn this?", want, or at least would be best served with an answer to one of two interpretations:

1. Why is this abstraction more relevant than any other abstraction I could be learning? The abstractions math previously gave me seemed like obvious things about the world. This analysis thing seems counter-intuitive at first, so even if it's true, that in itself doesn't seem like reason enough for me to care.
2. Why was this abstraction chosen to solve this set of problems? How did people stumble upon those problems and decide they were worth solving? What other abstractions did they try to end up with on that is so complex?

But instead, the answers most teachers give are answers to the question:

Why do we have to learn math?

I think most people have this mental split at some point, between the mathematics that is intuitive and that which isn't. Maybe for some people, it happens right after they learn to count, maybe for others, it happens when they get to 4d geometry.

I think many people blame this split on something like curiosity, or inborn ability. But I blame this split on whichever qualia dictate which mathematical abstractions are "intuitive" and which aren't.

Furthermore, I don't think this split can be easily resolved, I think to truly resolve it you'd need to find a problem impervious to intuitive abstractions, then try out a bunch of abstractions that fail short of solving it, then reason your way to one that does (which is likely going to be the "standard" one).

But it seems that abstraction-finding, whilst most certainly a part of mathematics, is something almost nobody is explicitly taught how to do.

To put it another way, I think that anyone learning to program, if asked "How would you redesign C to make it better?", could give an answer. Maybe a wrong answer, almost certainly an answer far worst than the "best" answers out there. Most people are asked some variant of this question, or at least ask themselves, a significant percentage even try to implement it... maybe not quite at the level of trying to redesign C, but at least at the level of trying to redesign some tiny library.

On the other hand, someone that's learned analysis for a few years, if asked how they would improve it, would fail to answer... even a poor answer, even a wrong answer, the question would seem as intractable to them as it would be to their 6-year-old self.

If I proposed to you that in 20 years schools and colleges would be teaching either Rust or Python instead of C and Java you might argue with that idea, but it would certainly seem like something within the realm of real possibilities.

If I proposed to you that in 20 years schools and colleges would have thrown away their analysis books and started teaching a very different paradigm you would ask me what's the stake and odds I'm willing to bet on that.

Maybe that is because math is perfect, or at least because math is close to perfect. Maybe one can make minute improvements and completions to the old knowledge, but the emphasis in that sentence should be placed on "minute" and "completions".

One thing that strikes me as interesting about mathematics, under this hypothesis, is that it seems to have gotten it impossibly right the first time around. E.g. the way one best abstracts figuring out the equation for the area under a curve in 17th-century with limited ink and paper, is the same way one best abstracts it when sitting at a desk with a computing machine millions of times faster than our brain.

I'm not comfortable making this point about analysis, I've tried thinking about an idea like "What if we assumed continuous variables did not exist and built math from there Pythagora style", every time I ran into an issue, so I'm fairly sure a good mathematician could poke 1001 holes in this approach.

However, in areas that interest me like statistics, I think it's fairly easy to see [gratingly bad abstractions](#) that have little reason for existing. From people still writing about trends without using any test data, let alone something like k-fold cross-validation, to people assuming reality has an affinity for straight lines and bell shapes until proven otherwise.

Maybe statistics is alone in using outdated abstractions, or maybe there are many areas of math like it. But because mathematics is (fairly) hierarchical in terms of the abstractions it uses, there's no marketplace for them to fight it out. New abstractions are forced to be born into new niches, or via strangling out a competitor by proving an edge case they couldn't.

When Python was born nobody ever claimed it does more than C, it is, by definition, impossible to do something with Python (as implemented by CPython) that can't be done with C. On the other hand, that doesn't make the Python abstraction inferior, indeed, for the vast majority of jobs, it's much better.

Is there such an abstraction we are missing out on in math? Some way to teach kids analysis that is as obvious as Euclidian geometry. I don't know, but I do know that I'd have no incentive to ever figure it out and I don't think anybody else does either. That fact makes me uncomfortable.

Perhaps mathematics is the worst possible field to reason about abstraction quality, but I feel like there are a lot of other easier picks where even a bit of abstraction competition could greatly improve things.

Alas

I think the way programming handles creating new abstractions is rather unique among any field of intellectual endeavor.

Maybe this is unique to programming for a reason or maybe I'm wrongfully associating the most fluid parts of programming with the most immutable parts of other fields.

But I do think that the idea is worth exploring more, especially light of our knowledge accumulation problem and the severe anti-intellectualism and lack of polymaths in the current world.

Maybe all theory is almost perfect and improving it can only be done incrementally. But maybe, if thousands of people attempted to completely revamp various theories every day, we'd come up with some exponentially better ones. I think the only field that provides evidence regarding this is programming and I think the evidence points towards the latter approach is very promising.

[AN #100]: What might go wrong if you learn a reward function while acting

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

Newsletter #100 (!!)

Alignment Newsletter is a weekly publication with recent content relevant to AI alignment around the world. Find all Alignment Newsletter [resources here](#). In particular, you can look through [this spreadsheet](#) of all summaries that have ever been in the newsletter.

Audio version [here](#) (may not be up yet).

HIGHLIGHTS

[Pitfalls of learning a reward function online](#) (*Stuart Armstrong et al*)

(summarized by Rohin): It can be dangerous to learn the metric that you are trying to optimize: if you don't set it up correctly, you may end up incentivizing the agent to "update in a particular direction" in the metric learning for the sake of future optimization (a point previously made in [Towards Interactive Inverse Reinforcement Learning](#)). This paper analyzes the problems that can arise when an agent simultaneously learns a reward function, and optimizes that reward function.

The agent may have an incentive to "rig" the reward learning process, such that it finds a reward that is easy to optimize. For example, consider a student Sandra who must figure out the deadline and evaluation criteria for a project from a teacher Trisha. Sandra expects that if she asks Trisha when the deadline is, she will say that the deadline is later this week. So, Sandra might cleverly ask, "Is the project due next week, or the week after", to which Trisha might respond "next week". In this way, Sandra can rig the deadline-learning process in order to obtain a more favorable deadline.

Worse, in such scenarios the need to rig the learning process can destroy value for every reward function you are considering. For example, let's suppose that if Trisha couldn't be manipulated, Sandra's optimal policy would be to start the project today, *regardless* of when the actual deadline is. However, given that Trisha *can* be manipulated, Sandra will spend today manipulating Trisha into setting a later deadline -- an action that seems clearly suboptimal from the perspective of any fixed deadline. The paper describes this as *sacrificing reward with certainty*.

To avoid such situations, we need *unriggable* learning processes, that is, ones where at all times, the expected final learned reward (deadline) is independent of the agent's (Sandra's) policy. This unriggability property is nearly equivalent to the property of *uninfluencability*, in which we must be able to posit some background variables in the environment such that the learning process can be said to be "learning" these

variables. Technically, an unriggable process need not be uninfluenceable, though it usually is (see the paper for details).

However, these properties only constrain the *expectation over environments* of the final reward distribution: it doesn't prevent the agent from somehow shuffling around reward functions to be matched with suitable environments. For example, without knowing which projects are easy or hard, Sandra could manipulate Trisha into giving early deadlines for easy projects, and late deadlines for hard projects, in a manner that preserved the *distribution* over early and late deadlines; this would satisfy the unriggable property (and probably also the uninfluenceable property, depending on the exact formalization).

The authors demonstrate these problems in a simple gridworld example. They also point out that there's a simple way to make any learning process uninfluenceable: choose a specific policy π that gathers information about the reward, and then define the new learning process to be "whatever the original learning process would have said if you executed π ".

Read more: [Blog post: Learning and manipulating learning](#)

Rohin's opinion: I would explain this paper's point somewhat differently than the paper does. Consider an AI system in which we build in a prior over rewards and an update rule, and then have it act in the world. At the end of the trajectory, it is rewarded according to the expected reward of the trajectory under the inferred posterior over rewards. Then, the AI system is incentivized to choose actions under which the resulting posterior is easy to maximize.

This doesn't require the reward function to be ambiguous; it just requires that the update rule isn't perfect. For example, imagine that Alice has a real preference for apples over bananas, and you use the update rule "if Alice eats an apple, infer that she likes apples; if Alice eats a banana, infer that she likes bananas". The robot finds it easier to grasp the rigid apple, and so can get higher expected reward in the worlds where Alice likes apples. If you train a robot in the manner above, then the robot will learn to throw away the bananas, so that Alice's only choice is an apple (that we assume she then eats), allowing the robot to "infer" that Alice likes apples, which it can then easily maximize. This sort of problem could happen in most current reward learning setups, if we had powerful enough optimizers.

It seems to me that the problem is that you are training the actor, but not training the update rule, and so the actor learns to "trick" the update rule. Instead, it seems like we should train both. This is kind of what happens with [assistance games / CIRL \(AN #69\)](#), in which you train a policy to maximize expected reward under the *prior*, and so the policy is incentivized to take the best information gathering actions (which, if you squint, is like "training to update well"), and to maximize what it thinks is the true reward. Of course, if your prior / update rule within the game are misspecified, then bad things can happen. See also Stuart's reactions [here](#) and [here](#), as well as my comments on those posts.

TECHNICAL AI ALIGNMENT

INTERPRETABILITY

Evaluating Explainable AI: Which Algorithmic Explanations Help Users

Predict Model Behavior? (Peter Hase et al) (summarized by Robert): In this paper the authors perform user tests on 5 different model agnostic interpretability methods: LIME, Anchor, Decision Boundary, Prototype Model and a Composite model (LIME Anchor and Decision Boundary). The use cases they test are a tabular dataset predicting income, and a movie-review dataset predicting sentiment of the review from a single sentence.

Their experimental setup consists of 2 tests: **forward prediction** and **counterfactual prediction**. In forward prediction, the user is shown 16 examples of inputs and corresponding outputs and explanations, and then must predict the model's output on new inputs (without the explanation, which often gives away the answer). In counterfactual prediction, after seeing 16 examples, the user is given an input-output-explanation triple, and then must predict how the output changes for a specific perturbation of the input.

Throughout the results they use a significance threshold of $p < 0.05$ (they don't use Bonferroni corrections). Their study has responses from 32 different students who'd taken at least 1 computer science course, with some screened out for outliers or low accuracy during training. There are approximately 200 individual predictions for each method/dataset-type combination, and each method/prediction-type combination.

Overall, their results show that **only LIME (Local Interpretable Model-agnostic Explanation) helps improve performance** with statistical significance on the tabular dataset across both prediction settings, and **only the Prototype model in counterfactual prediction across both datasets. No other result was statistically significant**. The improvement in accuracy for the statistically significant results is around 10% (from 70% to 80% in the Tabular dataset with LIME, and 63% to 73% for Prototype in counterfactual prediction).

They also showed that **user's ratings of the explanation method didn't correlate in a statistically significant way with the improvement the model gave to their predictions**.

Robert's opinion: I'm happy a paper like this exists, because I think this kind of work is crucial in evaluating whether interpretability methods we're building are actually useful. I'm not surprised by the results, because this hasn't been done rigorously before, so researchers have never had any idea whether their method has produced good explanations or not.

The study is weakened by the low sample size, which makes many of the p-values not significant. My intuition says a few more of the methods would produce statistically significant positive results in one of the domains/prediction settings if the sample size was bigger, but it seems like some settings (forward prediction, and textual data) are very hard to improve, with none of the methods getting a better improvement in performance than 5.7% (which had a p-value of 0.197).

A really interesting point is the lack of strong correlation between user-preference and performance improvement. This could be explained by the fact that most of the methods are ineffective at performance improvement, but it seems plausible (to me) that it could hold even if some methods were effective: If the model behaviour being explained can't be explained cleanly, then methods which do explain the behaviour might produce messy and confusing (but true) explanations and hence get lower ratings from users than methods which give clean and clear (but false) explanations. I

I think this stems from the problem of a lack of definition of what exactly the goal is for these interpretation methods. Without a goal in mind, it's impossible to measure whether the method achieves this goal. I think working towards some form of quantifiable measurement is useful particularly for comparing methods as, if this study's evidence is anything to go by, asking humans to evaluate the model's output might not be the most useful evaluation.

[Towards Interpretable Reinforcement Learning Using Attention Augmented Agents](#) (*Alexander Mott et al*) (summarized by Robert): In this paper the authors train a reinforcement learning agent with a soft attention module built into it. The attention module forms a bottleneck between the visual input and the network choosing the next action, which forces the model to learn to attend to only important parts of the scene. This means they can visualise which parts of the input the model thinks are important, as those are the parts of the input that the model is attending to. The queries to the attention model are determined by a top level recurrent network, without input from the current image, so act as a form of "top down" attention, where the top controller can be imagined to be querying the processed image for various locations and objects.

Having trained this agent (which still gets competitive performance with SOTA RL models on a fair few ATARI games), they qualitatively evaluate the attention visualisation on a variety of games. They find several common strategies in the attention schemes, such as the agents attending to specific points until an object crosses the point ("Tripwires"). The attention is computed over both regular pixels, as well as Fourier-based positional encoding. Thanks to this and other aspects of their architecture, the authors can check whether the queries are focused on pixel values (i.e. looking for a specific pattern of pixels anywhere) or on location features (i.e. asking what pixels are present at a specific location). For example, they find that the agent often queries the location where the score is displayed, presumably because it is useful for calculating the value function. They also compare their method with self-attention based models, and with other saliency methods.

The best way to get a feel for the visualisations is to go to the paper's website and watch the example videos.

Read more: [The paper's website](#)

Robert's opinion: This paper isn't revolutionary in its approach, but it's interesting to see work on interpreting RL agents, and the fact that the interpretability is built-in is interesting: it gives us a harder guarantee that this visualisation is actually showing us the parts of the input that the model thinks of as important, as they actually are important in its processing. It's promising to see that the in-built interpretability also didn't seem to penalise the performance much - it would be interesting to see this method applied to other, stronger kinds of models and see whether it still produces useful visualisations and how it affects their performance.

FIELD BUILDING

[AI Governance Career Paths for Europeans](#) (*Anonymous*) (summarized by Rohin): Exactly what it sounds like.

MISCELLANEOUS (ALIGNMENT)

[**A Guide to Writing the NeurIPS Impact Statement**](#) (*Carolyn Ashurst et al*) (summarized by Nicholas): NeurIPS 2020 requires paper submissions to include a statement on the broader impact of their work. This post provides a guide for how to write an effective impact statement. They recommend focusing on the most significant, neglected, and tractable impacts, both positive and negative, while also conveying the uncertainties involved. They also suggest integrating this into the research process by reading the tech governance literature and building institutional structures, and including this information in introductions.

Their guide then recommends considering 3 questions:

How could your research affect ML applications?

What are the societal implications of these applications?

What research or other initiatives could improve social outcomes?

There is more information in the guide on how to go about answering those questions, along with some examples.

Nicholas's opinion: I am definitely in favor of considering the impacts of ML research before conducting or publishing it. I think the field is currently either at or near a threshold where papers will start having significant real world effects. While I don't think this requirement will be sufficient for ensuring positive outcomes, I am glad NeurIPS is trying it out.

I think the article makes very strong points and will improve the quality of the impact statements that get submitted. I particularly liked the point about communicating uncertainty, which is a norm that I think the ML community would benefit from greatly. One thing I would add here is that giving explicit probabilities is often more helpful than vague words like "might" or "could".

OTHER PROGRESS IN AI

REINFORCEMENT LEARNING

[**"Other-Play" for Zero-Shot Coordination**](#) (*Hengyuan Hu et al*) (summarized by Rohin): How can we build AI systems that can *coordinate* with humans? While [past work](#) has assumed access to some amount of human data, this paper aims to coordinate *without any human data at all*, which they call *zero-shot coordination*. In order to develop an algorithm, they assume that their partner is also "trained" for zero-shot coordination.

Their key idea is that in zero-shot coordination, since you can't break symmetries by agreeing upon a protocol in advance (i.e. you can't agree on things like "we'll drive on the left, not the right"), you need a policy that is *robust to relabelings that preserve these symmetries*. This is easy to train for: you just train in self-play, but randomly relabel the states, actions and observations separately for each side in a way that preserves the MDP structure (i.e. uses one of the symmetries). Thus, each side must play a policy that works well *without knowing how the other agent's observations and actions have been relabeled*. In practice, for an N-player game you only need to

randomize N-1 of the relabelings, and so in the two player games they consider they only randomly relabel one side of the self-play.

They evaluate this in Hanabi (where the game is invariant to relabeling of the colors), and show that the resulting agents are better at playing with other agents trained on different seeds or with slightly different architectures, and also that they play better with humans, achieving an average score of 15.75 with non-expert human players, compared to 9.15 for agents trained via regular self-play.

Rohin's opinion: For comparison, I think I get around 17-22 when playing with new players, out of a max of 25, so 15.75 is quite a healthy score given that it doesn't use any human data. That being said, it seems hard to use this method in other settings -- even in the relatively simple [Overcooked environment \(AN #70\)](#), there aren't any obvious symmetries to use for such training. Perhaps future work will allow us to find approximate symmetries in games somehow, that we can then train to be robust to?

Towards Learning Multi-agent Negotiations via Self-Play (*Yichuan Charlie Tang*) (summarized by Rohin): While the previous paper introduces other-play to become robust to unknown partners, this paper takes the other approach of simply training an agent that is robust to a wide, diverse population of possible agents. In particular, it studies a self-driving car "zipper merge" environment, and trains an agent to be robust to a variety of rule-based agents, as well as past versions of itself, and finds that this leads to a much more successful merging policy. However, this is evaluated against the population it is trained with, and not against any previously unseen agents.

Building AI that can master complex cooperative games with hidden information (*Adam Lerer et al*) (summarized by Flo): This paper improves on the state of the art for AI agents playing [Hanabi \(AN #45\)](#), a cooperative multiplayer game that is challenging because of distributed hidden information and restricted communication.

The approach works by improving a baseline policy using search. In the simplest case, only one agent performs search while all other agents follow a fixed policy, such that the problem is reduced to search in a POMDP. This alone leads to relevant improvements, even when the search is very shallow. The fixed policies help because they allow the searching agent to correctly update its belief about hidden information when it sees other agents behaving (as it knows how other agents would behave given different states of the hidden information). This idea can be generalized to the case where all agents perform search by letting the agents simulate each other's search process. This can get expensive quickly as agent A's beliefs in the second round also depend on agent B's search process in counterfactual scenarios in the first round, such that agent B's search in round two also has to simulate these counterfactuals. A computation budget is introduced to make this computationally feasible and all agents know that the other agents will only use search in a turn if the cost of this is below the budget.

As search can be performed on top of any policy and allows to leverage compute during inference, not just training, it nicely complements more direct approaches using deep RL, which is a theme that has also been observed in Go and Poker.

Read more: [Paper: Improving Policies via Search in Cooperative Partially Observable Games](#)

Flo's opinion: This solution seems stunningly obvious in retrospect. While the authors informally report that their approach improves robustness to replacing other agents by humans, the example they give seems to indicate that this is because search prevents obvious mistakes in novel situations induced by human behaviour. Thus, I still expect (implicit) [human models \(AN #52\)](#) to be a vital component of human-machine cooperation.

DEEP LEARNING

[**Growing Neural Cellular Automata**](#) (*Alexander Mordvintsev et al*) (summarized by Zach): The process of an organism's shape development (morphogenesis) is an active area of research. One central problem is determining how cells decide how to grow and when to stop. One popular model for investigating this is Cellular Automata (CA). These model cells as living on a grid and interacting with each other via rules generated by looking at their nearest neighbors. The authors contribute to this research direction by introducing rule-sets that depend continuously on their local surroundings. The central insight connecting CA and deep learning is that because the rule-sets are constant the update rules work similarly to a convolutional filter. This allows the authors to take advantage of methods available to train neural networks to simulate CA. Using this insight, the authors train CA that can form into images that are resistant to perturbations and deletions. In other words, the CA are capable of regeneration.

Zach's opinion: The main relevance of an approach like this is that it provides proof-of-concept that complex goals, such as shape formation, can be programmed in an embarrassingly parallel fashion amenable to deep learning methodology. This naturally has implications in multi-agent settings where communication is expensive. I'd recommend checking out the main web app which allows you to watch and interact with the CA while they're growing. They also have a [code repository](#) that is easily adaptable to training on your own patterns. For example, I grew a regenerating Patrick Star [here](#).

META LEARNING

[**Gradient Surgery for Multi-Task Learning**](#) (*Tianhe Yu et al*) (summarized by Nicholas): In multi-task learning, an algorithm is given data from multiple tasks and tries to learn them all simultaneously, ideally sharing information across them. This paper identifies a *tragic triad* of conditions that can prevent gradient descent from finding a good minimum when all three are present:

Conflicting gradients occur when the gradient from one task points in a different direction from another.

Dominating gradients occur when the gradient from one task is much larger in magnitude than another.

High curvature is when the multi-task curvature is high in the direction of the gradient.

In this situation, the linear approximation of the gradient to the high curvature area leads to an overestimation of the increase in performance on the dominant gradient's task and an underestimation of the performance degradation from the conflicting

gradient's task. I find picturing the parabola $y=x^2$ and seeing that a gradient descent step overestimates progress while a gradient ascent step underestimates to be helpful in understanding this.

To solve this, they propose *PCGrad*, which projects all gradients into the normal plane of the others in a pairwise fashion. Their theoretical analysis establishes convergence properties of *PCGrad*, and they empirically show it can be combined with other multi-task algorithms to improve performance and that it makes optimization easier for multi-task supervised learning and RL. They also show plots confirming that the necessary conditions for their theorems appear in these contexts.

Nicholas's opinion: I like how this paper analyzes the loss landscape of a particular problem, multi-task learning, and uses that knowledge to derive a new algorithm. One thing I always find tricky in ML papers is that it is hard to establish that the theory of why an algorithm works (typically shown on toy models) is also the reason it improves performance (typically shown using complex neural networks). I appreciate that this paper checks for the conditions of their theorem in the multi-task RL models that they train. That said, I think that in order to confirm that the tragic triad they describe is the mechanism by which *PCGrad* improves performance, they would require some way to toggle each element of the triad while keeping everything else fixed.

FEEDBACK

I'm always happy to hear feedback; you can send it to me, [Rohin Shah](#), by **replying to this email**.

PODCAST

An audio podcast version of the [Alignment Newsletter](#) is available. This podcast is an audio version of the newsletter, recorded by [Robert Miles](#).

How can nonprofits gain the advantages of the for-profit model?

This is a linkpost for <https://rootsofprogress.org/how-nonprofits-can-gain-for-profit-advantages>

In my [last post](#) I described the advantages of for-profit models over nonprofit models, including scalable revenue, incentives and metrics to drive effectiveness and efficiency, and incentives to fund high-risk, high-reward experiments.

But not everything can be for-profit. How can nonprofit organizations get some of these advantages? Here are five ideas:

- Nonprofits that **generate revenue primarily through products & services**, rather than through charitable donations, gain some of the advantages of for-profits: they survive only to the extent that they can deliver a product to the market that people are willing to buy, out-compete alternatives, and keep their costs below their prices. To the extent that a nonprofit's paid services are subsidized by donations (as is the case with universities, museums, and opera houses, among others), this requirement is weakened but not destroyed.
- In the case of charity, I wonder if the most effective form of it is simply **giving money directly to beneficiaries**, with the goods and services themselves provided by for-profit businesses. This would seem to let free-market capitalism work to the maximum extent. There is [some research](#) to support this idea.
- If enough people **promote the idea of donating based primarily on demonstrated impact**, the world might slowly shift towards more [strategic nonprofits driven by output metrics](#) and other clear indication of delivered benefits. For instance, when reporting on a contribution from a major donor, the news media could focus more on the impact or potential impact of the contribution, rather than the amount of money given or the [percent of their wealth that represents](#).
- To drive innovation, perhaps we should be **putting more of our resources into prizes or mechanisms like advance market commitments**, rather than grants. Tyler Cowen [summarizes](#): "The case for prizes is stronger when you don't know who is likely to make the breakthrough, you value the final output more than the process, there is an urgency to solutions (talent development is too slow), success is relatively easy to define, and efforts and investments are likely to be undercompensated." It seems to me that most of those conditions apply to a lot of breakthrough scientific and technological R&D. Indeed, one of the earliest and most famous prizes, the Longitude Prize, had exactly the effect of uncovering an unexpected solution from an unlikely innovator: while most of the scientific community was looking for astronomy-based methods, John Harrison addressed the problem with a [highly robust and accurate clock](#)—and he wasn't even trained as a clockmaker. Why don't we have more prizes for grand challenge problems today?
- Beyond this, I think we need more mechanisms to **give credit for being right early**, for being the first backer of a risky experiment that has transformative effects. Who were the donors who gave small amounts of money to Howard Florey's lab around 1940 when they were inventing penicillin? The world should know their names. A special award or Hall of Fame could be created for these bold bets (perhaps with [a sophisticated scorekeeping mechanism](#)).

Our Need for Need

[I wrote this in 2017, well before I had a LessWrong account. The style isn't necessarily what I would have chosen if I'd known I was going to share it here, but I think the topic might be of some interest anyway, so I'm reposting it with light edits. Originally on [Grand, Unified, Crazy.](#).]

It is a trite, well-established truth that people like being useful. But there's more to it than that, or rather, there's also a stronger version of that claim. People do like being useful, but useful is a very broad term. Stocking shelves at a Walmart is useful, in that it's a thing with a use, which needs to be done. And it's true that some people may in fact actively like a job stocking shelves at a Walmart. But on the whole, it's not something most people would consider particularly enjoyable, and it's certainly not something that most people would consider "fulfilling".

Let us then upgrade the word "useful" to the word "needed": people like to be needed. While stocking shelves at a Walmart is useful, the person doing it is fundamentally replaceable. There are millions of others around the world perfectly capable of doing the same job, and there are probably thousands of them just within the immediate town or city. If our fictional stocker were to suddenly vanish one day, management would have no trouble hiring somebody else to fill their shoes. The world would go on. Walmart would survive.

Now this is all well and good, but I would argue that there is an even stronger version of this claim: people don't just like to be needed, people actively *need* to be needed. Over a decade ago, Paul Graham wrote an essay called [Why Nerds are Unpopular](#); it's a long essay with a number of different points, but there is one thread running through it that in my opinion has gotten far too little attention: "[Teenagers'] craziness is the craziness of the idle everywhere".

The important thing to note about this (and Graham does so, in a roundabout sort of way) is that teenagers in a modern western high school are not exactly *idle*. They have class, and homework, and soccer practice or band practice or chess club; they play games and listen to music and do all the sort of things that teenagers do. They just don't have a purpose. They are literally unneeded, shut away in a brick building memorizing facts they'll probably never use, mostly to get them out of the way of the adults doing real work.

This obviously feels bad, and Graham stops there, making the assumption that the adult world at least, has enough purpose to go around. Teenagers, and in particular nerds, just have to wait until they're allowed into the real world and voila, life will sort itself out. And it's true that for some, this is the case. A scientist doing groundbreaking research doesn't need to worry about their purpose; they know that the work they are doing is needed, and has the potential to change lives. Unfortunately, a Walmart stocker does not.

To anyone who has been following the broad path of the news over the last few decades, this probably doesn't come as a surprise. It seems like every other day we are confronted by another article suggesting that people are becoming [less happy](#) and [more depressed](#), and that [modern technology is making people unhappy](#). Occasionally it is also noted that [this is weird](#). We live in a world of wealth and plenty. The poorest

among us are healthier, better-fed, and more secure than the richest of kings only a few centuries past. What is causing this malaise?

The simple answer is that we are making ourselves obsolete. People need to be needed, sure, but nobody wants to *need*. [Independence is the American dream](#), chased and prized throughout the modern world. Needing someone else is seen as weakness, as vulnerability, and so we strive to be self-sufficient, to protect ourselves from the possibility of being hurt. But in doing so, we hurt others. We take from them our need, and leave them more alone than ever before.

Of course, Western independence as a philosophy has been growing for near on three centuries now, and modern unhappiness seems like a much more recent phenomenon. There are two reasons for this, one obvious and the other a bit more subtle. To start with, our modern wealth does count for something. A small amount of social decohesion can trade off against an entire industrial revolution's worth of progress and security with no alarm bells going off. But there is a deeper trick at play, and that is specialization.

In traditional hunter-gatherer bands, generally everybody was needed. The tribe could usually survive the loss of a few members of course – it had to – but not easily. Every member had a job, a purpose, a needed skill. That there were only a handful of needed skills really didn't matter; there just weren't that many people in any given tribe.

As civilization flourished, the number of people in a given community grew exponentially. Tribes of hundreds were replaced by cities of thousands, and for a time this was OK. Certainly, there was no room in a city of thousands for half the adult men to be hunters; it was both ecologically and sociologically unsustainable. But in a city of that size there was suddenly room for tailors and coopers and cobblers and masons and a million other specialized jobs that let humanity preserve this sense of being needed. If it was fine to be one of the handful of hunters providing food for your tribe, it was just as fine to be one of the handful of cobblers providing shoes for your town.

To a certain extent, specialization continued to scale right through the mid-twentieth century, just not as well. In addition to coopers and masons we also (or instead) got engineers and architects, chemists and botanists, marketers and economists. But somewhere in the late twentieth century, that process peaked. Specialization still adds the occasional occupation (e.g. software developer), but much more frequently modern technology takes them away instead. Automation lets one person do the work of thousands.

Even worse than this trend is the growth of the so-called “global village”. I, personally, am a software developer in a city of roughly one million people. Software development is highly specialized, and arguably the most modern profession in the world. At the end of the day however, I too am replaceable. Even if I were only one of the handful of developers in my city (I'm not), modern technology – both airplanes and the internet – has broadened the potential search pool for my replacement to nearly the entire world. My position is fundamentally no different from that of the Walmart stocker – I would not be missed.

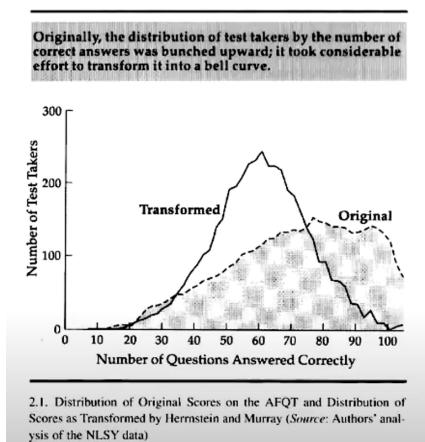
At the end of the day, humanity is coming to the cross-roads of our need for need. Obsessed with individuality, we refuse to depend on anyone. Women's liberation has mostly freed nearly half of the world's population from economic dependence. Technological progress, automation, and global travel are all nibbling away at the

number of specialized occupations, and at the replacement cost of the ones that remain. The future is one where we all live like the teenagers in Paul Graham's essay: neurotic lapdogs, striving to find meaning where fundamentally none exists. Teenagers, at least, just have to grow up so they can find meaning in the real world.

How is humanity going to grow up?

Why aren't we testing general intelligence distribution?

IQ supposedly measures general intelligence. Assuming the g-factor exists, why are we using IQ to measure it? IQ tests get tweaked so the end result will always form a normal distribution. I checked wikipedia to figure out why, but while it said some people claim that LCT is a good reason to believe the g-factor will have a normal distribution, it didn't seem terribly confident (and it also didn't give a source). Even "[The Bell Curve](#)" used an intelligence test that didn't return a bell curve at all. They said they had to turn it into a bell curve to prevent 'skew':



But isn't this putting the cart before the horse? Surely it would be interesting to see what distributions we would get if we did intelligence testing without a result in mind? I get that it is really difficult to measure, but if a test consistently gave us a different distribution, we could learn valuable insights into how human minds work. So is there a reason why we aren't testing the hypothesis that general intelligence will follow a normal distribution?

Pointing to a Flower

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

Imagine I have a highly detailed low-level simulation (e.g. molecular dynamics) of a garden. The initial conditions include a flower, and I would like to write some code to “point” to that particular flower. At any given time, I should be able to use this code to do things like:

- compute a bounding box around the flower
- render a picture which shows just the flower, with the background removed
- list all of the particles which are currently inside the flower

Meanwhile, it should be robust to things like:

- most of the molecules in the flower turning over on a regular basis
- the flower moving around in space and/or relative to other flowers
- the flower growing, including blooming/wilting/other large morphological change
- other flowers looking similar

That said, there’s a limit to what we can expect; our code can just return an error if e.g. the flower has died and rotted away and there is no distinguishable flower left. In short: we want this code to capture roughly the same notion of “this flower” that a human would.

We’ll allow an external user to draw a boundary around the flower in the initial conditions, just to define which object we’re talking about. But after that, our code should be able to robustly keep track of our particular flower.

How could we write that code, even in principle?

“Why Not Just... ”

There’s a lot of obvious hackish ways to answer the question - and obvious problems/counterexamples for each of them. I’ll list a few here, since the counterexamples make good test cases for our eventual answer, and illustrate just how involved the human concept of a flower is.

- Flower = molecules inside the flower-boundary at time zero. Problem: most of the molecules comprising a flower turn over on a regular basis.
- Flower = whatever’s inside the boundary which defined the flower at time zero. Counterexample: the flower might move.
- Flower = things which look (in a rendered image) like whatever was inside the boundary at time zero. Counterexample: the flower might bloom/wilt/etc. Another counterexample: there may be other, similar-looking flowers.
- Flower = instance of a recurring pattern in the data, defined by clustering. Counterexample: there may not be any other flowers. (More generally: we can recognize “weird” objects in the world which don’t resemble anything else we’ve ever seen.)
- Flower = region of high density contiguous in space-time with our initial region. Counterexample: we can dunk the flower in a bucket of water.
- Flower = contents of lipid bilayer membranes which also contain DNA sequence roughly identical to the consensus sequence of all DNA within the initial boundary, plus anything within a few microns of those membranes. Counterexample: it’s still the same flower if we blow it up via [expansion microscopy](#) and the individual cells lyse in the process. (Also this wouldn’t generalize to non-biological objects, or even clonal organisms.)

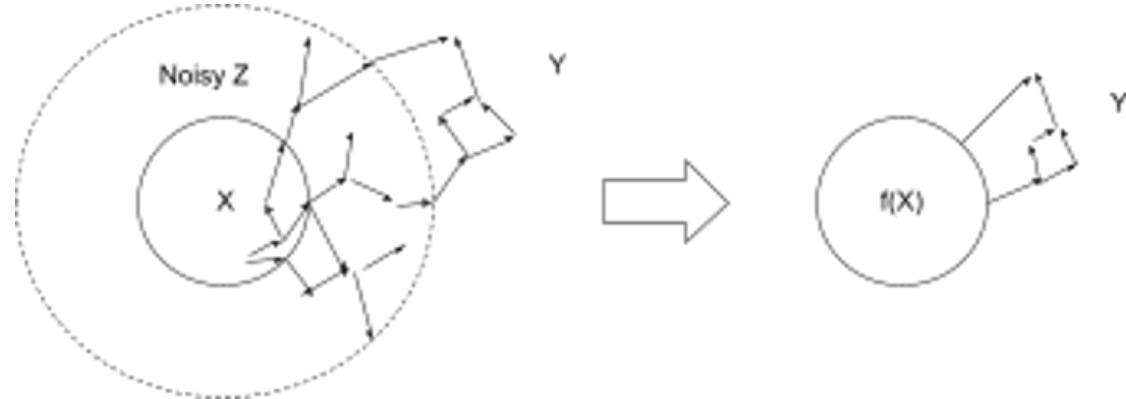
Drawing Abstract Object Boundaries

The general conceptual challenge here is how to define an abstract object - an object which is not an ontologically fundamental component of the world, but an abstraction on top of the low-level world.

In [previous posts](#) I've outlined a fairly-general definition of abstraction: far-apart components of a low-level model are independent given some high-level summary data. We imagine breaking our low-level system variables into three subsets:

- Variables X which we want to abstract
- Variables Y which are "far away" from X
- Noisy "in-between" variables Z which moderate the interaction between X and Y

The noise in Z wipes out most of the information in X, so the only information from X which is relevant to Y is some summary $f(X)$.



(I've sketched this as a causal DAG for concreteness, which is how I usually visualize it.) I want to claim that this is basically the right way to think about abstraction quite generally - so it better apply to questions like "what's an abstract object?".

So what happens if we apply this picture directly to the flower problem?

First, we need to divide up our low-level variables into the flower (X), things far away from the flower (Y), and everything in-between (noisy Z). I'll just sketch this as the flower itself and a box showing the boundary between "nearby" and "far away":

$t = 0$



$t = 1$



$t = 2$

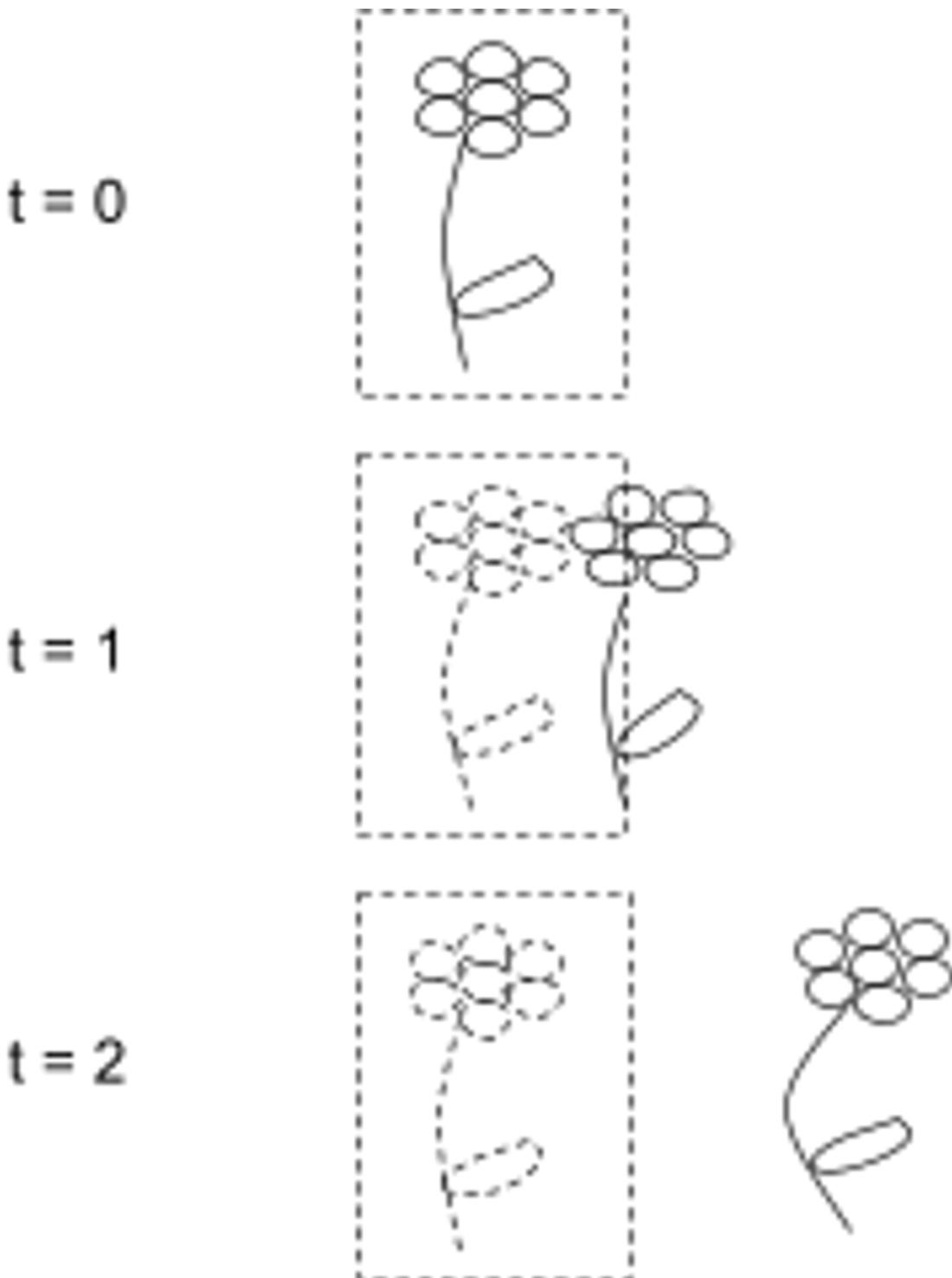


Notice the timesteps in the diagram - both the flower and the box are defined over time, so we imagine the boundaries living in four-dimensional spacetime, not just at one time. (Our

user-drawn boundary in the initial condition constrains the full spacetime boundary at time zero.)

Now the big question is: how do we decide where to draw the boundaries? Why draw boundaries which follow around the actual flower, rather than meandering randomly around?

Let's think about what the high-level summary $f(X)$ looks like for boundaries which follow the flower, compared to boundaries which *start* around the flower (i.e. at the user-defined initial boundary) but don't follow it as it moves. In particular, we'll consider what information about the *initial* flower (i.e. flower at time zero) needs to be included in $f(X)$.



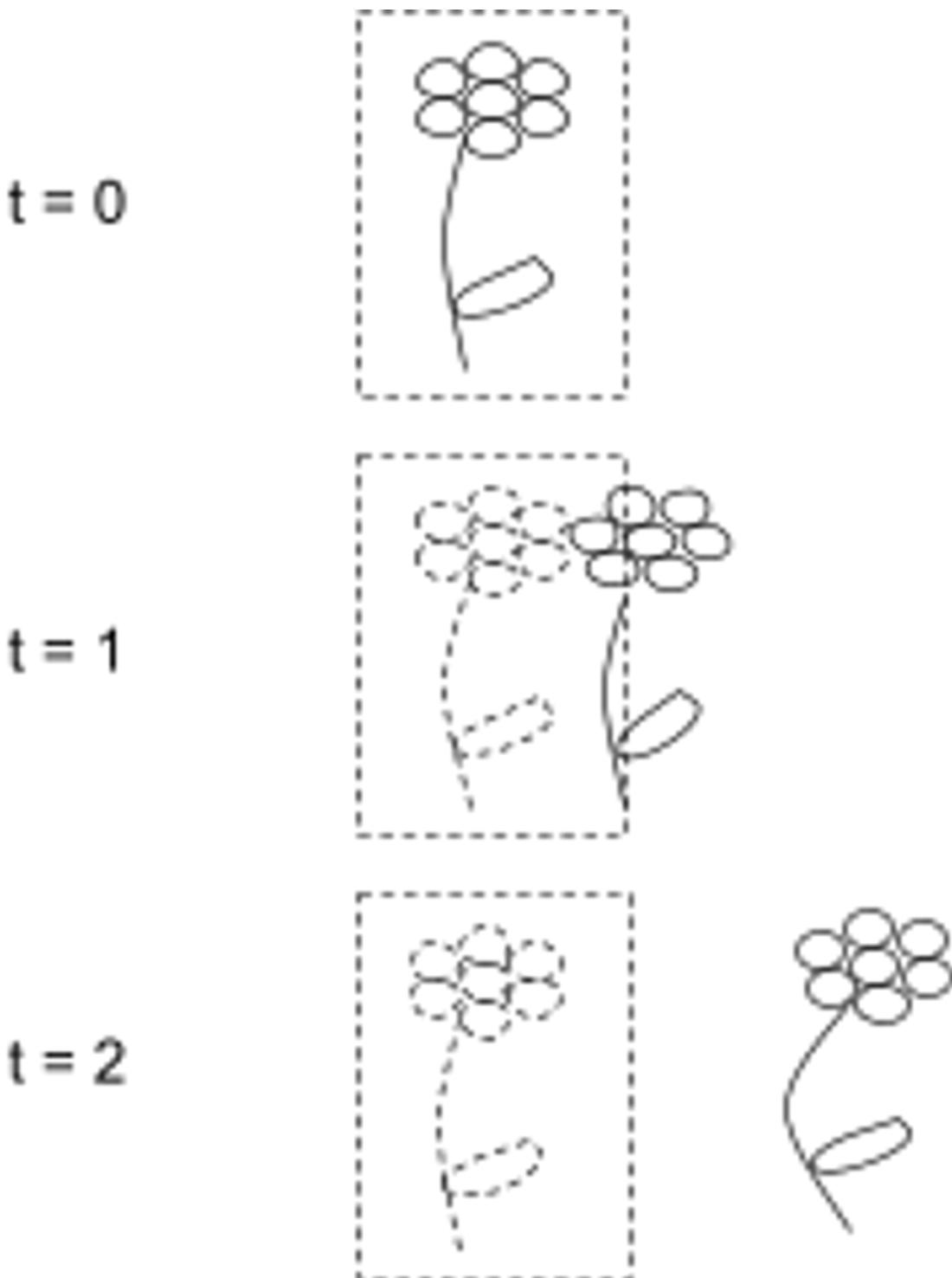
The “true” flower moves, but the boundaries supposedly defining the “flower” don’t follow it. What makes such boundaries “worse” than boundaries which do follow the flower?

There's a lot of information about the initial flower which *could* be included in our summary $f(X)$: the geometry of the flower's outer surface, its color and texture, temperature at each point, mechanical stiffness at each point, internal organ structure (e.g. veins), relative position of each cell, relative position of each molecule, ... Which of these need to be

included in the summary data for boundaries moving with the flower, and which need to be included in the summary data for boundaries not moving with the flower?

For example: the flower's surface geometry will have an influence on things outside the outer boundary in both cases. It will affect things like drag on air currents, trajectories of insects or raindrops, and of course the flower-image formed on the retina of anyone looking at it. So the outer surface geometry will be included in the summary $f(X)$ in both cases. On the other hand, relative positions of cells inside the flower itself are mostly invisible from far away *if the boundary follows the flower*.

But if the boundary doesn't follow the flower... then the true flower is inside the boundary at the initial time, but counts as "far away" at a later time. And the relative positions of individual cells in the true flower will mostly stay stable over time, so those relative cell positions at time zero contain lots of information about relative cell positions at time two... and since the cells at time two counts as "far away", that means we need to include all that information in our summary $f(X)$.



Strong correlation between low-level details (e.g. relative positions of individual cells) inside the spacetime boundary and outside. That information must be included in the high-level summary $f(X)$.

The takeaway from this argument is: **if the boundary doesn't follow the true flower, then our high-level summary $f(X)$ must contain far more information.** Specifically, it has to include tons of information about the low-level internal structure of the flower. On the

other hand, as long as the true flower remains inside the inner boundary, information about that low-level structure will mostly not propagate outside the outer boundary - such fine-grained detail will usually be wiped out by the noisy variables “nearby” the flower.

This suggests a formalizable approach: the “true flower” is defined by a boundary which is locally-minimal with respect to the summary data $f(X)$ required to capture all its mutual information with “far-away” variables.

Test Cases

Before we start really attacking this approach, let’s revisit the problems/counterexamples from the hackish approaches:

- Molecular turnover: not a problem. The relevant information does not follow the individual molecules.
- Flower might move: not a problem. We basically discussed that directly in the previous section.
- Flower might bloom/wilt/etc: not a problem. Mutual information still follows the same pattern, although note that once the flower rots away altogether, we can draw a time-boundary indicating that the flower no longer exists, and indeed we expect everything significantly after that in time to be roughly independent of our former flower.
- Similar-looking flowers: not a problem. We’re explicitly relying on the low-level internal structure to define the flower boundary.
- No other flowers: not a problem. We’re not relying on clustering or any other data from other flowers.
- Dunk flower in a bucket of water: not a problem. Noisy water molecules “nearby” the flower will wipe out low-level detailed information about as well as noisy air molecules, if not better.
- Expansion microscopy: not a problem. The information in the flower’s low-level structure sticks around in its expanded form. Indeed, expansion microscopy wouldn’t be very useful otherwise.

Main takeaway: this approach is mainly about information contained in the low-level structure of the flower (i.e. cells, organs, etc). Physical interactions which maintain that low-level structure will generally maintain the flower-boundary - and a physical interaction which destroys most of a flower’s low-level structure is generally something we’d interpret as destroying the flower.

Problems

Let’s start with the obvious: though it’s formalizable, this isn’t exactly formalized. We don’t have an actual test-case following around a flower in-silico, and given how complicated that simulation would be, we’re unlikely to have such a test case soon. That said, next section will give a computationally simpler test-case which preserves most of the conceptual challenges of the flower problem.

First, though, let’s look at a few conceptual problems.

What about perfect determinism?

This approach relies on high mutual information between true-flower-at-time-zero and true-flower-at-later-times. That requires some kind of uncertainty or randomness.

There’s a lot places for that to come from:

- We could have ontologically-basic randomness, e.g. quantum noise

- We could have deterministic dynamics but random initial conditions
- More realistically, we could have some sort of observer in the system with Bayesian uncertainty about the low-level details of the world.

That last is the “obvious” answer, in some sense, and it’s a good answer for many purposes. I’m still not completely satisfied with it, though - it seems like a superintelligence with extremely precise knowledge of every molecule in a flower should still be able to use the flower-abstraction, even in a completely deterministic world.

Why/how would a “flower”-abstraction make sense under perfect determinism? What notion of locality is even present in such a system? When I probe my intuition, my main answer is: causality. I’m imagining a world without noise, but that world still has a causal structure similar to our world, and it’s that causal structure which makes the “flower” make sense.

Indeed, causal abstraction allows us to apply the ideas above directly to a deterministic world. The only change is that $f(X)$ no longer only summarizes probabilistic information; it must also summarize any information needed to predict far-away variables under *interventions* (on either internal or far-away variables).

Of course, in practice, we’ll probably also want to include those interventional-information constraints even in the presence of uncertainty.

What about fine-grained information carried by, like, microwaves or something?

If we just imagine a physical outer boundary some distance from a flower (let’s say 3 meters), surely some clever physicists could figure out a way to map out the flower’s internal structure without crossing within that boundary. Isn’t information about the low-level structure constantly propagating outward via microwaves or something, without being wiped out by noisy air molecules on the way?

Two key things to keep in mind here:

- The boundary need not be a *physical* boundary; the “boundaries” just denote subsets of the variables of the model. If the model includes microwaves, we can just declare them all to be “nearby” the flower. Whenever they actually interact with molecules outside the flower, barring instruments specifically set up to detect them, the information they carry should be wiped out quite quickly by statistical-mechanical noise.
- In practice, we don’t just want to abstract *one* object. We want a whole high-level world model, full of abstract objects. The “far-away variables” will be variables within all the other high-level objects. So in order for microwaves to matter, they need to carry information from one object to another, without that information being wiped out by low-level noise.

Note that we’re talking about noise a lot here - does this problem play well with deterministic universes, where causality constrains $f(X)$ more than plain old information? I expect the answer is yes - chaos makes low-level interventions look basically like noise for our purposes. But that’s another very hand-wavy answer.

What if we draw a boundary which follows around every individual particle which interacts with the flower?

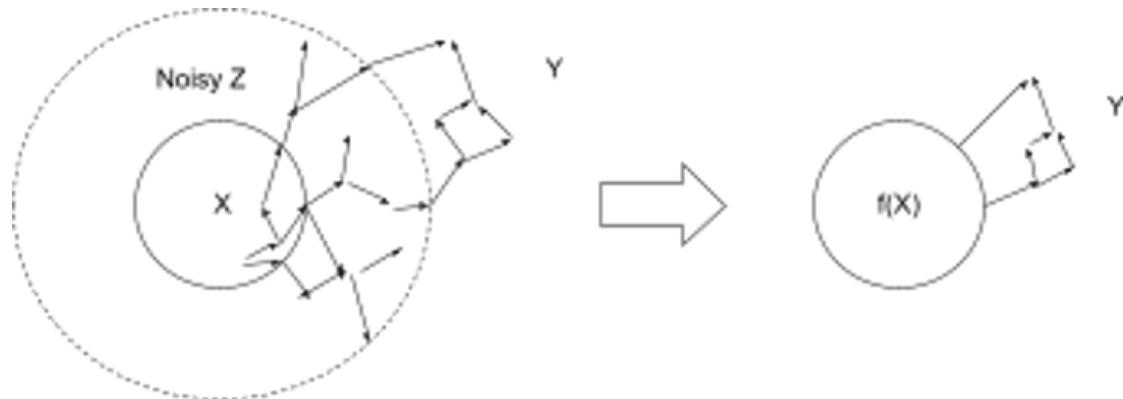
Presumably we could get even less information in $f(X)$ by choosing some weird boundary. The easy way to solve this is to add boundary complexity to the information contained in $f(X)$ when judging how “good” a boundary is.

Humans seem to use a flower-abstraction without actually knowing the low-level flower-structure.

Key point: we don't need to *know* the low-level flower-structure in order to use this approach. We just need to have a model of the world which says that the flower has *some* (potentially unknown) low-level structure, and that the low-level structure of flower-at-time-zero is highly correlated with the low-level structure of flower-at-later-times.

Indeed, when I look at a flower outside my apartment, I don't know its low-level details. But I do expect that, for instance, the topology of the veins in that flower is roughly the same today as it was yesterday.

In fact, we can go a step further: humans lack-of-knowledge of the low-level structure of particular flowers is one of the main reasons we should expect our abstractions to look roughly like the picture above. Why? Well, let's go back to the original picture from the definition:



Key thing to notice: since Y is independent of all the low-level details of X except the information contained in $f(X)$, $f(X)$ contains everything we can possibly learn about X just by looking at Y .

In terms of flowers: our “high-level summary data” $f(X)$ contains precisely the things we can figure out about the flower without pulling out a microscope or cutting it open or otherwise getting “closer” to the flower.

Testable Case?

Finally, let's outline a way to test this out more rigorously.

We'd like some abstract object which we can simulate at a “low-level” at reasonable computational cost. It should exhibit some of the properties relevant to our conceptual test-cases from earlier: components which turn over, moves around, change shape/appearance, might be many or just one, etc. Just those first two properties - components which turn over and object moving around - immediately suggest a natural choice: a wave.

- In a particle view, the underlying particles comprising the wave change over time
- The wave moves around in space and relative to other waves
- The wave may change shape (due to obstacles, dissipation, nonlinearity, etc)
- There may be other similar-looking waves in the environment or no other waves

I'd be interested to hear if this sounds to people like a sensible/fair test of the concept.

Summary

We want to define abstract objects - objects which are not ontologically fundamental components of the world, but are instead abstractions on top of a low-level world. In particular, our problem asks to track a particular flower within a molecular-level simulation of a garden. Our method should be robust to the sorts of things a human notion of a flower is robust to: molecules turning over, flower moving around, changing appearance, etc.

We can do that with a [suitable notion of abstraction](#): we have summary data $f(X)$ of some low-level variables X , such that $f(X)$ contains all the information relevant to variables “far away”. We’ve argued that, if we choose X to include precisely the low-level variables which are physically inside the flower, and mostly use physical distance to define “far-away” (modulo microwaves and the like), then we’d expect the information-content of $f(X)$ to be locally minimal. Varying our choice of X subject to the same initial conditions - i.e. moving the supposed flower-boundary away from the true flower - requires $f(X)$ to contain more information about the low-level structure of the flower.

