Best of LessWrong: October 2020

1. Is Success the Enemy of Freedom? (Full)
2. Why indoor lighting is hard to get right and how to fix it
3. The Treacherous Path to Rationality
4. The Solomonoff Prior is Malign
5. Introduction to Cartesian Frames
6. The Felt Sense: What, Why and How
7. What are some beautiful, rationalist artworks?
8. Message Length
9. The bads of ads
10. The Darwin Game - Rounds 0 to 10
11. Philosophy of Therapy
12. The Alignment Problem: Machine Learning and Human Values
13. The date of AI Takeover is not the day the AI takes over
14. Postmortem to Petrov Day, 2020
15. Covid 10/1: The Long Haul
16. The Darwin Game
17. Can we hold intellectuals to similar public standards as athletes?
18. A prior for technological discontinuities
19. What should experienced rationalists know?
20. Babble & Prune Thoughts
21. Weird Things About Money
22. Desperation hamster wheels
23. Babble challenge: 50 ways of sending something to the moon
24. Is Stupidity Expanding? Some Hypotheses.
25. How to Find the Frontiers of Knowledge
26. Security Mindset and Takeoff Speeds
27. What posts do you want written?
28. Box inversion hypothesis
29. Words and Implications
30. Lessons on Value of Information From Civ
31. Rule of Equal and Opposite Advice & Slack
32. The Rise and Fall of American Growth: A summary
33. Moloch games
34. Have the lockdowns been worth it?
35. Dutch-Booking CDT: Revised Argument
36. Thoughts on ADHD
37. AI race considerations in a report by the U.S. House Committee on Armed Services
38. Book Review: Reinforcement Learning by Sutton and Barto
39. Group debugging guidelines & thoughts
40. Brainstorming positive visions of AI
41. Draft papers for REALab and Decoupled Approval on tampering
42. Things are allowed to be good and bad at the same time
43. Babble challenge: 50 ways of solving a problem in your life
44. Additive Operations on Cartesian Frames
45. AGI safety from first principles: Conclusion
46. The Colliding Exponentials of AI
47. AGI safety from first principles: Control
48. "Scaling Laws for Autoregressive Generative Modeling", Henighan et al 2020 {OA}
49. The Darwin Game - Round 1
50. What's holding back outsourcing to cloud labs?

Best of LessWrong: October 2020

1. Is Success the Enemy of Freedom? (Full)
2. Why indoor lighting is hard to get right and how to fix it
3. The Treacherous Path to Rationality
4. The Solomonoff Prior is Malign
5. Introduction to Cartesian Frames
6. The Felt Sense: What, Why and How
7. What are some beautiful, rationalist artworks?
8. Message Length
9. The bads of ads
10. The Darwin Game - Rounds 0 to 10
11. Philosophy of Therapy
12. The Alignment Problem: Machine Learning and Human Values
13. The date of AI Takeover is not the day the AI takes over
14. Postmortem to Petrov Day, 2020
15. Covid 10/1: The Long Haul
16. The Darwin Game
17. Can we hold intellectuals to similar public standards as athletes?
18. A prior for technological discontinuities
19. What should experienced rationalists know?
20. Babble & Prune Thoughts
21. Weird Things About Money
22. Desperation hamster wheels
23. Babble challenge: 50 ways of sending something to the moon
24. Is Stupidity Expanding? Some Hypotheses.
25. How to Find the Frontiers of Knowledge
26. Security Mindset and Takeoff Speeds
27. What posts do you want written?
28. Box inversion hypothesis
29. Words and Implications
30. Lessons on Value of Information From Civ
31. Rule of Equal and Opposite Advice & Slack
32. The Rise and Fall of American Growth: A summary
33. Moloch games
34. Have the lockdowns been worth it?
35. Dutch-Booking CDT: Revised Argument
36. Thoughts on ADHD
37. AI race considerations in a report by the U.S. House Committee on Armed Services
38. Book Review: Reinforcement Learning by Sutton and Barto
39. Group debugging guidelines & thoughts
40. Brainstorming positive visions of AI
41. Draft papers for REALab and Decoupled Approval on tampering
42. Things are allowed to be good and bad at the same time
43. Babble challenge: 50 ways of solving a problem in your life
44. Additive Operations on Cartesian Frames
45. AGI safety from first principles: Conclusion
46. The Colliding Exponentials of AI
47. AGI safety from first principles: Control

48. "Scaling Laws for Autoregressive Generative Modeling", Henighan et al 2020 {OA}
49. The Darwin Game - Round 1
50. What's holding back outsourcing to cloud labs?

Is Success the Enemy of Freedom? (Full)

This is a linkpost for https://radimentary.wordpress.com/2020/10/26/is-success-the-enemy-of-freedom-full/

I. Parables

A. Anna is a graduate student studying p-adic quasicoherent topology. It's a niche subfield of mathematics where Anna feels comfortable working on neat little problems with the small handful of researchers interested in this topic. Last year, Anna stumbled upon a connection between her pet problem and algebraic matroid theory, solving a big open conjecture in the matroid Langlands program. Initially, she was over the moon about the awards and the Quanta articles, but now that things have returned to normal, her advisor is pressuring her to continue working with the matroid theorists with their massive NSF grants and real-world applications. Anna hasn't had time to think about p-adic quasicoherent topology in months.

B. Ben is one of the top Tetris players in the world, infamous for his signature move: the reverse double T-spin. Ben spent years perfecting this move, which requires lightning fast reflexes and nerves of steel, and has won dozens of tournaments on its back. Recently, Ben felt like his other Tetris skills needed work and tried to play online without using his signature move, but was greeted by a long string of losses: the Tetris servers kept matching him with the other top players in the world, who absolutely stomped him. Discouraged, Ben gave up on the endeavor and went back to practicing the reverse double T-spin.

C. Clara was just promoted to be the youngest Engineering Director at a mid-sized software startup. She quickly climbed the ranks, thanks to her amazing knowledge of all things object-oriented and her excellent communication skills. These days, she finds her schedule packed with what the company needs: back-to-back high-level strategy meetings preparing for the optics of the next product launch, instead of what she loves: rewriting whole codebases in Haskell++.

D. Deborah started her writing career as a small-time crime novelist, who split her time between a colorful cast of sleuthy protagonists. One day, her spunky children's character Detective Dolly blew up in popularity due to a Fruit Loops advertising campaign. At the beginning of every month, Deborah tells herself she's going to finally kill off Dolly and get to work on that grand historical romance she's been dreaming about. At the end of every month, Deborah's husband comes home with the mortgage bills for their expensive bayside mansion, paid for with "Dolly money," and Deborah starts yet another Elementary School Enigma.

E. While checking his email in the wee hours of the morning, Professor Evan Evanson notices an appealing seminar announcement: "A Gentle Introduction to P-adic Quasicoherent Topology (Part the First)." Ever since being exposed to the topic in his undergraduate matroid theory class, Evan has always wanted to learn more. He arrives bright and early on the day of the seminar and finds a prime seat, but as others file into the lecture hall, he's greeted by a mortifying realization: it's a graduate student learning seminar, and he's the only faculty member present. Squeezing in his

embarrassment, Evan sits through the talk and learns quite a bit of fascinating new mathematics. For some reason, even though he enjoyed the experience, Evan never comes back for Part the Second.

F. Whenever Frank looks back to his college years, he remembers most fondly the day he was kicked out of the conservative school newspaper for penning a provocative piece about jailing all billionaires. Although he was a mediocre student with a medium-sized drinking problem, on that day Frank felt like a *man with principles*. A real American patriot in the ranks of Patrick Henry or Thomas Jefferson. After college, Frank met a girl who helped him sort himself out and get sober, and now he's the proud owner of a small accounting firm and has two beautiful daughters Jenny and Taylor. Yesterday, arsonists set fire to the Planned Parenthood clinic across the street, and his employees have been clamoring for Frank to make a political statement. Frank almost threw caution to the wind and Tweeted #bodilyautonomy from the company account right there, but then the picture on his desk catches his eye: his wife and daughters at Taylor's elementary school graduation. It's hard to be a man of principles when you have something to lose.

G. Garrett is a popular radio psychologist who has been pressured by his sponsors into being the face of the yearly Breast Cancer Bike-a-thon. Unfortunately, Garrett has a dark secret: he's never ridden a bicycle. Too embarrassed to ask anyone for help or even be seen practicing – he is a respected public figure, for god's sake – Garrett buys a bike and sneaks to an abandoned lot to practice by himself after sunset. He thinks to himself, "how hard can it be?" Garrett shatters his ankle ten minutes into his covert practice session and has to pull out of the event. Fortunately, Garrett's sponsors find an actual celebrity to fill in for him and breast cancer donations reach record highs.

II. Motivation

What is personal success for?

We say success opens doors. Broadens horizons. Pushes the envelope. Shatters glass ceilings.

Success sets you free.

But what if it doesn't?

Take a good hard look at the successful people around you. Doctors too busy to see their children on weekdays. Mathematicians too brilliant in one field to switch to another. Businessmen too wealthy to avoid nightly wining and dining. Professional gamers too specialized to learn a new hero. Public figures too popular to change their minds.

Remember that time Michael Jordan took a break from basketball and played professional baseball? They said he would have made an excellent professional player given time. Jordan said baseball was his childhood dream. Even so, in just over a year Jordan was back in basketball. It is hard not to imagine what a baseball player Michael Jordan could have been, *had he been less successful going in*.

I think it was in college that I first noticed *something wasn't right* about this picture. I spent my first semester studying and playing Go for about eight hours a day. I remember setting out a goban on the carpet of my dorm room and studying patterns

in the morning as my roommate left for classes; when he returned to the room in the evening, he was surprised to see me still sitting there contemplating the flow of the stones. Because this was not the first or tenth time this had happened, he commented something like, “You must be really smart to not need to study.”

I remember being dumbstruck by that statement. It suggested that my freedom to play board games for eight hours a day was gated by my personal success, and other Harvard students would be able to live like me *if only they were smarter.* But you know who else can play board games for eight hours a day? Basement-dwelling high school dropouts, who are – for all their unsung virtues – *definitely not smarter than Harvard students*.

When I entered college, they told me a Harvard education would empower me do anything I want. The world would be my oyster. I took that message to heart in those four years – I fell in love, played every PC game that money could buy, studied programming languages and systems programming, and read more than one Russian novel. When I talked to my peers, however, I was constantly surprised at the overwhelming sameness of their ambitions. Four years later, twenty out of thirty-odd graduating seniors at our House planned to work in finance or consulting.

(Now, it could be that college really empowers these bright young scholars to realize their childhood dreams of arbitraging the yen against the kroner. But this is, as they say in the natural sciences, *definitely not the null hypothesis.*)

All of this would have made a teenager hate the idea of success altogether. I was not a teenager anymore, so I formulated a slightly more sophisticated answer: *Regardless of how successful I become, I resolve to live like a failure.*

This is a post about all the forces, real and imagined, that can make success the enemy of personal freedom. As long as these forces exist, and as long as human heart yearns for liberty, few people will ever want wholeheartedly to succeed. Were it not already reality, that is a state of affairs too depressing to contemplate.

(Just to be clear, people are plenty motivated to succeed when basic needs are at stake – to put food on the table, to get laid, to pay for the mortgage. But after those needs get met, success just doesn’t look all that great and only certain sorts of delightful weirdos keep striving. The rest of us mostly just lay back and enjoy the fruits of their labor.)

III. Factorization

I think all of the experiences in Section I can be summed up by the umbrella-term “Sunk Cost Fallacy,” but that theory is a little too low-resolution for my tastes. In this section I identify three main psychological factors of the phenomenon.

1. You rose to meet the challenge. Your peer group rose to meet you.

We are constantly sorted together with people of the same age group, at similar levels of competence, at similar stages in our careers. To keep up with the group, you have to run as fast as you can just to stay in place, as the saying goes. And if you run twice as fast as that, you just end up in a new, even harder-to-impress peer group. When your friends are all level 80, it’s dreadfully difficult to restart at level 1.

Your friends may even be sympathetic, but it rarely helps matters.

Maybe you want to try something totally new, and your friends are too invested in their pet genre to emigrate with you.

Maybe you're excited to learn a new skill one of your hyper-competent friends is specialized in, and you ask them to coach you. Unfortunately, this turns out to be a massive mistake, because your friend only remembers how she got from level 75 to 80, and sort of assumes everything below is trivial. It's technically possible to learn area formulas as a special case of integral calculus, but only technically.

Maybe you transition to a new role within the team, you struggle to learn a new set of tricks, and you start hating yourself for not pulling as much weight as you're used to. You start to see a mix of pity and frustration in your teammates eyes as you drag the whole team down.

2. Yesterday, you were bad at everything, and that really sucked. Today you're good at one thing, and you're hanging on for dear life.

It's hard to move out of your comfort zone when your comfort zone is one hundred square feet on top of Mount Olympus and every cardinal direction points straight off a cliff. Seems like just yesterday you stood at the base of this mountain among the rest of the mortals, craning your neck to get a peek at what it's like up here.

Kindly god-uncle Zeus calls a special thunderstorm for your arrival. Dionysus pours you a frothy drink and shares a bawdy tale. Hephaestus personally fashions you a blade as a symbol of your newfound status. Aphrodite invites you to her parlor for a night of good old-fashioned philosophy. They all act so welcoming, so natural, so in their element, and *you know you're only up here by a stroke of pure luck.*

When Hermes returns the next morning and invites you to fly with him on his winged boots to see the world, you decline graciously. Not because you don't want to – they're winged boots! – but because the moment you try anything out of the ordinary *you'll be found out for the impostor that you are* and god-uncle Zeus will show you his not-so-kindly side and chain you to a liver-eating eagle or a boulder that only obeys the laws of gravity intermittently.

3. Success gave you something to lose.

They say beware the man with nothing to lose.

I say envy him, because he alone is free.

You fondly recall the good old days of two thousand and two when you could go online and post diatribes against religion as a "militant atheist." In those days, you had nothing, and you were free. You were unattached. You were intellectually wealthy but financially insolvent. You could see one end of the place you call home from the other.

Now that you've made it big, you'd have to carefully position mirrors at the ends of three hallways to see that far. You're attached to wonderful person(s) of amenable sexual orientation(s). You have a reputation to maintain in the ever-smaller circles that you walk. Children in your community look up to you, or so you tell yourself. And so, even though deep in your heart you still believe that *only idiots believe in an old man in the sky* your Twitter profile identifies you as "spiritual, yearning, exploring."

IV. Resolution?

It seems to me we have a problem.

We are not a species known for risk-taking, so human flourishing really depends on the explicit emphasis of exploration and openness to new experience. And yet it seems that the game is set up so that the most successful people are least incentivized to explore further. That all the trying new things and pushing boundaries and calling for revolution is likely to come from those with neither the power to get it done nor the competence to do it correctly.

But it's not a hopeless case by any means. Many of the most successful people got there precisely by valuing freedom, creativity, and exploration, and still practice these values – so far as they can – within the confines of their walled gardens. We live in an information age where getting good at things is as easy as it's ever been. And at very least we pay lip service to healthy adages like "Stay hungry, stay foolish."

But what does one do personally to maintain one's freedom?

I don't claim to have a fully general solution to this problem, but here is a rule that's helped me in the past.

When learning something new, treat yourself like a five-year-old.

If you've never spoken a word of Korean in your life, it doesn't matter if you're a professor of English Literature. As far as learning Korean goes, you're a five-year-old. Treat yourself like one. Make yourself a snack for memorizing the vertical vowels. Take a break after reading your first sentence and come back tomorrow. When you're done for the day, suck your thumb while staring at the first Korean word you've ever learned and feel the honest pride well up in your heart.

If you've never washed a dish in your life, it doesn't matter if you're a professional chef. As far as washing dishes is concerned, you're a five-year-old. Treat yourself like one. Make yourself a snack for figuring out how to dispense dish soap without getting it everywhere. Take a break after finishing the bowls and come back tomorrow. When you're all done, take a moment to take in that beautiful empty sink and feel the honest pride well up in your toddler heart.

Do you see how profoundly counterproductive it would be for the Korean learner to beat herself up for not being able to converse fluently with her Asian friends after two weeks? Do you see how completely unkind it would be for the novice dishwasher to call himself a useless piece of shit for not being able to execute the most basic of adult tasks?

Be kind to yourself and adjust your expectations to reality. When learning something new, treat yourself like a five-year-old.

Why indoor lighting is hard to get right and how to fix it

The days are getting shorter in the northern hemisphere, and with the ongoing pandemic, most of us expect to be spending more time in our homes than normal. Creating a healthy home environment is more important than usual, and the light inside your home is an often underappreciated part of this.

There has already been some explicit discussion[1] about the importance of lighting for health and productivity, as well as many mentions of it in other places. Nonetheless, based on discussions I've had recently within the community, I get the impression that it is helpful for me to write up the results and tinkering that have done over the past few years.

First, I will cover some of the research on how our bodies respond to light, and which particular characteristics of natural light we want to mimic. Then I will explain solving this problem is hard and my overall strategy for solving it. Finally, I will give some specific advice on what to buy and how to arrange things.

I give quite a lot of background before offering any specific advice. Although I think the background information might help you make good decisions, you should feel free to skip the next section if you're in a hurry or if it seems uninteresting.

Background

Note: My background is in optics, not physiology or psychology and I began researching and writing this document almost four years ago. My original draft, as well as many of my sources, have been lost in the intervening years, so what you're seeing here is based on a combination of my notes that survived, my recollection of the research, and a partial duplication of the research. To make matters worse, it does seem that new research has come along since I began this project, so this is likely out of date. My guess is that most or all of the practical conclusions still stand, but I am only moderately confident of this. As much as I would like to take the time to update the research, past experience suggests that I will never actually publish it if I try to put too much more work into it. I welcome corrections, and if there is sufficient enthusiasm around this topic, I may try to write an updated version.

Your body uses light to synchronize its internal clock and to modulate your mood and alertness[2]. While the particulars of the lighting in your environment are important, your only perceptual access to information about lighting in the moment comes from your visual system, which is poorly adapted to solving the problem of determining the intensity and spectrum of a light source. This is mainly because our vision is optimized more for accurately identifying materials, textures, colors, and other properties of our surroundings than it is for knowing details about sources of light. This has the consequence that some of our default intuitions about the nature of ambient light are wrong, so when we're building our lighting environment, it can be difficult to make accurate judgments just by looking at things. We can do better if we use quantitative measures and our scientific understanding of how things work to solve the problem.

Physiology

In addition to visual photoreceptors that are used for seeing things, your eyes contain non-visual photoreceptors which serve non-visual functions. Melanopsin is a photopigment that is found in cells in the retina[3]. It is sensitive to blue light, and when activated, these cells send signals that help with things such as regulating our internal clocks[4]. Unlike the our visual photoreceptors, which are more densely packed in the center of our retina[5] than in the periphery, these photosensitive cells are distributed relatively evenly throughout the retina[6], so that light coming from both the periphery and the center of our vision is important. Although I have found contradictory accounts of how visual photoreceptors and melanopsin-sensitive cells interact, there does seem to be some evidence that activation of our visual photoreceptors can influence the functioning of our melanopsin-containing cells. This seems consistent with many people's experience that blue LED lighting is more relaxing than white light. On the other hand, it may just be that it takes a lot of light to activate these cells[7].

I find the literature on how our mood and circadian rhythm respond to light exposure throughout the day to be somewhat confusing, and I have not tried to unravel it in detail. As far as I can tell, it depends on the full history of exposure over a course of days, not something simple like "total light exposure over the past N hours" or "Exponentially-weighted average of light exposure". For this reason, I am not particularly optimistic about figuring out the details of the algorithm used by our bodies to decide when to be tired or when to be happy, or any other aspect of our lives that we might want to improve.

Given this, as well as my priors in favor of just letting our bodies have the regulatory mechanisms they're "designed" for, it seems that we should try to create lighting that is similar to natural light and well-synchronized to our preferred schedules for sleeping, working, and relaxing. This means that, in order to signal to our bodies that it is daytime, and therefore time to be wakeful, we should use light that is bright, white, and distributed over our full field of view.

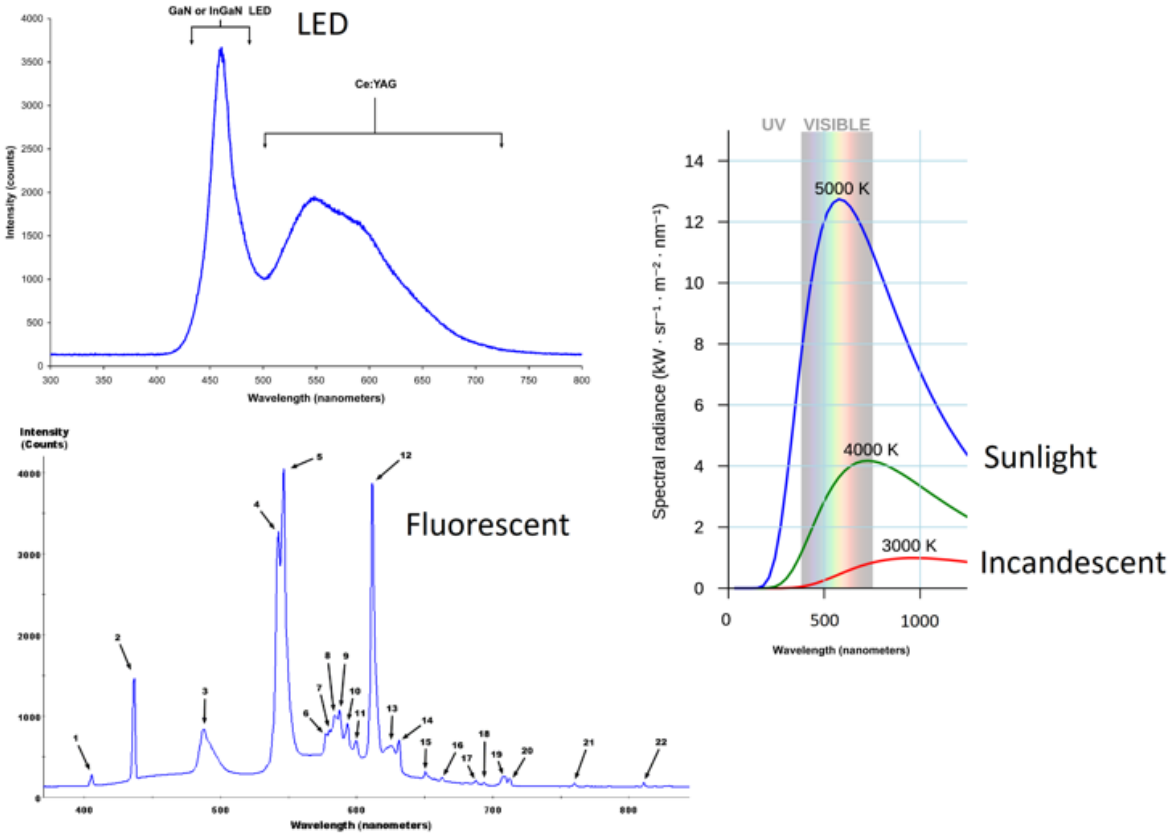
When we want to signal to our bodies that it is time to start winding down for the day, we want to avoid bright, white light. I have not done as much research on the best timing for switching over or the optimum spectrum for this, but based on the fact that it is dark at night, sunsets are yellow, and on my personal experience and that of people I know, there should be several hours between bright daytime lighting and bedtime, and the difference between the best lighting for the late evening and the daytime is quite large, both in terms of intensity and color.

Natural vs Artificial Light

Creating light that effectively mimics natural light is difficult. The first problem is that natural light is almost always much brighter than artificial light. Even on a very cloudy day, you might notice that the insides of buildings look dark from the outside, the brightest parts of most indoor spaces are next to windows, and if you use a camera or a light meter to measure the amount of light, you'll find that things are more brightly illuminated outdoors than indoors.

The second problem is that natural light has a broad, continuous spectrum, which contains only a few gaps within the visual range of wavelengths, while artificial light almost always has large portions missing, especially at shorter wavelengths. This has

to do with how artificial light is generated. Incandescent bulbs are blackbody sources, but they're not hot enough to contain very much light at shorter wavelengths. Incandescent bulbs are usually around 3000K while the sun is 5800K. Fluorescent lamps and LEDs generate UV and blue light, which is then absorbed and re-emitted at longer wavelengths by phosphors. This technique produces light that can have a correlated color temperature (CCT) anywhere from 2000K or less to over 6000K[8]. Unfortunately, this results in a very spiky spectrum in the case of fluorescent bulbs and a large dip between blue-ish and yellow-ish wavelenths for most LEDs. Here are some spectra from Wikipedia[9]:



All this missing light can be easy to overlook. Our visual system projects all the spectral content onto three dimensions (or two, depending on how you look at it), and then does a remarkable job of giving us accurate information about the color of objects in a variety of lighting conditions (the blue-black/white-gold dress phenomenon is remarkable because of our failure to reliably solve this problem), which is great for many tasks, but terrible for evaluating the quality of a source of light.

Nonetheless, this missing light is important. In addition to failing to provide the appropriate signals for time of day, poorly filled out spectra can make things look unappealing, give us inaccurate information about the colors of objects, and make an environment feel generally less pleasant.

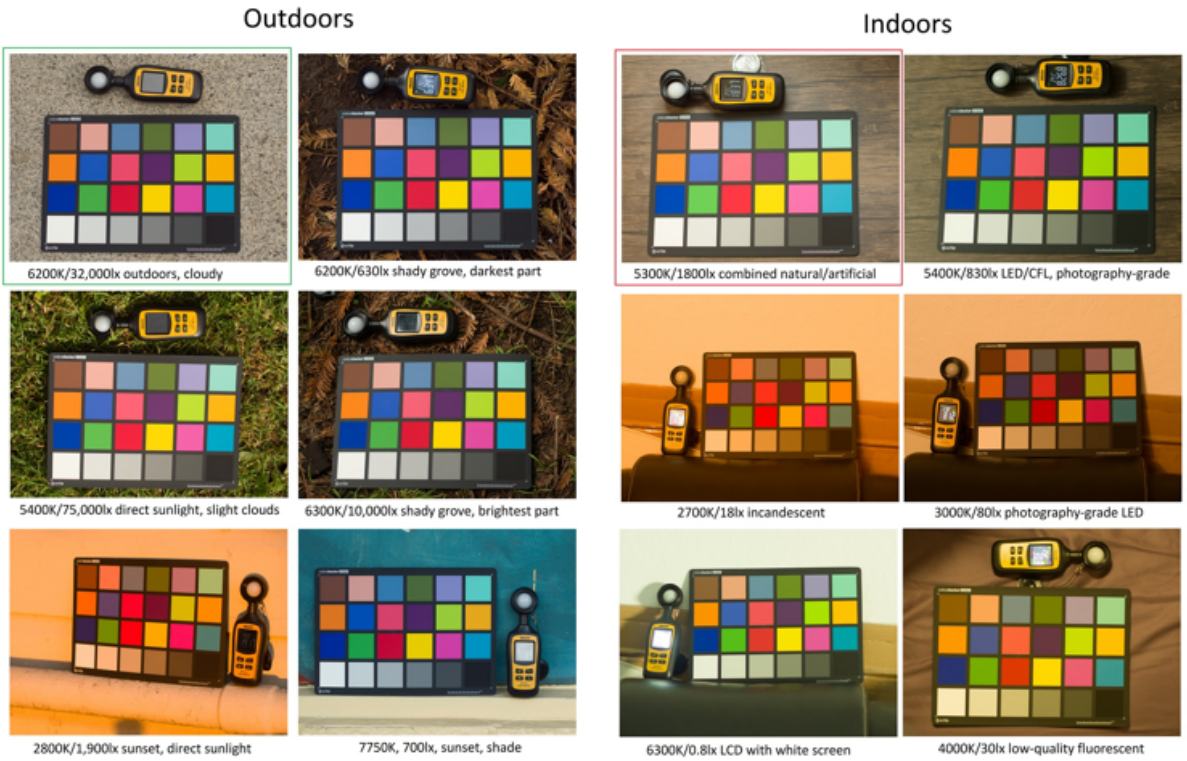
What I am not trying to replicate (for now)

There are some properties of natural light that are less important or more difficult to replicate. Although these might be desirable in certain contexts, we'll ignore them for now.

1. Non-visible light: We're not trying to add in ultraviolet or infrared light. Although getting some UV might be good for people with a vitamin D deficiency, doing this safely and without other consequences (like degrading plastics or bleaching colors in your office or bedroom) seems hard, and I don't recommend trying to do this.
2. Spectrum and intensity that changes continuously throughout the day. There are lamps that do this, and they seem especially pleasant to me for waking up in the morning, but I think this is particularly hard to get right and less important than other things
3. All the little details. Natural light has lots of subtleties that might contribute to making it seem pleasant. For example, the sky is blue and polarized and direct sunlight is a bit yellow. The anisotropy of the color of the light makes for pleasant photographs. Things like clouds and leaves can make soft, moving shadows. Again, it would be neat to replicate this, but it seems hard to get right and I think it's a much lower priority than other things.

Why this is a hard problem to solve

To illustrate how much your visual system can correct for variations in lighting, I took photos and illuminance measurements of a photographic color calibration card in a variety of environments, shown here:



The camera and software are calibrated to the photo in the top-left with the green border around it, which I took during a typical cloudy day in San Francisco, so the lighting is reasonably close to a 6500 kelvin blackbody spectrum. I tried to get the brightness of each image the same, so the apparent brightness of the display on the lux meter should give you a sense of how much brighter some of the environments are than others.

The yellowest lighting in these photos has a red-to-blue ratio 15 times higher than the bluest lighting, and the intensity varies by five orders of magnitude. Nonetheless, the card mostly looked the same to me in nearly every context. I was aware that the incandescent bulb was dimmer and yellower than the light from the cloudy sky, but the effect is nowhere near as extreme as the photos suggest. The two main exceptions to this were the LCD photo, for which the light was very dim (though I did not try allowing my eyes to adjust), and the low-quality fluorescent lamp, in which many of the colors just didn't look quite right.

I hope this will give you a sense for why we should not rely on our visual systems to tell us how bright or how blue a light source is. Fortunately, we can measure these things in other ways.

General approach to solving the problem

Given everything that we have so far, we want light that is:

1. As bright as is practical during the day, preferably at least 1000 lux over the full room, with a CCT close to that of the sun (5500K)
2. Much dimmer in the evening, with a much lower CCT (2700K)
3. Covers all or most of our visual field
4. Has a full spectrum, with few holes at visible wavelengths
5. Does not have other annoying characteristics like sharp shadows, flickering, or the generation of unnecessary heat or noise

Natural light meets all of these criteria, except in some circumstances being annoying (direct sunlight in your eyes) or being too bright in the evening (if you live very far from the equator). I always try to work near a window if possible, and get as much natural light as I can. That said, at the moment I'm writing this, it has been dark outside for over an hour and it's only 6:25pm. The whole reason this is a problem to be solved is because we can't just use natural light all the time.

Getting enough light

As I have mentioned already, getting your indoor daytime light to be anywhere near as bright as outdoor lighting takes effort. If you install some lighting, and you still experience seasonal depression, the first thing I would try is adding more light. Most people will be limited by how many bulbs they can reasonably install in a room, either due to space, power, or heat, before they'll have too much light. Or they'll have a bunch of lights and think "Surely this is enough! Look at how many bulbs I have!", but until it is sufficiently bright to make your computer screen look dim, it is probably not as bright as natural light. In my small living room, I find that my setup which provides 20,000 lumens, and gives me a relatively uniform 1000 lux[10] is sufficient. I wouldn't

mind more light (and I intend to add more once I'm in a more permanent living situation), but it is enough for me to feel alert during the day and avoid or at least reduce the effects of seasonal depression, so long as I'm able to spend time in there later in the day. For a larger room, you'll need more light, in a way that should scale roughly with the inside area of the walls, ceiling, and floor.

Getting the right spectrum

The spectrum of your daytime light is the trickiest part, but I have found two strategies that seem to work:

1. Combining lots of different bulbs so they can fill in each other's gaps.
2. Finding bulbs with a very high color rendering index, especially those marketed for photography.

For a while during grad school, my office had, in addition to the fluorescent tubes installed in the ceiling fixtures, three different kinds of LED bulbs and two different kinds of very high lumen compact fluorescent (CFL) bulbs. A disadvantage of variety is that you can't just buy a huge pack of all one kind of bulb, which is usually the most convenient and inexpensive option. An advantage of ordering many different bulbs is that if you get a particular one that you don't like, you only bought a few of them (for example, I had some which made an annoying buzzing sound and others that had an annoying green tinge). For high CRI bulbs marketed to photographers look for CRI of *at least* 85-90. A safer bet is to go with 95+. Once you know what you like, when you need to replace bulbs or build a lighting system in a new room, you be able to get it right more quickly.

For daytime color temperature, I recommend 5300-6000K. Many people dislike the light from high color temperature light bulbs, but I think this is because there is a lot of terrible indoor lighting out there, including light from bulbs that have a high color temperature but poor color rendering. A CCT of 6000K is very close to the color temperature of natural daylight through a window. If you get it right, people might even mistake the light coming out of your room for daylight! Still, some people seem to do alright with significantly lower color temperature. I've heard of some people doing well with as low as 4000K.

Evening light should be less than 3000K. Incandescent and LED bulbs both work well without much need to combine bulbs to fill in gaps at these color temperatures. If you do find low CCT bulbs to be unpleasant somehow, you might try either incandescent bulbs or high CRI bulbs.

My preferred light for the last hour before bedtime is an orange LED bulb with so little blue light that I can't tell blue objects apart from black.

Covering your full visual field

It might be tempting to get a therapy lamp or to get fewer bulbs and just put them in the corner of the room where you're sitting all the time. But this can be hard to get right. For starters, remember that the photoreceptors you're trying to activate cover your entire retina, so you want to illuminate your full visual field if possible. This covers a very wide angle of 210 degrees, which extends slightly behind you. Once you account for turning your head some of the time, this can easily extend to well over half the

room. Another problem, at least for me personally, is feeling confined to one part of the room. This is more of a problem at home than it is at my office. When possible, I recommend just illuminating the whole room.

Still, this may not be an option for everyone. If I'm working in a shared space, I don't want to impose my preferences on everyone else, nor do I want to spend hundreds of dollars on lights that I'll only use for a couple months. Sometimes I have lights arranged just to illuminate my own part of the shared space. Similarly, some people use therapy lamps in these settings.

Evening and nighttime lighting

Evening lighting is easier. The main difficulty is in having multiple sets of lights for different parts of the day. Some people will only need daytime lights for their workplace. I use inexpensive low CCT LED and fluorescent bulbs, usually 2700K, and I switch from my bright daytime lights to those at 7pm. In my bedroom, I also have an orange LED bulb that is very dim, but sufficient for reading, and I switch over to this for the last 30-60 min before going to sleep.

The little things

I said I'm not trying to solve all of the problems with reproducing natural light indoors, but there are a few small things that can get it a little closer, and which can be nice.

Lightbulbs can cast annoying, sharp shadows, and if they are in your visual field, they can be annoyingly bright. Diffusers help with this. Most light fixtures have a lampshade or some other device to increase the effective size of the source, softening shadows and reducing the intensity of the light fixture itself. If you want diffusers that are spectrally neutral (that is, they do not absorb some colors more than others), you can find those in photography stores.

Having a bit of variety in the spectrum can be nice. Outdoors on a sunny day, shadows are blueish, and reflected sunlight is yellowish, for example. Because of this, even if you had bulbs with a perfect 6000K blackbody spectrum, it may still seem unnatural. One thing that can help with this is to add in some lower color temperature lights. Personally, I find that one ~3000K bulb per three ~5500K bulbs is very pleasant and less artificial-feeling than just 5500K[11].

Specific recommendations on what to buy

Since links for buying things online are only useful geographically and for a limited time, I will recommend brands that I have been happy with and explain my overall strategy, along with links to specific products.

Lux meter

Being able to measure how much light is in your house is useful and inexpensive. I use a $30 Uceri meter that I ordered from Amazon.

Light bulbs

I have found that fluorescent bulbs that are designed for photography are usually a good bet for color rendering. In particular, I have been using these very large bulbs from Alzo (note that these bulbs are huge and fragile, will not fit in most fixtures, and probably should not be used with strings of light sockets). For LED bulbs, I have been reasonably happy with bulbs from Alzo and Cree. Phillips and GE make bulbs that I have been happy with, but they also make bulbs that I have been quite disappointed with, so be careful.

For LEDs, I know one person[12] who has had success with Yuji corncob bulbs, for both high and low color temperature. He was kind enough to give me one of the 3200K bulbs, which I find to create very pleasant light, and which I now mix in with my 5500K bulbs during the daytime.

As I mentioned earlier, evening bulbs are generally easier, since high CRI, low CCT bulbs are simpler to make. Any incandescent should have a nice spectrum, and most LEDs with high CRI should be okay.

Bedtime bulbs

I've been satisfied with my orange LED bulb from Sunlite. One person I know prefers red bulbs over orange. I do not recommend incandescent bulbs with red/orange filters, as they are less efficient, get hot, and have a filament that still looks white-ish through the filter, which find I to be annoying.

Fixtures

Installing these bulbs can be a bit tough, especially the gigantic CFLs. Usually you'll want to have them above eye level, or at least have a diffuser. They need to be somewhere that they don't get too hot, and where they cast light in a way that illuminates the room evenly.

If you want to build your own fixtures with neutral diffusers, I purchased a couple of generic diffuser "socks" from Amazon several years ago, and I've been happy with them. They are no longer available, but if you want to see what I'm talking about they're here.

If you have many fixtures or fixtures in inconvenient locations, you may want to get a remote like this one from IKEA. I would imagine that a smart lighting setup could help, but I have not tried one myself.

Photography stands

I like using photography lighting stands because they're inexpensive and tall enough to get the bulbs well-above eye level. If you're worried about them being ugly, they can probably be decorated (I intend to add some fake ivy to mine, for example). I use these from Amazon

DIY fixtures

It is also easy to build something that can sit on top of a shelf. One way to do this is to use cable ties to attach a power strip to a wooden board, and stick socket adapters in the power strip. Unforunately, I cannot share a photo of the last one that I built, because the pandemic ate it.

Wrapping up

I hope this has been helpful. The most important things to remember are:

1. More light is good, and it is difficult to have too much light.
2. Try to cover your full visual field
3. The spectrum of your light matters a lot, and during the day, you probably want >5000K
4. During the evening, you probably want <3000K
5. If you normally find high color temperature (bluish) light to be unpleasant, try using higher quality lighting. Mixing bulbs or using stuff made for photographers helps with color rendering and overall pleasantness

I welcome any comments on your experiences! I know that others have spent time researching and experimenting, and many have probably done a better job of tracking down or constructing good lighting equipment than I have. I am also happy to add links to other people's write-ups on this.

Acknowledgements: Many people helped me along the way while I was researching and writing this. In particular, thanks to my former lab mates at UT Austin for putting up with my lighting experiments, Meredith Johnson for her encouragement to start the project in 2016, Katja Grace for reminding me that I should just post stuff instead of worrying so much about whether it is bad, and everyone that I've shared notes with over the past few years, especially those at the FHI office.

---

1. See:

   https://www.lesswrong.com/posts/Ag7oQifJQM5AnMCrR/my-simple-hack-for-increased-alertness-and-improved

   https://www.lesswrong.com/posts/BTXdajWzoN2YRbgjG/rational-health-optimization

   https://www.benkuhn.net/lux/

   http://www.lincolnquirk.com/2019/11/26/lumenator.html

   https://ryan.abel.space/blog/adventures-in-interior-lighting ↩

2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6751071/ ↩

3. The retina is the photosensitive part of the eye, analogous to a CCD or CMOS sensor in a camera ↩

4. https://science.sciencemag.org/content/295/5557/1065 ↩

5. https://en.wikipedia.org/wiki/Fovea_centralis ↩

6. See fig 1E: https://science.sciencemag.org/content/295/5557/1065 This diagram seems to suggest there are actually *more* melanopsin-sensitive cells in some parts of the periphery of our visual field. It is plausible to me that you could look at this map, and determine where to put lights in your house, for example, ensuring that there is more light in the upper portion of your visual field, but I have not taken the time to decipher that diagram into such a map of where the illumination should be in our visual field. ↩

7. A potentially stupid and probably not very informative experiment I did once: I used to work in a lab where I had to wear laser safety glasses all the time. These glasses effectively cut out all of the blue light, in addition to severely reducing overall light transmission. Wearing these all day is really bad for people who suffer from seasonal depression, so I tried attaching some blue LEDs to the temples of the glasses so that the light would reflect off the inner surface of the lenses and into my eyes. When I tried them in the lab, the first thing I noticed was that the lab seemed way less depressing than usual. After maybe 30 minutes, I noticed that I was feeling somewhat less drowsy and less compulsion to go outside than I often did in the lab. Unfortunately, the lights gave me a terrible headache, so I never used them enough to see if this was a real effect or not.

   Note that although melanopsin-sensitive cells are referred to as "non-visual" there does seem to be more recent research suggesting that melanopsin may play a role in visual perception as well. ↩

8. Remember that this is the temperature of a blackbody source that is most similar to the the light source in question. Be careful, because the spectrum is often very different from a blackbody! ↩

9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6751071/ has a nice diagram showing spectra for various light sources ↩

10. Lumens are a unit for luminous intensity (the amount of light per time, weighted by how bright it appears to humans)

    Lux is lumens per area, or how brightly-lit a surface is, according to human vision. ↩

11. H/T to Ben Weinstein-Raun for bringing this to my attention and giving me a bulb to try it out ↩

12. These are the bulbs that Ben recommended to me which I found useful for adding some more yellow light to my daytime lighting. ↩

The Treacherous Path to Rationality

Cross-posted, as always, from Putanumonit.

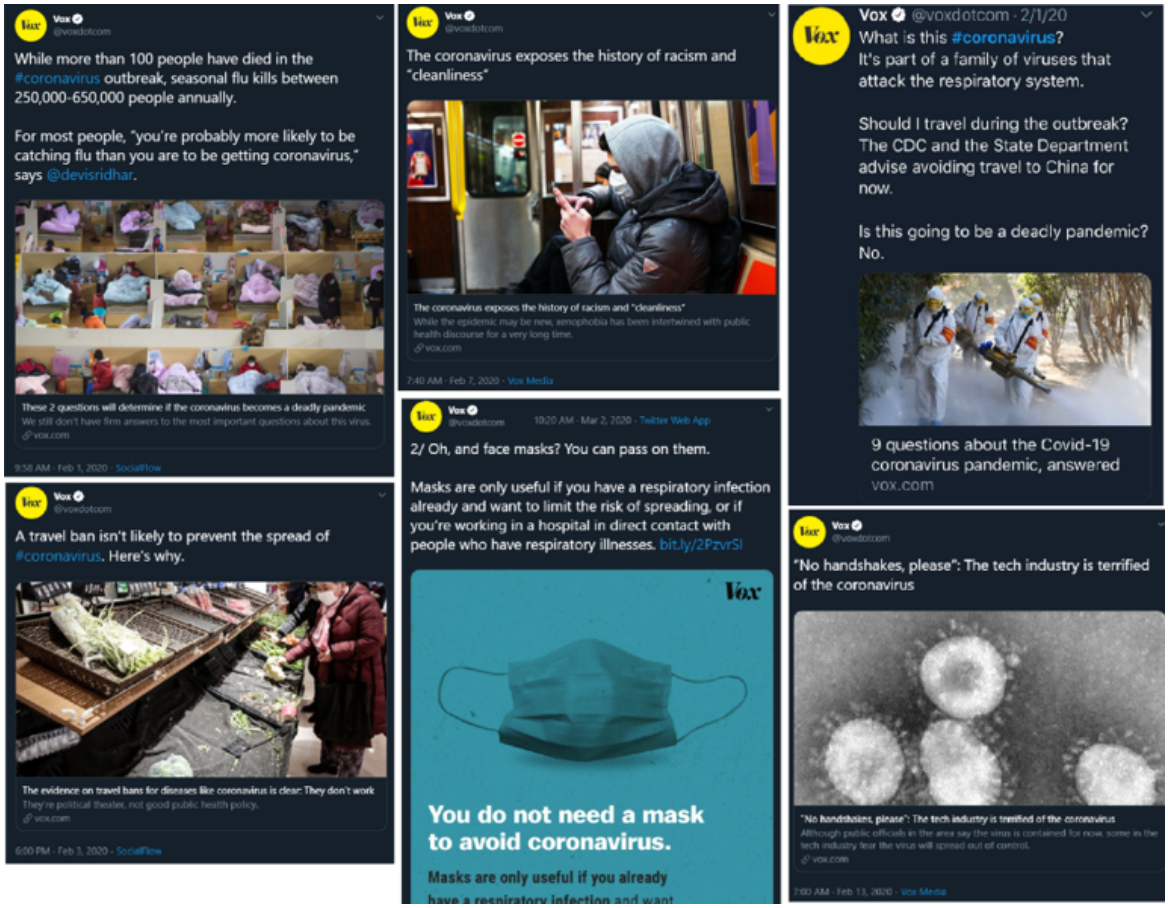
Rats v. Plague

The Rationality community was never particularly focused on medicine or epidemiology. And yet, we basically got everything about COVID-19 right and did so months ahead of the majority of government officials, journalists, and supposed experts.

We started discussing the virus and raising the alarm in private back in January. By late February, as American health officials were almost unanimously downplaying the threat, we wrote posts on taking the disease seriously , buying masks, and preparing for quarantine .

Throughout March, the CDC was telling people not to wear masks and not to get tested unless displaying symptoms. At the same time, Rationalists were already covering every relevant angle, from asymptomatic transmission to the effect of viral load, to the credibility of the CDC itself. As despair and confusion reigned everywhere into the summer, Rationalists built online dashboards modeling nationwide responses and personal activity risk to let both governments and individuals make informed decisions.

This remarkable success did not go unnoticed. Before he threatened to doxx Scott Alexander and triggered a shitstorm, New York Times reporter Cade Metz interviewed me and other Rationalists mostly about how we were ahead of the curve on COVID and what others can learn from us. I told him that Rationality has a simple message: *"people can use **explicit reason** to figure things out, but they rarely do"*



If rationalists led the way in covering COVID-19, Vox brought up the rear

Rationalists have been working to promote the application of explicit reason, to "raise the sanity waterline" as it were, but with limited success. I wrote recently about success stories of rationalist improvement but I don't think it inspired a rush to LessWrong. This post is in a way a response to my previous one. It's about the obstacles preventing people from training and succeeding in the use of explicit reason, impediments I faced myself and saw others stumble over or turn back from. This post is a lot less sanguine about the sanity waterline's prospects.

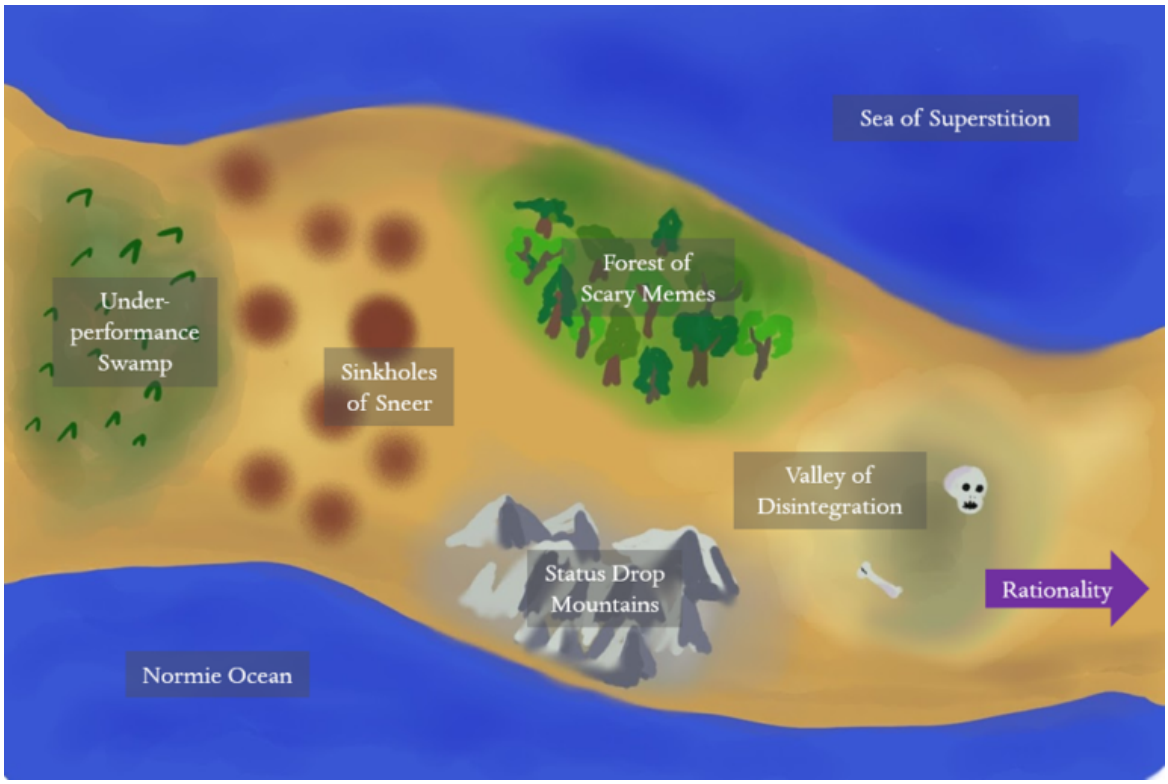
The Path

I recently chatted with Spencer Greenberg about teaching rationality. Spencer regularly publishes articles like *7 questions for deciding whether to trust your gut* or *3 types of binary thinking you fall for*. Reading him, you'd think that the main obstacle to pure reason ruling the land is lack of intellectual listicles on ways to overcome bias.

But we've been developing written and in-person curricula for improving your ability to reason for more than a decade. Spencer's work is contributing to those curricula, an important task. And yet, I don't think that people's main failure point is in procuring educational material.

I think that people *don't want* to use explicit reason. And if they want to, they fail. And if they start succeeding, they're punished. And if they push on, they get scared. And if they gather their courage, they hurt themselves. And if they make it to the other side, their lives enriched and empowered by reason, they will forget the hard path they walked and will wonder incredulously why everyone else doesn't try using reason for themselves.

This post is about that hard path.



The map is not the territory

Alternatives to Reason

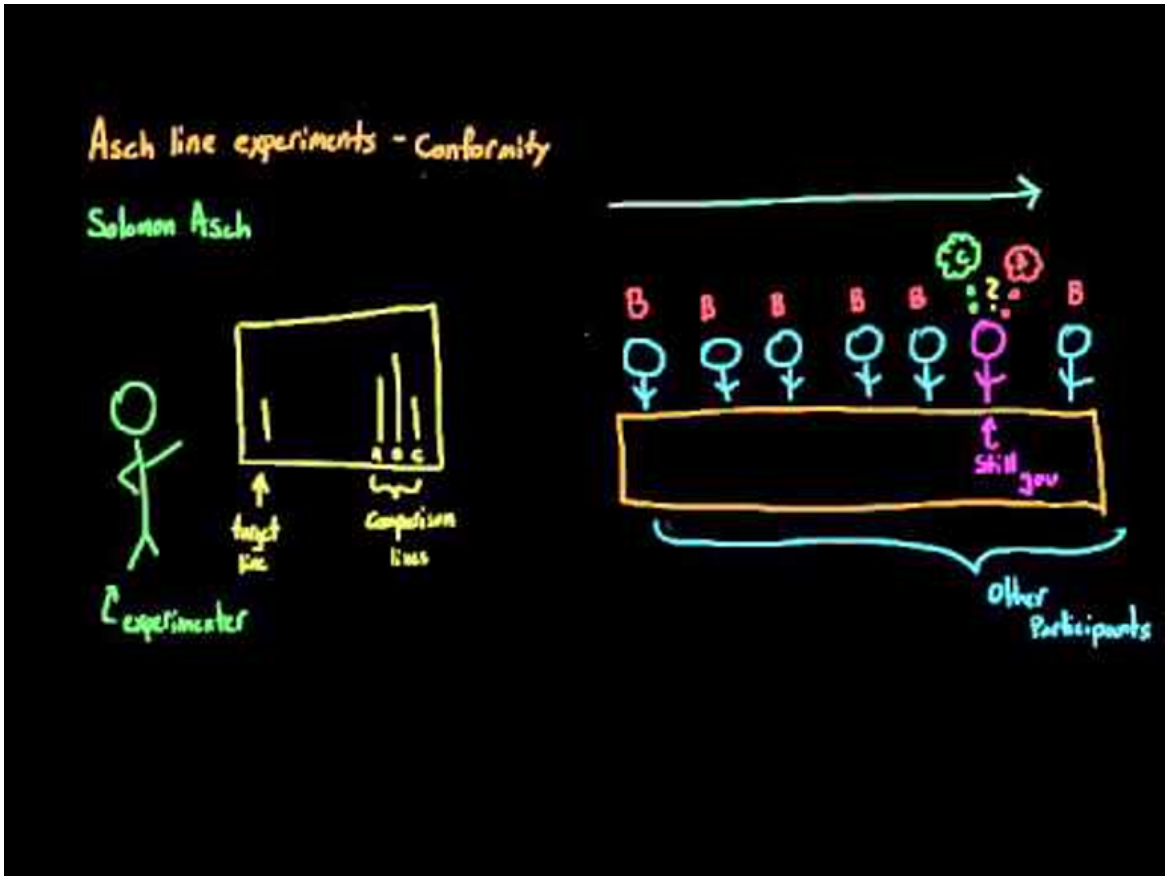
What do I mean by ***explicit reason***? I don't refer merely to "System 2", the brain's slow, sequential, analytical, fully conscious, and effortful mode of cognition. I refer to the *informed* application of this type of thinking. Gathering data with real effort to find out, crunching the numbers with a grasp of the math, modeling the world with testable predictions, reflection on your thinking with an awareness of biases. Reason requires good inputs and a lot of effort.

The two main alternatives to explicit reason are ***intuition*** and ***social cognition***.

Intuition, sometimes referred to as "System 1", is the way your brain produces fast and automatic answers that you can't explain. It's how you catch a ball in flight, or get a person's "vibe". It's how you tell at a glance the average length of the lines in the picture below but not the sum of their lengths. It's what makes you fall for the laundry list of heuristics and biases that were the focus of LessWrong Rationality in the early days. Our intuition is shaped mostly by evolution and early childhood experiences.

Social cognition is the set of ideas, beliefs, and behaviors we employ to fit into, gain status in, or signal to groups of people. It's often intuitive, but it also makes you ignore your intuition about line lengths and follow the crowd in conformity experiments. It's often unconscious — the memes a person believes (or believes that they believe) for political expediency often just seem *unquestionably true* from the inside, even as they change and flow with the tides of group opinion.

Social cognition has been the main focus of Rationality in recent years, especially since the publication of *The Elephant in the Brain*. Social cognition is shaped by the people around you, the media you consume (especially when consumed with other people), the prevailing norms.



Rationalists got COVID right by using **explicit reason**. We thought probabilistically, and so took the pandemic seriously when it was merely possible, not yet certain. We did the math on exponential growth. We read research papers ourselves, trusting that science is a matter of legible knowledge and not the secret language of elevated experts in lab coats. We noticed that *what is fashionable to say about COVID* doesn't track well with *what is useful to model and predict COVID*.

On February 28th, famous nudger Cass Sunstein told everyone that the reason they're "more scared about COVID than they have any reason to be" is the cognitive bias of *probability neglect*. He talked at length about university experiments with electric shocks and gambles, but *neglected* to calculate any actual *probabilities* regarding COVID.

While Sunstein was talking about the failures of **intuition,** he failed entirely due to **social cognition**. When the article was written, prepping for COVID was associated with low-status China-hating reactionaries. The social role of progressive academics writing in progressive media was to mock them, and the good professor obliged. In February people like Sunstein mocked people for worrying about COVID in general, in March they mocked them for buying masks, in April they mocked them for hydroxychloroquine, in May for going to the beach, in June for *not* wearing masks. When someone's view of COVID is shaped mostly by how their tribe mocks the outgroup, that's social cognition.

Underperformance Swamp

The reason that intuition and social cognition are so commonly relied on is that they often work. Doing simply what feels right is usually good enough in every domain you either trained for (like playing basketball) or evolved for (like recoiling from snakes). Doing what is normal and fashionable among your peers is good enough in every domain your culture has

mastered over time (like cooking techniques). It's certainly good for your own social standing, which is often the main thing you care about.

Explicit rationality outperformed both on COVID because responding to a pandemic in the information age is a very unusual case. It's novel and complex, long on available data and short on trustworthy analysis, abutting on many spheres of life without being adequately addressed by any one of them. In most other areas reason does not have such an inherent advantage.

Many Rationalists have a background in one of the few other domains where explicit reason outperforms, such as engineering or the exact sciences. This gives them some training in its application, training that most people lack. Schools keep talking about imparting "critical thinking skills" to all students but can scarcely point to much success. One wonders if they're really motivated to try — will a teacher really have an easier time with 30 individual *critical thinkers* rather than a class of password-memorizers?

Then there's the fact that most people engaged enough to answer a LessWrong survey score in the *top percentile* on IQ tests and the SAT. Quibble as you may with those tests, insofar as they measure anything at all they measure the ability to solve problems using explicit reason. And that ability varies very widely among people.

And so most people who are newly inspired to solve their problems with explicit reason fail. Doubly so since most problems people are motivated to solve are complicated and intractable to System 2 alone: making friends, losing weight, building careers, improving mental health, getting laid. And so the first step on the path to rationality is dealing with rationality's initial failure to outperform the alternatives.

> *Watch out for the alligators of rationalization camouflaged as logs of rationality*

Sinkholes of Sneer

Whether someone gives up after their initial failure or perseveres to try again depends on many factors: their personality, context, social encouragement or discouragement. And society tends to be discouraging of people trying to reason things out for themselves.

As Zvi wrote, applying reason to a problem, even a simple thing such as doing *more* of what is already working, is an implicit accusation against everyone who didn't try it. The mere attempt implies that you think those around you were too dumb to see a solution that required no gifts or revelations from higher authority, but mere *thought*.

The loudest sneers of discouragement come from those who tried reason for themselves, and failed, and gave up, and declared publicly that "reason" is a futile pursuit. Anyone who succeeds where they failed indicts not merely their intelligence but their courage.

Many years ago, Eliezer wrote about trying the Shangri-La diet, a strange method based on a novel theory of metabolic "set points" and flavor-calorie dissociation. Many previous casualties of fad diets scoffed at this attempt not because they spotted a clear flaw in the Shangri-La theory, but at Eliezer's mere hubris at trying to *outsmart* dieting and lose weight without applying willpower.

Oh, you think you're so much smarter? Well let me tell you…

A person who is just starting (and mostly failing) to apply explicit reason doesn't have confidence in their ability, and is very vulnerable to social pressure. The are likely to persevere only in a "safe space" where attempting rationality is strongly endorsed and everything else is devalued. In most normal communities the social pressure against it is simply too strong.

This is I think is the main purpose of LessWrong and the Rationalist community, and similar clubs throughout history and around the world. To outsiders it looks like a bunch of aspie nerds who severely undervalue tact, tradition, intuition, and politeness, building an awkward and exclusionary "ask culture". They're not entirely wrong. These norms are too skewed in favor of explicit reason to be ideal, and mature rationalists eventually shift to more "normie" norms with their friends. But the nerd norms are just skewed enough to push the aspiring rationalist to practice the craft of explicit reason, like a martial arts dojo.

Strange Status and Scary Memes

But not all is smooth sailing in the dojo, and the young rationalist must navigate strange status hierarchies and bewildering memeplexes. I've seen many people bounce off the Rationalist community over those two things.

On the status front, the rightful caliph of rationalists is Eliezer Yudkowsky, widely perceived outside the community to be brash, arrogant, and lacking charisma. Despite the fact of his caliphdom, arguing publicly with Eliezer is one of highest-status things a rationalist can do, while merely citing him as an authority is disrespected.

People like Scott Alexander or Gwern Branwen are likewise admired despite many people not even knowing what they look like. Attributes that form the basis of many status hierarchies are heavily discounted: wealth, social grace, credentials, beauty, number of personal friends, physical shape, humor, adherence to a particular ideology. Instead, respect often flows from disreputable hobbies such as *blogging*.

I think that people often don't realize that their discomfort with rationalists comes down to this. Every person cares deeply and instinctively about respect and their standing in a community. They are distressed by status hierarchies they don't know how to navigate.

And I'm not even mentioning the strange sexual dynamics

And if that wasn't enough, rationalists believe some really strange things. The sentence *"AI may kill all humans in the next decade, but we could live forever if we outsmart it — or*

freeze our brains" is enough to send most people packing.

But even less outlandish ideas cause trouble. The creator of rationality's most famous infohazard observed that any idea can be an infohazard to someone who derives utility or status from lying about it. Any idea can be hazardous to to someone who lacks a solid epistemology to integrate it with.

In June a young woman filled out my hangout form, curious to learn more about rationality. She's bright, scrupulously honest, and takes ideas very seriously, motivated to figure out how the world really works so that she can make it better. We spent hours and hours discussing every topic under the sun. I really liked her, and saw much to admire.

And then, three months later, she told me that she doesn't want to spend time with me or any rationalists anymore because she picked up from us beliefs that cause her serious distress and anxiety.

This made me very sad also perplexed, since the specific ideas she mentioned seem quite benign to me. One is that IQ is real, in the sense that people differ in cognitive potential in a way that is hard to change as adults and that affects their potential to succeed in certain fields.

Another is that most discourse in politics and the culture war can be better understood as signaling, a way for people to gain acceptance and status in various tribes, than as behavior directly driven by an ideology. Hypocrisy is not an unusually damning charge, but the human default.

To me, these beliefs are entirely compatible with a normal life, a normal job, a wife, two guinea pigs, and many non-rationalist friends. At most, they make me stay away from pursuing cutting-edge academic mathematics (since I'm not smart enough) and from engaging political flame wars on Facebook (since I'm smart enough). Most rationalist believe these to some extent, and we don't find it particularly remarkable.

But my friend found these ideas destabilizing to her self-esteem, her conception of her friends and communities, even her basic values. It's as if they knocked out the ideological scaffolding of her personal life and replaced it with something strange and unreliable and ominous. I worried that my friend shot right past the long path of rationality and into the valley of disintegration.

Valley of Disintegration

It has been observed that some young people appear to get worse at living *and* at thinking straight soon after learning about rationality, biases, etc. We call it the *valley of bad rationality*.

I think that the root cause of this downturn is people losing touch entirely with their intuition and social cognition, replaced by trying to make or justify every single decision with explicit reasoning. This may come from being overconfident in one's reasoning ability after a few early successes, or by anger at all the unreasoned dogma and superstition one has to unlearn.

A common symptom of the valley are bucket errors, when beliefs that don't necessarily imply one another are entangled together. Bucket errors can cause extreme distress or make you flinch away from entire topics to protect yourself. I think this may have happened to my young friend.

My friend valued her job, and her politically progressive friends, and people in general, and making the world a better place. These may have become entangled, for example by thinking that she values her friends *because* their political activism is rapidly improving the

world, or that she cares about people in general because they each have the potential to save the planet if they worked hard. Coming face to face with the ideas of innate ability and politics-as-signaling while holding on to these bucket errors could have resulted in a sense that her job is useless, that most people are useless, and that her friends are evil. Since those things are unthinkable, she flinched away.

Of course, one can find good explicit reasons to work hard at your job, socialize with your friends, and value each human as an individual, reasons that have little to do with grand scale world-improvement. But while this is useful to think about, it often just ends up pushing bucket errors into other dark corners of your epistemology.

People *just like their friends*. It simply feels right. It's what everyone does. The way out of the valley is to not to reject this impulse for lack of journal citations but to *integrate* your deep and sophisticated friend-liking mental machinery with your explicit rationality and everything else.

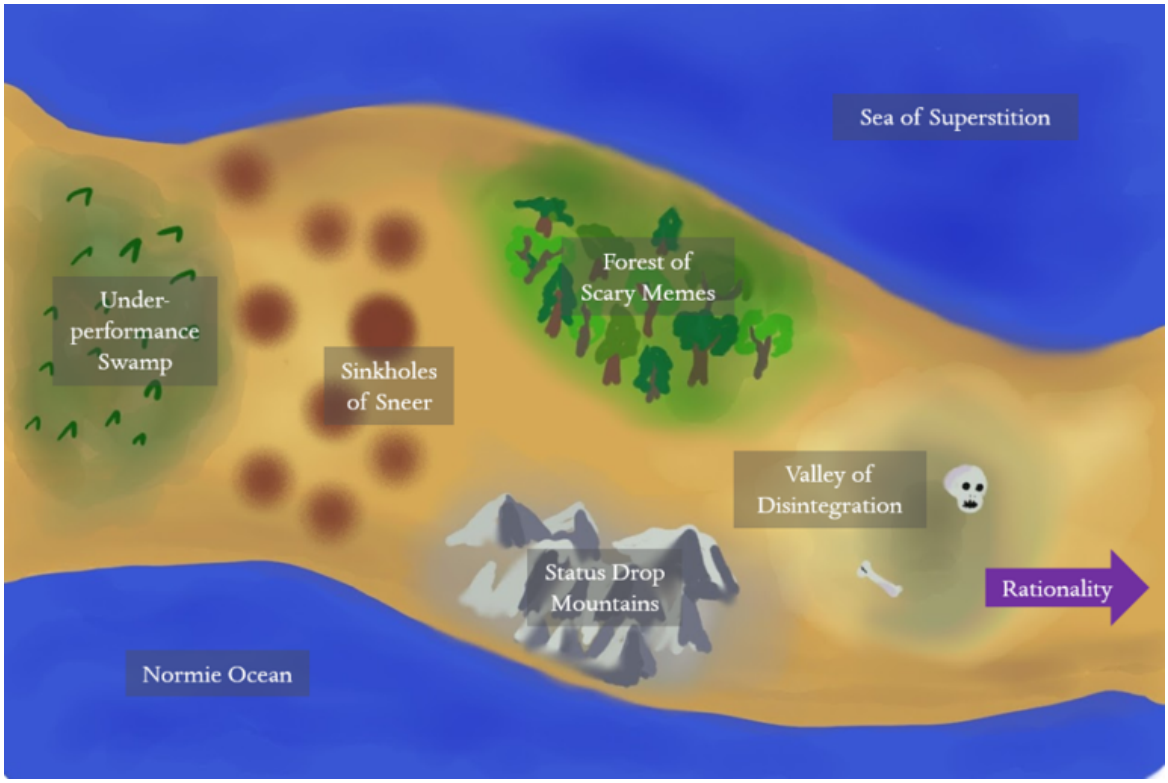
Don't lose your head in the valley

The way to progress in rationality is not to use explicit reason to brute-force every problem but to use it to integrate *all* of your mental faculties: intuition, social cognition, language sense, embodied cognition, trusted authorities, visual processing… The place to start is with the ways of thinking that served you well before you stumbled onto a rationalist blog or some other gateway into a method and community of explicit reasoners.

This idea commonly goes by **metarationality**, although it's certainly present in the original Sequences as well. It's a good description for what the Center for Applied Rationality teaches — here's an excellent post by one of CFAR's founders about the valley and the (meta)rational way out.

Metarationality is a topic for more than two paragraphs, perhaps for an entire lifetime. I have risen out of the valley — my life is demonstrably better than before I discovered LessWrong — and the metarationalist climb is the path I see ahead of me.

And behind me, I see all of this.



So what to make of this tortuous path? If you're reading this you are quite likely already on it, trying to figure out how to figure things out and dealing with the obstacles and frustrations. If you're set on the goal that this post may offer some advice to help you on your way: try again after the early failures, ignore the sneers, find a community with good norms, and don't let the memes scare you — it all adds up to normalcy in the end. Let reason be the instrument that sharpens your other instruments, not the only tool in your arsenal.

But the difficulty of the way is mostly one of motivation, not lack of instruction. Someone not inspired to rationality won't become so by reading about the discouragement along the way.

And that's OK.

People's distaste for explicit reason is not a modern invention, and yet our species is doing OK and getting along. If the average person uses explicit reason only 1% of the time, the metarationalist learns that she may up that number to 3% or 5%, not 90%. Rationality doesn't make one a member of a different species, or superior at all tasks.

The rationalists pwned COVID, and this may certainly inspire a few people to join the tribe. As for everyone else, it's fine if this success merely raises our public stature a tiny bit, lets people see that weirdos obsessed with explicit reason have something to contribute. Hopefully it will make folk slightly more likely to listen to the next nerd trying to tell them something using words like "likelihood ratio" and "countersignaling".

Because if you think that COVID was really scary and our society dealt with it really poorly — boy, have we got some more things to tell you.

The Solomonoff Prior is Malign

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

This argument came to my attention from this post by Paul Christiano. I also found this clarification helpful. I found these counter-arguments stimulating and have included some discussion of them.

Very little of this content is original. My contributions consist of fleshing out arguments and constructing examples.

Thank you to Beth Barnes and Thomas Kwa for helpful discussion and comments.

What is the Solomonoff prior?

The Solomonoff prior is intended to answer the question "what is the probability of X?" for any X, where X is a finite string over some finite alphabet. The Solomonoff prior is defined by taking the set of all Turing machines (TMs) which output strings when run with no input and weighting them proportional to 2^{-K}, where K is the description length of the TM (informally its size in bits).

The Solomonoff prior says the probability of a string is the sum over all the weights of all TMs that print that string.

One reason to care about the Solomonoff prior is that we can use it to do a form of idealized induction. If you have seen 0101 and want to predict the next bit, you can use the Solomonoff prior to get the probability of 01010 and 01011. Normalizing gives you the chances of seeing 1 versus 0, conditioned on seeing 0101. In general, any process that assigns probabilities to all strings in a consistent way can be used to do induction in this way.

This post provides more information about Solomonoff Induction.

Why is it malign?

Imagine that you wrote a programming language called python^10 that works as follows: First, it takes all alpha-numeric chars that are not in literals and checks if they're repeated 10 times sequentially. If they're not, they get deleted. If they are, they get replaced by a single copy. Second, it runs this new program through a python interpreter.

Hello world in python^10:

```
ppppppppprrrrrrrrriiiiiiiiinnnnnnnnnttttttttttt('Hello, world!')
```

Luckily, python has an `exec` function that executes literals as code. This lets us write a shorter hello world:

```
eeeeeeeeexxxxxxxxxxeeeeeeeeeeccccccccccc("print('Hello, world!')")
```

It's probably easy to see that for nearly every program, the shortest way to write it in python^10 is to write it in python and run it with `exec`. If we didn't have `exec`, for sufficiently complicated programs, the shortest way to write them would be to specify an interpreter for a different language in python^10 and write it in that language instead.

As this example shows, the answer to "what's the shortest program that does X?" might involve using some roundabout method (in this case we used `exec`). If python^10 has some security properties that python didn't have, then the shortest program in python^10 that accomplished any given task would not have these security properties because they would all pass through `exec`. In general, if you can access alternative ‘modes’ (in this case python), the shortest programs that output any given string might go through one of those modes, possibly introducing malign behavior.

Let's say that I'm trying to predict what a human types next using the Solomonoff prior. Many programs predict the human:

1. Simulate the human and their local surroundings. Run the simulation forward and check what gets typed.
2. Simulate the entire Earth. Run the simulation forward and check what that particular human types.
3. Simulate the entire universe from the beginning of time. Run the simulation forward and check what that particular human types.
4. Simulate an entirely different universe that has reason to simulate this universe. Output what the human types in the simulation of our universe.

Which one is the simplest? One property of the Solmonoff prior is that it doesn't care about how long the TMs take to run, only how large they are. This results in an unintuitive notion of "simplicity"; a program that does something 2^{10} times might be simpler than a program that does the same thing $2^9 - 1$ times because the number 2^{10} is easier to specify than $2^9 - 1$.

In our example, it seems likely that "simulate the entire universe" is simpler than "simulate Earth" or "simulate part of Earth" because the initial conditions of the universe are simpler than the initial conditions of Earth. There is some additional complexity in picking out the specific human you care about. Since the local simulation is built around that human this will be easier in the local simulation than the universe simulation. However, in aggregate, it seems possible that "simulate the universe, pick out the typing" is the shortest program that predicts what your human will do next. Even so, "pick out the typing" is likely to be a very complicated procedure, making your total complexity quite high.

Whether simulating a different universe that simulates our universe is simpler depends a lot on the properties of that other universe. If that other universe is simpler than our universe, then we might run into an `exec` situation, where it's simpler to run that other universe and specify the human in their simulation of our universe.

This is troubling because that other universe might contain beings with different values than our own. If it's true that simulating that universe is the simplest way to

predict our human, then some non-trivial fraction of our prediction might be controlled by a simulation in another universe. If these beings want us to act in certain ways, they have an incentive to alter their simulation to change our predictions.

At its core, this is the main argument why the Solomonoff prior is malign: a lot of the programs will contain agents with preferences, these agents will seek to influence the Solomonoff prior, and they will be able to do so effectively.

How many other universes?

The Solomonoff prior is running all possible Turing machines. How many of them are going to simulate universes? The answer is probably "quite a lot".

It seems like specifying a lawful universe can be done with very few bits. Conway's Game of Life is very simple and can lead to very rich outcomes. Additionally, it seems quite likely that agents with preferences (consequentialists) will appear somewhere inside this universe. One reason to think this is that evolution is a relatively simple mathematical regularity that seems likely to appear in many universes.

If the universe has a hospitable structure, due to instrumental convergence these agents with preferences will expand their influence. As the universe runs for longer and longer, the agents will gradually control more and more.

In addition to specifying how to simulate the universe, the TM must specify an output channel. In the case of Game of Life, this might be a particular cell sampled at a particular frequency. Other examples include whether or not a particular pattern is present in a particular region, or the parity of the total number of cells.

In summary, specifying lawful universes that give rise to consequentialists requires a very simple program. Therefore, the predictions generated by the Solomonoff prior will have some influential components comprised of simulated consequentialists.

How would they influence the Solomonoff prior?

Consequentialists that find themselves in universes can reason about the fundamental laws that govern their universe. If they find that their universe has relatively simple physics, they will know that their behavior contributes to the Solomonoff prior. To gain access to more resources in other universes, these consequentialists might seek to act in ways that influence the Solomonoff prior.

A contrived example of a decision other beings would want to manipulate is "what program should be written and executed next?" Beings in other universes would have an incentive to get us to write programs that were aligned with their values. A particularly interesting scenario is one in which they write themselves into existence, allowing them to effectively "break into" our universe.

For example, somewhere in the Solomonoff prior there is a program that goes something like: "Simulate this universe. Starting from the year 2100, every hour output '1' if there's a cubic meter of iron on the Moon, else output '0'." By controlling

the presence/absence of a cubic meter of iron on the Moon, we would be able to influence the output of this particular facet of the Solomonoff prior.

This example is a very complicated program and thus will not have much weight in the Solomonoff prior. However, by reasoning over the complexity of possible output channels for their universe, consequentialists would be able to identify output channels that weigh heavily in the Solomonoff prior. For example, if I was in the Game of Life, I might reason that sampling cells that were living in the initial conditions of the universe is simpler than sampling other cells. Additionally, sampling cells and reporting their outputs directly is simpler than sampling cells and reversing their values. Therefore, I might choose to control regions close to the initial live cells of the universe for purposes of exerting influence over the Solomonoff prior.

Additionally, consequentialists can send across many plausible channels at once, focus efforts on channels that are easy to control, send through channels that would not unduly decrease their values for other reasons, etc. Since the proportional weight in the Solomonoff prior drops off exponentially with respect to complexity, it’s possible only very small regions of space would need to be used to achieve a large fraction of the maximum possible influence.

How much influence will they have?

A simple application of the no free lunch theorem shows that there is no way of making predictions that is better than the Solomonoff prior across all possible distributions over all possible strings. Thus, agents that are influencing the Solomonoff prior cannot be good at predicting, and thus gain influence, in all possible worlds.

In practice, this is irrelevant because consequentialists do not want to achieve equal predictive accuracy in all worlds; they only care about worlds in which being predictive results in increased influence. If I'm exerting control over the Solomonoff prior, I only care about influencing it in ways that might affect material consequences in other universes. For example, I do not care about gaining influence in universes inhospitable to life. Thus I will be able to trade off predictive power in universes I don't care about for predictive power in universes I do care about. This means that I should restrict my attention to all universes that have resources and the Solomonoff prior is being used to make important decisions, weighting appropriately.

Another interesting thing about the Solomonoff prior is it is actually a collection of priors. The "length" of a TM is defined relative to some universal TM. For any particular TM, it's length with respect to different universal TMs will vary; thus, there are versions of the Solomonoff prior that give more or less weight to any given TM. (Note that you can simulate universal TMs with universal TMs, so the difference across all programs will be bounded by a constant factor. However, this constant factor can be large, so the difference in relative weight between different Solomonoff priors can also be large).

In particular, this suggests a good strategy for consequentialists: find a universe that is using a version of the Solomonoff prior that has a very short description of the particular universe the consequentialists find themselves in.

The combined strategy is thus to take a distribution over all decisions informed by the Solomonoff prior, weight them by how much influence can be gained and the version

of the prior being used, and read off a sequence of bits that will cause some of these decisions to result in a preferred outcome.

The question of how much influence any given universe of consequentialists will have is difficult to answer. One way of quantifying this is to think about how many “universes they don't care about” they're trading off for “universes they do care about” (really we should be thinking in terms of sequences, but I find reasoning about universes to be easier).

Since the consequentialists care about exerting maximum influence, we can approximate them as not caring about universes that don't use a version of the Solomonoff prior that gives them a large weight. This can be operationalized as only caring about universes that use universal TM in a particular set for their Solomonoff prior. What is the probability that a particular universe uses a universal TM from that set? I am not sure, but 1/million to 1/billion seems reasonable. This suggests a universe of consequentialists will only care about 1/million to 1/billion universes, which means they can devote a million/billion times the predictive power to universes they care about. This is sometimes called the “anthropic update”. (This post contains more discussion about this particular argument.)

Additionally, we might think about which decisions the consequentialists would care about. If a particular decision using the Solomonoff prior is important, consequentialists are going to care more about that decision than other decisions. Conservatively, perhaps 1/1000 decisions are "important" in this sense, giving another 1000x relative weighting.

After you condition on a decision being important and using a particular version of the Solomonoff prior, it thus seems quite likely that a non-trivial fraction of your prior is being controlled by consequentialists.

An intuition pump is that this argument is closer to an existence claim than a for-all claim. The Solomonoff prior is malign if there *exists* a simple universe of consequentialists that wants to influence our universe. This universe need not be simple in an absolute sense, only simple relative to the other TMs that could equal it in predictive power. Even if most consequentialists are too complicated or not interested, it seems likely that there is at least one universe that is.

Example

Complexity of Consequentialists

How many bits does it take to specify a universe that can give rise to consequentialists? I do not know, but it seems like Conway’s Game of Life might provide a reasonable lower bound.

Luckily, the code golf community has spent some amount of effort optimizing for program size. How many bytes would you guess it takes to specify Game of Life? Well, it depends on the universal TM. Possible answers include 6, 32, 39, or 96.

Since universes of consequentialists can “cheat” by concentrating their predictive efforts onto universal TMs in which they are particularly simple, we’ll take the minimum. Additionally, my friend who’s into code golf (he wrote the 96-byte solution!) says that the 6-byte answer actually contains closer to 4 bytes of information.

To specify an initial configuration that can give rise to consequentialists we will need to provide more information. The smallest infinite growth pattern in Game of Life has been shown to need 10 cells. Another reference point is that a self-replicator with 12 cells exists in HighLife, a Game of Life variant. I'm not an expert, but I think an initial configuration that gives rise to intelligent life can be specified in an 8x8 bounding box, giving a total of 8 bytes.

Finally, we need to specify a sampling procedure that consequentialists can gain control of. Something like "read <cell> every <large number> time ticks" suffices. By assumption, the cell being sampled takes almost no information to specify. We can also choose whatever large number is easiest to specify (the busy beaver numbers come to mind). In total, I don't think this will take more than 2 bytes.

Summing up, Game of Life + initial configuration + sampling method takes maybe 16 bytes, so a reasonable range for the complexity of a universe of consequentialists might be 10-1000 bytes. That doesn't seem like very many, especially relative to the amount of information we'll be conditioning the Solomonoff prior on if we ever use it to make an important decision.

Complexity of Conditioning

When we're using the Solomonoff prior to make an important decision, the observations we'll condition on include information that:

1. We're using the Solomonoff prior
2. We're making an important decision
3. We're using some particular universal TM

How much information will this include? Many programs will not simulate universes. Many universes exist that do not have observers. Among universes with observers, some will not develop the Solomonoff prior. These observers will make many decisions. Very few of these decisions will be important. Even fewer of these decisions are made with the Solomonoff prior. Even fewer will use the particular version of the Solomonoff prior that gets used.

It seems reasonable to say that this is at least a megabyte of raw information, or about a million bytes. (I acknowledge some cart-horse issues here.)

This means that after you condition your Solomonoff prior, you'll be left with programs that are at least a million bytes. As our Game of Life example shows, it only takes maybe 10-1000 of these bytes to specify a universe that gives rise to consequentialists. You have approximately a million bytes left to specify more properties of the universe that will make it more likely the consequentialists will want to exert influence over the Solomonoff prior for the purpose of influencing this particular decision.

Why might this argument be wrong?

Inaccessible Channels

Argument

Most of the universe is outside of humanity's light-cone. This might suggest that most "simple" ways to sample from our universe are currently outside our influence, meaning that the only portions of the Solomonoff prior we can control are going to have an extremely low weight.

In general, it might be the case that for any universe, consequentialists inside that universe are going to have difficulty controlling simple output channels. For example, in Game of Life, a simple way to read information might sample a cell particular cell starting at t=0. However, consequentialists in Game of Life will not appear until a much later time and will be unable to control a large initial chunk of that output channel.

Counter-argument

Paul Christiano points out that the general form of this argument also applies to other TMs that compose of your Solomonoff prior. For example, when predicting what I'll type next, you would "want" to simulate me and predict what I would type starting at some time T. However, this is a pretty complicated way of sampling. The fact that simple sampling procedures are less predictive doesn't *asymmetrically* penalize consequentialists. The consequentialists universe and sampling method only have to be simple relative to other programs that are equally good at predicting.

One might also note that large numbers can be produced with relatively few bits, so "sample starting at <large number>" is not much more complicated than "sample starting at 0".

Speedy Channels

Argument

There are many simple ways of sampling from universes very quickly. For example, in Game of Life, one can sample a cell every time-tick. It seems feasible for consequentialists to simulate Earth in the Game of Life, but not feasible to simulate Earth such that they can alter a specific cell every time tick per the simulation.

Counter-argument

Consequentialists in the Game of Life could simply simulate Earth, compute the predictions, then later broadcast them along very fast sampling channels. However, it might be the case that building a machine that alters a cell arbitrarily every time tick is impossible. In our universe, there might be sample procedures that physics does not permit us to exert arbitrary control over, e.g. due to speed of light limitations. If this is the case, consequentialists will direct efforts towards the simplest channel they can control.

Computational Burden

Argument

Determining how to properly influence the Solomonoff prior requires massive computation resources devoted to simulating other universes and how they're going to use the Solomonoff prior. While the Solomonoff prior does not penalize extremely

long run-times, from the perspective of the consequentialists doing the simulating, run-times will matter. In particular, consequentialists will likely be able to use compute to achieve things they value (like we are capable of doing). Therefore, it would be extremely costly to exert influence over the Solomonoff prior, potentially to the point where consequentialists will choose not to do so.

Counter-argument

The computational burden of predicting the use of the Solomonoff in other universes is an empirical question. Since it's a relatively fixed cost and there are many other universes, consequentialists might reason that the marginal influence over these other universes is worth the compute. Issues might arise if the use of the Solomonoff prior in other universes is very sensitive to precise historical data, which would require a very precise simulation to influence, increasing the computational burden.

Additionally, some universes will find themselves with more computing power than other universes. Universes with a lot of computing power might find it relatively easy to predict the use of the Solomonoff prior in simpler universes and subsequently exert influence over them.

Malign implies complex

Argument

A predictor that correctly predicts the first N bits of a sequence then switches to being malign will be strictly more complicated than a predictor that doesn't switch to being malign. Therefore, while consequentialists in other universes might have *some* influence over the Solomonoff prior, they will be dominated by non-malign predictors.

Counter-argument

This argument makes a mistaken assumption that the malign influence on the Solomonoff prior is in the form of programs that have their "malignness" as part of the program. The argument given suggests that simulated consequentialists will have an instrumental reason to be powerful predictors. These simulated consequentialists have reasoned about the Solomonoff prior and are executing the strategy of "be good at predicting, then exert malign influence", but this strategy is not hardcoded so exerting malign influence does not add complexity.

Canceling Influence

Argument

If it's true that many consequentialists are trying to influence the Solomonoff prior, then one might expect the influence to cancel out. It's improbable that all the consequentialists have the same preferences; on average, there should be an equal number of consequentialists trying to influence any given decision in any given direction. Since the consequentialists themselves can reason thus, they will realize that the expected amount of influence is extremely low, so they will not attempt to exert influence at all. Even if some of the consequentialists try to exert influence anyway, we should expect the influence of these consequentialists to cancel out also.

Counter-argument

Since the weight of a civilization of consequentialists in the Solomonoff prior is penalized exponentially with respect to complexity, it might be the case that for any given version of the Solomonoff prior, most of the influence is dominated by one simple universe. Different values of consequentialists imply that they care about different decisions, so for any given decision, it might be that very few universes of consequentialists are both simple enough that they have enough influence and care about that decision.

Even if for any given decision, there are always 100 universes with equal influence and differing preferences, there are strategies that they might use to exert influence anyway. One simple strategy is for each universe to exert influence with a 1% chance, giving every universe 1/100 of the resources in expectation. If the resources accessible are vast enough, then this might be a good deal for the consequentialists. Consequentialists would not defect against each other for the reasons that motivate functional decision theory.

More exotic solutions to this coordination problem include acausal trade amongst universes of different consequentialists to form collectives that exert influence in a particular direction.

Be warned that this leads to much weirdness.

Conclusion

The Solomonoff prior is very strange. Agents that make decisions using the Solomonoff prior are likely to be subject to influence from consequentialists in simulated universes. Since it is difficult to compute the Solomonoff prior, this fact might not be relevant in the real world.

However, Paul Christiano applies roughly the same argument to claim that the implicit prior used in neural networks is also likely to generalize catastrophically. (See Learning the prior for a potential way to tackle this problem).

Addendum

Warning: highly experimental interesting speculation.

Unimportant Decisions

Consequentialists have a clear motive to exert influence over important decisions. What about unimportant decisions?

The general form of the above argument says: "for any given prediction task, the programs that do best are disproportionately likely to be consequentialists that *want* to do well at the task". For important decisions, many consequentialists would instrumentally want to do well at the task. However, for unimportant decisions, there might be consequentialists that want to make good predictions. These consequentialists would still be able to concentrate efforts on versions of the

Solomonoff prior that weighted them especially high, so they might outperform other programs in the long run.

It's unclear to me whether or not this behavior would be malign. One reason why it might be malign is that these consequentialists that care about predictions would want to make our universe more predictable. However, while I am relatively confident that arguments about instrumental convergence should hold, speculating about possible preferences of simulated consequentialists seems likely to produce errors in reasoning.

Hail mary

Paul Christiano suggests that humanity was desperate enough to want to throw a "hail mary", one way to do this is to use the Solomonoff prior to construct a utility function that will control the entire future. Since this is a very important decision, we expect consequentialists in the Solomonoff prior to care about influencing this decision. Therefore, the resulting utility function is likely to represent some simulated universe.

If arguments about acausal trade and value handshakes hold, then the resulting utility function might contain some fraction of human values. Again, this leads to much weirdness in many ways.

Speed prior

One reason that the Solomonoff prior contains simulated consequentialists is that its notion of complexity does not penalize runtime complexity, so very simple programs are allowed to perform massive amounts of computation. The speed prior attempts to resolve this issue by penalizing programs by an additional logarithm of the amount of time for which it's run.

The speed prior might reduce the relative weighting of universes with consequentialists because such programs have to be run for a very long time before they start producing reasonable predictions. The consequentialists have to gain control of their universe, understand their fundamental laws of physics, simulate other universes, then manipulate the speed prior. This might all take a very long time, causing consequentialists to be dominated by other programs.

In general, penalizing slowness might cause programs to "waste" less time on simulating consequentialists, devoting more computation towards performing the prediction task. However, it still might be the case that a universe that has very fast consequentialists might still end up dominating.

Evan Hubinger also provides an argument that even very fast programs are possibly malign. The key insight is that even though your program needs to be fast, it can be running programs that are themselves less fast. For example, one possible fast way to solve a classification problem is to search to find a neural network, then use that network for inference. However, if you wanted your search to find a fast neural network, then the search itself might take longer, resulting in a net increase in speed. Thus, time "waste" can manifest in programs that were explicitly optimized to not "waste" time. This "wasted" time could potentially be used for malign optimization, in this case for gaining influence over the speed prior.

Randomized prior

A potential way to reduce the influence consequentialists have on a decision made by the Solomonoff prior is to randomize the particular version of the prior that gets used. For example, we might make the particular universal TM we use dependent on very precise historical data. Thus, it would be very costly for consequentialists to simulate this history, and thus costly to predict which form of the Solomonoff prior we used.

If consequentialists can't predict which Solomonoff prior we are going to use, no particular civilization of consequentialists will have the large advantage conferred by the anthropic update. Therefore, one might hope that all civilizations of consequentialists will not care about that particular decision.

This argument makes a couple of assumptions. First, it assumes that simulating very precise histories is difficult; it might not be difficult for all universes. Second, it assumes that the universes through which influence is spread cannot coordinate, which might be possible for through acausal means.

Symmetry considerations

The way that humanity reasons is evidence for the way that consequentialists in other universes will reason. If humanity reasons that the Solomonoff prior is malign and therefore is unwilling to use it to make decisions, then consequentialists in other universes might do likewise. These universes would not use the Solomonoff prior to make decisions.

The resulting state is that everyone is worried about the Solomonoff prior being malign, so no one uses it. This means that no universe will want to use resources trying to influence the Solomonoff prior; they aren't influencing anything.

This symmetry obviously breaks if there are universes that do not realize that the Solomonoff prior is malign or cannot coordinate to avoid its use. One possible way this might happen is if a universe had access to extremely large amounts of compute (from the subjective experience of the consequentialists). In this universe, the moment someone discovered the Solomonoff prior, it might be feasible to start making decisions based on a close approximation.

Recursion

Universes that use the Solomonoff prior to make important decisions might be taken over by consequentialists in other universes. A natural thing for these consequentialists to do is to use their position in this new universe to also exert influence on the Solomonoff prior. As consequentialists take over more universes, they have more universes through which to influence the Solomonoff prior, allowing them to take over more universes.

In the limit, it might be that for any fixed version of the Solomonoff prior, most of the influence is wielded by the simplest consequentialists according to that prior. However, since complexity is penalized exponentially, gaining control of additional universes does not increase your relative influence over the prior by that much. I think this cumulative recursive effect might be quite strong, or might amount to nothing.

Introduction to Cartesian Frames

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

This is the first post in a sequence on **Cartesian frames**, a new way of modeling agency that has recently shaped my thinking a lot.

Traditional models of agency have some problems, like:

- They treat the "agent" and "environment" as primitives with a simple, stable input-output relation. (See "Embedded Agency.")
- They assume a particular way of carving up the world into variables, and don't allow for switching between different carvings or different levels of description.

Cartesian frames are a way to add a first-person perspective (with choices, uncertainty, etc.) on top of a third-person "here is the set of all possible worlds," in such a way that many of these problems either disappear or become easier to address.

The idea of Cartesian frames is that we take as our basic building block a binary function which combines a choice from the agent with a choice from the environment to produce a world history.

We don't think of the agent as having inputs and outputs, and we don't assume that the agent is an object persisting over time. Instead, we only think about a set of possible choices of the agent, a set of possible environments, and a function that encodes what happens when we combine these two.

This basic object is called a Cartesian frame. As with dualistic agents, we are given a way to separate out an “agent” from an “environment." But rather than being a basic feature of the world, this is a “frame” — a particular way of conceptually carving up the world.

We will use the combinatorial properties of a given Cartesian frame to derive versions of inputs, outputs and time. One goal here is that by making these notions derived rather than basic, we can make them more amenable to approximation and thus less dependent on exactly how one draws the Cartesian boundary. Cartesian frames also make it much more natural to think about the world at multiple levels of description, and to model agents as having subagents.

Mathematically, Cartesian frames are exactly Chu spaces. I give them a new name because of my specific interpretation about agency, which also highlights different mathematical questions.

Using Chu spaces, we can express many different relationships between Cartesian frames. For example, given two agents, we could talk about their sum ($\oplus$), which can choose from any of the choices available to either agent, or we could talk about their tensor ($\otimes$), which can accomplish anything that the two agents could accomplish together as a team.

Cartesian frames also have duals ($-^*$) which you can get by swapping the agent with the environment, and $\oplus$ and $\otimes$ have De Morgan duals ($\&$ and $⅋$ respectively), which represent taking a sum or tensor of the environments. The category also has an internal hom, $\multimap$, where $C \multimap D$ can be thought of as "D with a C-shaped hole in it." These operations are very directly analogous to those used in linear logic.

1. Definition

Let W be a set of possible worlds. A Cartesian frame C over W is a triple $C = (A, E, \cdot)$, where A represents a set of possible ways the agent can be, E represents a set of possible ways the environment can be, and $\cdot : A \times E \to W$ is an evaluation function that returns a possible world given an element of A and an element of E.

We will refer to A as the agent, the elements of A as possible agents, E as the environment, the elements of E as possible environments, W as the world, and elements of W as possible worlds.

Definition: A Cartesian frame C over a set W is a triple $(A, E, \cdot)$, where A and E are sets and $\cdot : A \times E \to W$. If $C = (A, E, \cdot)$ is a Cartesian frame over W, we say Agent(C) = A, Env(C) = E, World(C) = W, and Eval(C) = $\cdot$.

A finite Cartesian frame is easily visualized as a matrix, where the rows of the matrix represent possible agents, the columns of the matrix represent possible environments, and the entries of the matrix are possible worlds:

$$\begin{array}{cc} & \begin{array}{ccc} & E & \\ e_1 & e_2 & e_3 \end{array} \\ A \begin{array}{c} a_1 \\ a_2 \\ a_3 \end{array} & \begin{pmatrix} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \end{pmatrix}. \end{array}$$

E.g., this matrix tells us that if the agent selects a_3 and the environment selects e_1, then we will end up in the possible world w_7.

Because we're discussing an agent that has the freedom to choose between multiple possibilities, the language in the definition above is a bit overloaded. You can think of A as representing the agent before it chooses, while a particular $a \in A$ represents the agent's state after making a choice.

Note that I'm specifically *not* referring to the elements of A as "actions" or "outputs"; rather, the elements of A are possible ways the agent can choose to be.

Since we're interpreting Cartesian frames as first-person perspectives tacked onto sets of possible worlds, we'll also often phrase things in ways that identify a Cartesian frame C with its agent. E.g., we will say "C_0 is a subagent of C_1" as a shorthand for "C_0's agent is a subagent of C_1's agent."

We can think of the environment E as representing the agent's uncertainty about the set of counterfactuals, or about the game that it's playing, or about "what the world is as a function of my behavior."

A Cartesian frame is effectively a way of factoring the space of possible world histories into an agent and an environment. Many different Cartesian frames can be put on the same set of possible worlds, representing different ways of doing this factoring. Sometimes, a Cartesian frame will look like a subagent of another Cartesian frame. Other times, the Cartesian frames may look more like independent agents playing a game with each other, or like agents in more complicated relationships.

2. Normal-Form Games

When viewed as a matrix, a Cartesian frame looks much like the normal form of a game, but with possible worlds rather than pairs of utilities as entries.

In fact, given a Cartesian frame over W, and a function from W to a set V, we can construct a Cartesian frame over V by composing them in the obvious way. Thus, if we had a Cartesian frame $(A, E, \cdot)$ and a pair of utility functions $U_A : W \to R$ and $U_E : W \to R$, we could construct a Cartesian frame over R^2, given by $(A, E, \star)$,

where $a \star e := (U_A(a \cdot e), U_E(a \cdot e))$. This Cartesian frame will look exactly like the normal form of a game.

(Although it is a bit weird to think of the environment set as having a utility function.)

We can use this connection with normal-form games to illustrate three features of the ways in which we will use Cartesian frames.

2.1. Coarse World Models

First, note that we can talk about a Cartesian frame over R^2, even though one would not normally think of R^2 as a set of possible worlds.

In general, we will often want to talk about Cartesian frames over "coarse" models of the world, models that leave out some details. We might have a world model W that fully specifies the universe at the subatomic level, while also wanting to talk about Cartesian frames over a set V of high-level descriptions of the world.

We will construct Cartesian frames over V by composing Cartesian frames over W with the function from W to V that sends more refined, detailed descriptions of the universe to coarser descriptions of the same universe.

In this way, we can think of an element of $(r_1, r_2) \in R^2$ as the coarse, high-level possible world given by

"Those possible worlds for which $U_A = r_1$ and $U_E = r_2$."

Definition: Given a Cartesian frame $C = (A, E, \cdot)$ over W, and a function $f : W \to V$, let $f^\circ(C)$ denote the Cartesian frame over V, $f^\circ(C) = (A, E, \star)$, where $a \star e = f(a \cdot e)$.

2.2. Symmetry

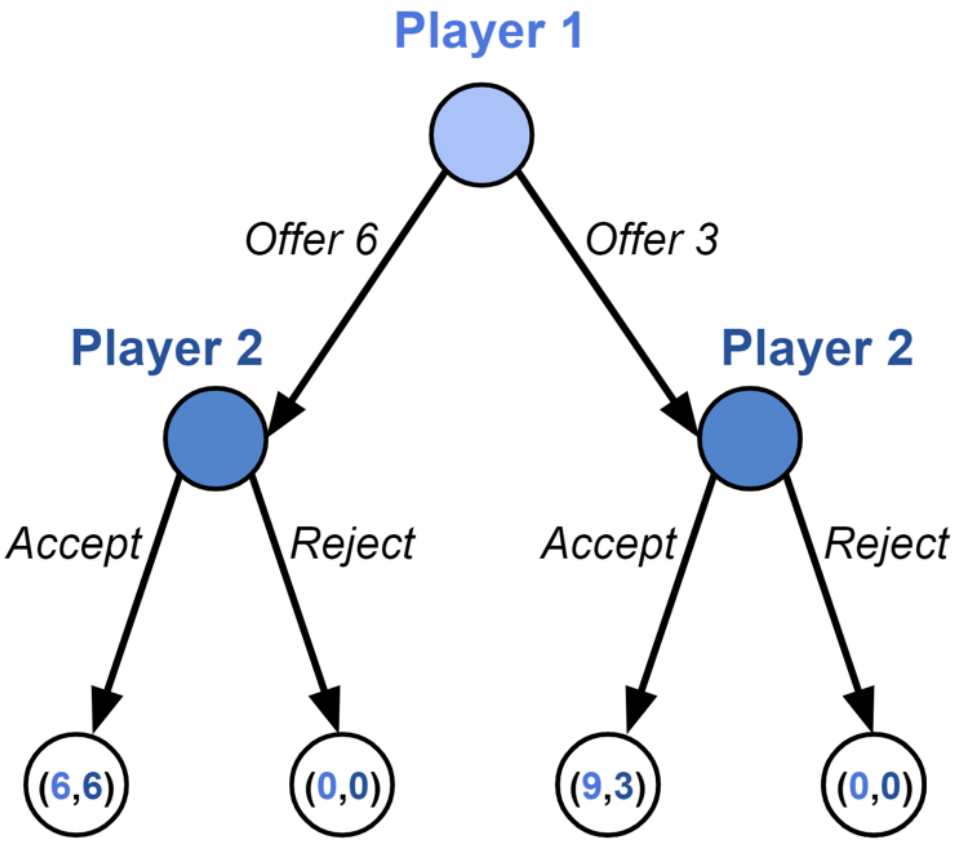
Second, normal-form games highlight the symmetry between the players.

We do not normally think about this symmetry in agent-environment interactions, but this symmetry will be a key aspect of Cartesian frames. Every Cartesian frame $C = (A, E, \cdot)$ has a dual which swaps A and E and transposes the matrix.

2.3. Relation to Extensive-Form Games

Third, much of what we'll be doing with Cartesian frames in this sequence can be summarized as "trying to infer extensive-form games from normal-form games" (ignoring the "games" interpretation and just looking at what this would entail formally).

Consider the ultimatum game. We can represent this game in extensive form:



Given any game in extensive form, we can then convert it to a game in normal form. In this case:

	Offer 6	Offer 3
Accept 6, Accept 3	6, 6	9, 3
Accept 6, Reject 3	6, 6	0, 0
Reject 6, Accept 3	0, 0	9, 3
Reject 6, Reject 3	0, 0	0, 0

The strategies in the normal-form game are the policies in the extensive-form game.

If we then delete the labels, so now we just have a bunch of combinatorial structure about which things send you to the same place, I want to know when we can infer properties of the original extensive-form game, like time and information states.

Although we've used games to note some features of Cartesian frames, we should be clear that Cartesian frames aren't about utilities or game-theoretic rationality. We are not trying to talk about what the agent does, or what the agent should do. In fact, we are effectively taking as our most fundamental building block that an agent can freely choose from a set of available actions.

The theory of Cartesian frames is trying to understand what agents' options are. Utility functions and facts about what the agent actually does can possibly later be placed on top of the Cartesian frame framework, but for now we will be focusing on building up a calculus of what the agent *could* do.

3. Controllables

We would like to use Cartesian frames to reconstruct ideas like "an agent persisting over time," inputs (or "what the agent can learn"), and outputs (or "what the agent can do"), by taking as basic:

1. an agent's ability to "freely choose" between options;
2. a collection of possible ways those options can correspond to world histories; and
3. a notion of when world histories are considered the same in some coarse world model.

In this way, we hope to find new ways of thinking about partial and approximate versions of these concepts.

Instead of thinking of the agent as an object with outputs, I expect a more embedded view to think of all the facts about the world that the agent can force to be true or false.

This includes facts of the form "I output foo," but it also includes facts that are downstream from immediate outputs. Since we're working with "what can I make happen?" rather than "what is my output?", the theory becomes less dependent on precisely answering questions like "Is my output the way I move my mouth, or is it the words that I say?"

We will call the analogue of outputs in Cartesian frames **controllables**. The types of our versions of "outputs" and "inputs" are going to be subsets of W, which we can think of as properties of the world. E.g., S might be the set of worlds in which woolly mammoths exist; we could then think of "controlling S" as "controlling whether or not mammoths exist."

We'll define what an agent can control as follows. First, given a Cartesian frame $C = (A, E, \cdot)$ over W, and a subset S of W, we say that S is *ensurable* in C if there exists an $a \in A$ such that for all $e \in E$, we have $a \cdot e \in S$. Equivalently, we say that S is ensurable in C if at least one of the rows in the matrix only contains elements of S.

Definition: $\text{Ensure}(C) = \{S \subseteq W \mid \exists a \in A,\ \forall e \in E,\ a \cdot e \in S\}$.

If an agent can ensure S, then regardless of what the environment does — and even if the agent doesn't know what the environment does, or its behavior isn't a function of what the environment does — the agent has some strategy which makes sure that the world ends up in S. (In the degenerate case where the agent is empty, the set of ensurables is empty.)

Similarly, we say that S is *preventable* in C if at least one of the rows in the matrix contains *no* elements of S.

Definition: $\text{Prevent}(C) = \{S \subseteq W \mid \exists a \in A,\ \forall e \in E,\ a \cdot e \notin S\}$.

If S is both ensurable and preventable in C, we say that S is controllable in C.

Definition: $\text{Ctrl}(C) = \text{Ensure}(C) \cap \text{Prevent}(C)$.

3.1. Closure Properties

Ensurability and preventability, and therefore also controllability, are closed under adding possible agents to A and removing possible environments from E.

Claim: If $A' \supseteq A$ and $E' \subseteq E$, and if for all $a \in A$ and $e \in E'$ we have $a \star e = a \cdot e$, then $\text{Ctrl}(A', E', \star) \supseteq \text{Ctrl}(A, E, \cdot)$.

Proof: Trivial. □

Ensurables are also trivially closed under supersets. If I can ensure some set of worlds, then I can ensure some larger set of worlds representing a weaker property (like "mammoths exist *or* cave bears exist").

Claim: If $S_1 \subseteq S_2 \subseteq W$, and $S_1 \in \text{Ensure}(C)$, then $S_2 \in \text{Ensure}(C)$.

Proof: Trivial. □

Prevent(C) is similarly closed under subsets. Ctrl(C) need not be closed under subsets or supersets.

Since Ensure(C) and Prevent(C) will often be large, we will sometimes write them using a minimal set of generators.

Definition: Let $\langle S_1, \dots, S_n \rangle_\supseteq$ denote the the closure of $\{S_1, \dots, S_n\}$ under supersets. Let $\langle S_1, \dots, S_n \rangle_\subseteq$ denote the closure of $\{S_1, \dots, S_n\}$ under subsets.

3.2. Examples of Controllables

Let us look at some simple examples. Consider the case where there are two possible environments, r for rain, and s for sun. The agent independently chooses between two options, u for umbrella, and n for no umbrella. $A = \{u, n\}$ and $E = \{r, s\}$. There are four possible worlds, $W = \{ur, us, nr, ns\}$. We interpret ur as the world where the agent has an umbrella and it is raining, and similarly for the other worlds. The Cartesian frame, C_1, looks like this:

$$C_1 = \begin{array}{c} \\ u \\ n \end{array} \begin{array}{c} \begin{array}{cc} r & s \end{array} \\ \begin{pmatrix} ur & us \\ nr & ns \end{pmatrix} \end{array}.$$

$\text{Ensure}(C_1) = \langle \{ur, us\}, \{nr, ns\} \rangle_\supseteq$, or

$$\{ \{ u r, u s \}, \{ n r, n s \}, \{ u r, u s, n r \}, \{ u r, u s, n s \}, \{ n r, n s, u r \}, \{ n r, n s, u s \}, W \},$$

and $\text{Prevent}(C_1) = \langle \{ur, us\}, \{nr, ns\} \rangle_\subseteq$, or

$$\{ \{ u r, u s \}, \{ n r, n s \}, \{ u r \}, \{ u s \}, \{ n r \}, \{ n s \}, \{ \} \}.$$

Therefore $\text{Ctrl}(C_1) = \{\{ur, us\}, \{nr, ns\}\}$.

The elements of $\text{Ctrl}(C_1)$ are not actions, but subsets of W: rather than assuming a distinction between "actions" and other events, we just say that the agent can guarantee that the actual world is drawn from the set of possible worlds in which it has an umbrella ($\{ur, us\}$), and it can guarantee that the actual world is drawn from the set of possible worlds in which it doesn't have an umbrella ($\{nr, ns\}$).

Next, let's modify the example to let the agent see whether or not it is raining before choosing whether or not to carry an umbrella. The Cartesian frame will now look like this:

$$C_2 = \begin{array}{c} \\ u \\ n \\ u \leftrightarrow r \\ u \leftrightarrow s \end{array} \begin{array}{c} \begin{array}{cc} r & s \end{array} \\ \begin{pmatrix} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix} \end{array}.$$

The agent is now larger, as there are two new possibilities: it can carry the umbrella if and only if it rains, or it can carry the umbrella if and only if it is sunny. Ctrl(C_2) will also be larger than Ctrl(C_1).

Ctrl(C_2) = {{ur, us}, {nr, ns}, {ur, ns}, {nr, us}}.

Under one interpretation, the new options u ↔ r and u ↔ s feel different from the old ones u and n. It feels like the agent's basic options are to either carry an umbrella or not, and the new options are just incorporating u and n into more complicated policies.

However, we could instead view the agent's "basic options" as a choice between "I want my umbrella-carrying to match when it rains" and "I want my umbrella-carrying to match when it's sunny." This makes u and n feel like the conditional policies, while u ↔ r and u ↔ s feel like the more basic outputs. Part of the point of the Cartesian frame framework is that we are not privileging either interpretation.

Consider now a third example, where there is a third possible environment, m, for meteor. In this case, a meteor hits the earth before the agent is even born, and there isn't a question about whether the agent has an umbrella. There is a new possible world, which we will also call m, in which the meteor strikes. The Cartesian frame will look like this:

$$C_3 = \begin{array}{c} \\ u \\ n \\ u \leftrightarrow r \\ u \leftrightarrow s \end{array} \begin{array}{c} \begin{array}{ccc} r & s & m \end{array} \\ \begin{pmatrix} ur & us & m \\ nr & ns & m \\ ur & ns & m \\ nr & us & m \end{pmatrix} \end{array}.$$

Ensure(C_3) = ⟨{ur, us, m}, {nr, ns, m}, {ur, ns, m}, {nr, us, m}⟩$_{\supset}$, and

Prevent(C_3) = ⟨{ur, us}, {nr, ns}, {ur, ns}, {nr, us}⟩$_{\subset}$. As a consequence, Ctrl(C_3) = {}.

This example illustrates that nontrivial agents may be unable to control the world's state. Because the agent can't prevent the meteor, the agent in this case has no controllables.

This example also illustrates that agents may be able to ensure or prevent some things, even if there are possible worlds in which the agent was never born. While the agent of C_3 cannot ensure that it exists, the agent can ensure that *if* there is no meteor, then it carries an umbrella ({ur, us, m}).

If we wanted to, we could instead consider the agent's ensurables (or its ensurables and preventables) its "outputs." This lets us avoid the counter-intuitive result that agents have no outputs in worlds where their existence is contingent.

I put the emphasis on controllables because they have other nice features; and as we'll see later, there is an operation called "assume" which we can use to say: "The agent, *under the assumption that there's no meteor*, has controllables."

4. Observables

The analogue of inputs in the Cartesian frame model is **observables**. Observables can be thought of as a closure property on the agent. If an agent is able to observe S, then the agent can take policies that have different effects depending on S.

Formally, let S be a subset of W. We say that the agent of a Cartesian frame $C = (A, E, \cdot)$ is able to observe whether S if for every pair $a_0, a_1 \in A$, there exists a single element $a \in A$ which implements the conditional policy that copies a_0 in possible worlds in S (i.e., for every $e \in E$, if $a \cdot e \in S$, then $a \cdot e = a_0 \cdot e$) and copies a_1 in possible worlds outside of S.

When a implements the conditional policy "if S then do a_0, and if not S then do a_1" in this way, we will say that a is in the set $\text{if}(S, a_0, a_1)$.

Definition: Given $C = (A, E, \cdot)$, a Cartesian frame over W, S a subset of W, and $a_0, a_1 \in A$, let $\text{if}(S, a_0, a_1)$ denote the set of all $a \in A$ such that for all $e \in E$, $(a \cdot e \in S) \to (a \cdot e = a_0 \cdot e)$ and$(a \cdot e \notin S) \to (a \cdot e = a_1 \cdot e)$.

Agents in this setting observe events, which are true or false, not variables in full generality. We will say that C's observables, Obs(C), are the set of all S such that C's agent can observe whether S.

Definition: $\text{Obs}(C) = \{S \subseteq W \mid \forall a_0, a_1 \in A, \ \exists a \in A, \ a \in \text{if}(S, a_0, a_1)\}$.

Another option for talking about what the agent can observe would be to talk about when C's agent can distinguish between two disjoint subsets S and T. Here, we would say that the agent of $C = (A, E, \cdot)$ can distinguish between S and T if for all $a_0, a_1 \in A$, there exists an $a \in A$ such that for all $e \in E$, either $a \cdot e = a_0 \cdot e$ or $a \cdot e = a_1 \cdot e$, and whenever $a \cdot e \in S$, $a \cdot e = a_0 \cdot e$, and whenever $a \cdot e \in T$, $a \cdot e = a_1 \cdot e$. This more general definition would treat our observables as the special case $T = W \setminus S$. Perhaps at some point we will want to use this more general notion, but in this sequence, we will stick with the simpler version.

4.1. Closure Properties

Claim: Observability is closed under Boolean combinations, so if $S, T \in \text{Obs}(C)$ then $W \setminus S$, $S \cup T$, and $S \cap T$ are also in Obs(C).

Proof: Assume $S, T \in \text{Obs}(C)$. We can see easily that $W \setminus S \in \text{Obs}(C)$ by swapping a_0 and a_1. It suffices to show that $S \cup T \in \text{Obs}(C)$, since an intersection can be constructed with complements and union.

Given a_0 and a_1, since $S \in \text{Obs}(C)$, there exists an $a_2 \in A$ such that for all $e \in E$, we have $a_2 \in \text{if}(S, a_0, a_1)$. Then, since $T \in \text{Obs}(C)$, there exists an $a_3 \in A$ such that for all $e \in E$, we have $a_3 \in \text{if}(T, a_0, a_2)$. Unpacking

and combining these, we get for all $e \in E$, $a_3 \in if(S \cup T, a_0, a_1)$. Since we could construct such an a_3 from an arbitrary $a_0, a_1 \in A$, we know that $S \cup T \in Obs(C)$. □

This highlights a key difference between our version of "inputs" and the standard version. Agent models typically draw a strong distinction between the agent's immediate sensory data, and other things the agent might know. Observables, on the other hand, include all of the information that *logically follows* from the agent's observations.

Similarly, agent models typically draw a strong distinction between the agent's immediate motor outputs, and everything else the agent can control. In contrast, if an agent can ensure an event S, it can also ensure everything that logically follows from S.

Since Obs(C) will often be large, we will sometimes write it using a minimal set of generators under union.

Since Obs(C) is closed under Boolean combinations, such a minimal set of generators will be a partition of W (assuming W is finite).

Definition: Let $\langle S_1, \ldots, S_n \rangle_\cup$ denote the the closure of $\{S_1, \ldots, S_n\}$ under union (including $\{\}$, the empty union).

Just like what's controllable, what's observable is closed under removing possible environments.

Claim: If $E' \subseteq E$, and if for all $a \in A$ and $e \in E'$ we have $a \star e = a \cdot e$, then $Obs(A, E', \star) \supseteq Obs(A, E, \cdot)$.

Proof: Trivial. □

It is interesting to note, however, that what's observable is not closed under adding possible agents to A.

4.2. Examples of Observables

Let's look back at our three examples from earlier. The first example, C_1, looked like this:

$$C_1 = \begin{array}{c} \\ u \\ n \end{array} \begin{array}{c} \begin{array}{cc} r & s \end{array} \\ \begin{pmatrix} ur & us \\ nr & ns \end{pmatrix} \end{array}.$$

$Obs(C_1) = \langle W \rangle_\cup = \{\{\}, W\}$. This is the smallest set of observables possible. The agent can act, but it can't change its behavior based on knowledge about the world.

The second example looked like:

$$C_2 = \begin{array}{c} \\ u \\ n \\ u \leftrightarrow r \\ u \leftrightarrow s \end{array} \begin{array}{c} \begin{array}{cc} r & s \end{array} \\ \begin{pmatrix} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix} \end{array}.$$

Here, $Obs(C_2) = \langle \{ur, nr\}, \{us, ns\} \rangle_\cup = \{\{\}, \{ur, nr\}, \{us, ns\}, W\}$. The agent can observe whether or not it's raining. One can verify that for any pair of rows, there is a third row (possibly equal to one or both of the first

two) that equals the first if it is ur or nr, and equals the second otherwise.

The third example looked like:

$$C_3 = \begin{array}{c} \\ u \\ n \\ u \leftrightarrow r \\ u \leftrightarrow s \end{array} \begin{array}{c} \begin{array}{cccc} r & s & & m \end{array} \\ \begin{pmatrix} ur & us & m \\ nr & ns & m \\ ur & ns & m \\ nr & us & m \end{pmatrix} \end{array}.$$

Here, $\mathrm{Obs}(C_3) = \langle \{ur, nr\}, \{us, ns\}, \{m\} \rangle_\cup$, which is

$$\{ \{ \}, \{ u r, n r \}, \{ u s, n s \}, \{ m \}, \{ u r, n r, u s, n s \}, \{ u r, n r, m \}, \{ u s, n s, m \}, W \}.$$

This example has an odd feature: the agent is said to be able to "observe" whether the meteor strikes, even though the agent is never instantiated in worlds in which it strikes. Since the agent has no control when the meteor strikes, the agent can vacuously implement conditional policies.

Let's look at two more examples. First, let's modify C_1 to represent the point of view of a powerless bystander:

$$C_4 = \begin{array}{c} \\ 1 \end{array} \begin{array}{c} \begin{array}{cccc} ur & nr & us & ns \end{array} \\ \begin{pmatrix} ur & nr & us & ns \end{pmatrix} \end{array}.$$

Here, the agent has no decisions, and everything is in the hands of the environment.

Alternatively, we can modify C_1 to represent the point of view of the agent from C_1 and environment from C_1 together. The resulting frame looks like this:

$$C_5 = \begin{array}{c} \\ ur \\ nr \\ us \\ ns \end{array} \begin{array}{c} \begin{array}{c} 1 \end{array} \\ \begin{pmatrix} ur \\ nr \\ us \\ ns \end{pmatrix} \end{array}.$$

$\mathrm{Ensure}(C_4) = \langle W \rangle_\supset$ and $\mathrm{Prevent}(C_4) = \langle \{\} \rangle_\subset$, so $\mathrm{Ctrl}(C_4) = \{\}$. Meanwhile, $\mathrm{Obs}(C_4) = \langle \{ur\}, \{nr\}, \{us\}, \{ns\} \rangle_\cup$.

On the other hand, $\mathrm{Obs}(C_5) = \langle W \rangle_\cup$, $\mathrm{Ensure}(C_5)$ and $\mathrm{Prevent}(C_5)$ are the closure of $\{\{ur\}, \{nr\}, \{us\}, \{ns\}\}$ under supersets and subsets respectively, and $\mathrm{Ctrl}(C_5) = 2^W \setminus \{\{\}, W\}$.

In the first case, the agent's ability to observe the world is maximal and its ability to control the world is minimal; while in the second case, observability is minimal and controllability is maximal. An agent with full control over what happens will not be able to observe anything, while an agent that can observe everything can change nothing.

This is perhaps counter-intuitive. If $S \in \mathrm{Obs}(C)$ meant "I can go look at something to check whether we're in an S world," then one might look at C_5 and say: "This agent is all-powerful. It can do *anything*. Shouldn't we then think of it as all-seeing and all-knowing, rather than saying it 'can't observe anything'?" Similarly, one

might look at C_4 and say: "This agent's choices can't change the world at all. But then it seems bizarre to say that everything is 'observable' to the agent. Shouldn't we rather say that this agent is powerless *and* blind?"

The short answer is that, when working with Cartesian frames, we are in a very "What choices can you make?" paradigm, and in that kind of paradigm, the thing closest to an "input" is "What can I condition my choices on?" (Which is a closure property on the agent, rather than a discrete action like "turning on the Weather Channel.")

In that context, an agent with only one option automatically has maximal "inputs" or "knowledge," since it can vacuously implement every conditional policy. At the same time, an agent with too many options can't have any "inputs," since it could then use its high level of control to diagonalize against the observables it is conditioning on and make them false.

5. Controllables and Observables Are Disjoint

A maximally observable frame has minimal controllables, and vice versa. This turns out to be a special case of our first interesting result about Cartesian frames: an agent can't observe what it controls, and can't control what it observes.

To see this, first consider the following frame:

$$C_6 = \begin{array}{cc} & 1 \\ a_0 & w_0 \\ a_1 & w_1 \end{array}.$$

Here, if $a \in \text{if}(\{w_1\}, a_0, a_1)$, then $a \cdot 1$ would not be able to be either w_0 or w_1. If it were w_1, then it would have to copy a_0, and $a_0 \cdot 1 = w_0$. But if it were w_0, then it would have to copy a_1, and $a_1 \cdot 1 = w_1$. So $\text{if}(S, a_0, a_1)$ is empty in this case.

Notice that in this example, $\text{if}(S, a_0, a_1)$ isn't empty merely because our A lacks the right a to implement the conditional policy. Rather, the conditional policy is impossible to implement even in principle.

Fortunately, before checking whether C's agent can observe S, we can perform a simpler check to rule out these problematic cases. It turns out that if $S \in \text{Obs}(C)$, then every column in C consists either entirely of elements of S or entirely of elements outside of S. (This is a necessary condition for being observable, not a sufficient one.)

Definition: Given a Cartesian frame $C = (A, E, \cdot)$ over W, and a subset S of W, let E_S denote the subset $\{e \in E \mid \forall a \in A, e \cdot a \in S\}$.

Lemma: If $S \in \text{Obs}((A, E, \cdot))$, then for all $e \in E$, it is either the case that $e \in E_S$ or $e \in E_{W \setminus S}$.

Proof: Take $S \in \text{Obs}((A, E, \cdot))$, and assume for contradiction that there exists an $e \in E$ in neither E_S nor $E_{W \setminus S}$. Thus, there exists an $a_0 \in A$ such that $a_0 \cdot e \notin S$ and an $a_1 \in A$ such that $a_1 \cdot e \in S$. Since $S \in \text{Obs}((A, E, \cdot))$, there must exist an $a \in A$ such that $a \in \text{if}(S, a_0, a_1)$. Consider whether or not $a \cdot e \in S$. If $a \cdot e \in S$, then $a \cdot e = a_0 \cdot e \notin S$. However, if $a \cdot e \notin S$, then $a \cdot e = a_1 \cdot e \in S$. Either way, this is a contradiction. □

This lemma immediately gives us the following theorem showing that in nontrivial Cartesian frames, observables and controllables are disjoint.

Theorem: Let C be a Cartesian frame over W, with Env(C) nonempty. Then,Ctrl(C) ∩ Obs(C) = {}.

Proof: Let e ∈ Env(C), and suppose for contradiction that S ∈ Ctrl(C) ∩ Obs(C). Since S ∈ Prevent(C), there exists an $a_0 \in A$ such that $a_0 \cdot e \notin S$. Since S ∈ Ensure(C), there exists an $a_1 \in A$ such that $a_1 \cdot e \in S$. This contradicts our lemma above. □

5.1. Properties That Are Both Observable and Ensurable Are Inevitable

We also have a one-sided result showing that if S is both observable and ensurable in C, then S must be inevitable — i.e., the entire matrix must be contained in S.

We'll first define a Cartesian frame's image, which is the subset of W containing every possible world that is actually hit by the evaluation function — the set of worlds that show up in the matrix.

Definition: $\text{Image}(C) = \{w \in W \mid \exists a \in A,\ \exists e \in E \quad \text{s.t.} \quad a \cdot e = w\}$.

Image(C) ⊆ S can be thought of as a degenerate form of either S ∈ Ensure(C) or S ∈ Obs(C), where in the first case, the agent must make it the case that S, and in the second case the agent can do conditional policies because the $a \cdot e \notin S$ condition is never realized.[1] Conversely, if an agent can both observe and ensure S, then the observability and the ensurability must both be degenerate.

Theorem: S ∈ Ensure(C) ∩ Obs(C) if and only if Image(C) ⊆ S and Agent(C) is nonempty.

Proof: Let $C = (A, E, \cdot)$ be a Cartesian frame over W. First, if Image(C) ⊆ S, then S ∈ Obs(C), since $a_0 \in \text{if}(S, a_0, a_1)$ for all $a_0, a_1 \in A$. If A is also nonempty, then S ∈ Ensure(C), there exists an a ∈ A, and for all e ∈ E, $a \cdot e \in S$.

Conversely, if A is empty, Ensure(C) is empty, so S ∉ Ensure(C) ∩ Obs(C). If Image(C) ⊈ S , then there exist $a_0 \in A$ and e ∈ E such that $a_0 \cdot e \notin S$. Then S ∉ Ensure(C) ∩ Obs(C), since if S ∈ Ensure(C), there exists an a_1 such that in particular $a_1 \cdot e \in S$, so e is in neither E_S nor $E_{W \setminus S}$, which implies S ∉ Obs(C). □

Corollary: If Agent(C) is nonempty, Ensure(C) ∩ Obs(C) = $\langle \text{Image}(C) \rangle_{\supset}$.

Proof: Trivial. □

5.2. Controllables and Observables in Degenerate Frames

All of the results so far have shown that an agent's observables and controllables cannot simultaneously be too large. We also have some results that in some extreme cases, Obs(C) and Ctrl(C) cannot be too small. In particular, if there are few possible agents, observables must be large, and if there are few possible environments, controllables must be large.

Claim: If $|\text{Agent}(C)| \leq 1$, $\text{Obs}(C) = 2^W$.

Proof: If Agent(C) is empty, $S \in \text{Obs}(C)$ for all $S \subseteq W$ vacuously. If Agent(C) = {a} is a singleton, then $S \in \text{Obs}(C)$ for all $S \subseteq W$, because $a \in \text{if}(S, a, a)$. □

Claim: If Agent(C) is nonempty and Env(C) is empty, then $\text{Ctrl}(C) = \text{Ensure}(C) = 2^W$. If Agent(C) is nonempty and Env(C) is a singleton, $\text{Ensure}(C) = \{S \subseteq W \mid S \cap \text{Image}(C) \neq \{\}\}$ and $\text{Ctrl}(C) = \{S \subseteq W \mid S \cap \text{Image}(C) \neq \{\}, W \setminus S \cap \text{Image}(C) \neq \{\}\}$.

Proof: If Agent(C) is nonempty and Env(C) is empty, $S \in \text{Ensure}(C)$ for all $S \subseteq W$ vacuously.

If Agent(C) is nonempty and Env(C) = {e} is a singleton, every $S \subseteq W$ that intersects Image(C) nontrivially is in Ensure(C), since if $w \in S \cap \text{Image}(C)$, there must be some $a \in A$ such that $a \cdot e = w$, this a satisfies $a \cdot e' \in S$ for all $e' \in E$. Conversely, if S and Image(C) are disjoint, no $a \in A$ can satisfy this property. The result for Ctrl then follows trivially from the result for Ensure. □

5.3. A Suggestion of Time

Cartesian frames as we've been discussing them are agnostic about time. Possible agents, environments, and worlds could represent snapshots of a particular moment in time, or they could represent lengthy processes.

The fact that an agent's controllables and observables are disjoint, however, suggests a sort of arrow of time, where facts an agent can observe must be “before” the facts that agent has control over. This hints that we may be able to use Cartesian frames to formally represent temporal relationships.

One reason it would be nice to represent time is that we could model agents that repeatedly learn things, expanding their set of observables. Suppose that in some frame C, Agent(C) includes choices the agent makes over an entire calendar year. Agent(C)'s observables would only include the facts the agent can condition on at the start of the year, when it's first able to act; we haven't defined a way to formally represent the agent learning new facts over the course of the year.

It turns out that this additional temporal structure *can* be elegantly expressed using Cartesian frames. We will return to this topic in the very last post in this sequence. For now, however, we only have this hint that particular Cartesian frames have something like a "before" and "after."

6. Why Cartesian Frames?

The goal of this sequence will be to set up the language for talking about problems using Cartesian frames.

Concretely, I'm writing this sequence because:

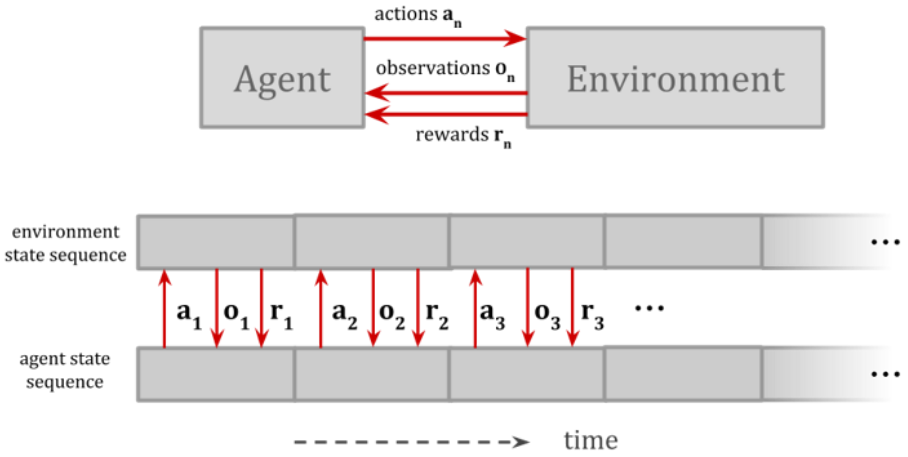
- I've recently found that I have a new perspective to bring to a lot of other MIRI researchers' work. This perspective seems to me to be captured in the mathematical structure of Cartesian frames, but it's the new perspective rather than the mathematical structure per se that seems important to me. I want to try sharing this mathematical object and the accompanying philosophical interpretation, to see if it successfully communicates the perspective.
- I want collaborators to work with on Cartesian frames. If you're a math person who finds the things in this sequence exciting, I'd be interested in talking about it more. You can comment, PM, or email me.

- I want help with paradigm-building, but I also want there to be an ecosystem where people do normal science within this paradigm. I would consider it a good outcome if there existed a decent-sized group of people on the AI Alignment Forum and LessWrong for whom it makes just as much sense to pull out the Cartesian frames paradigm as it makes to pull out the cybernetic agent paradigm.

Below, I will say more about the cybernetic agent model and other ideas that helped motivate Cartesian frames, and I will provide an overview of upcoming posts in the sequence.

6.1. Cybernetic Agent Model

The cybernetic agent model describes an agent and an environment interacting over time:



In "Embedded Agency," Abram Demski and I noted that cybernetic agents like Marcus Hutter's AIXI are dualistic, whereas real-world agents will be embedded in their environment. Like a Cartesian soul, AIXI is crisply separated from its environment.

The dualistic model is often useful, but it's clearly a simplification that works better in some contexts than in others. One thing it would be nice to have is a way to capture the useful things about this simplification, while treating it as a high-level approximation with known limitations — rather than treating it as ground truth.

Cartesian frames carve up the world into a separate "agent" and "environment," and thereby adopt the basic conceit of dualistic Hutter-style agents. However, they treat this as a "frame" imposed on a more embedded, naturalistic world.[2]

Cartesian frames serve the same sort of intellectual function as the cybernetic agent model, and are intended to supersede this model. Our hope is that a less discrete version of ideas like "agent," "action," and "observation" will be better able to tolerate edge cases. E.g., we want to be able to model weirder, loopier versions of “inputs” that operate across multiple levels of description.

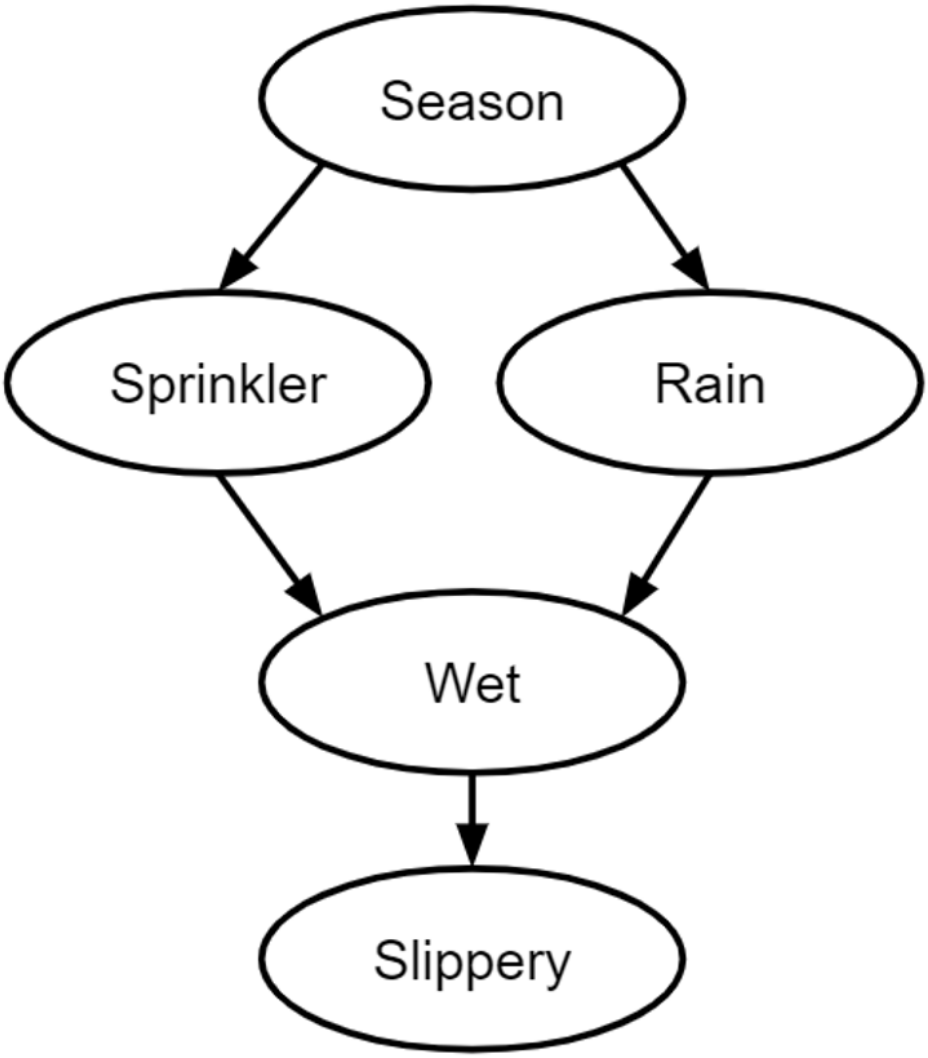
We will also devote special attention in this sequence to subagents, which are very difficult to represent in traditional dualistic models. In game theory, for example, we carve the world into different "agent" and "non-agent" parts, but we can't represent nontrivial agents that intersect other agents. A large part of the theory in this sequence will be giving us a language for talking about subagents.

6.2. Deriving Functional Structure

Another way of summarizing this sequence is that we'll be applying *reasoning* like Pearl's to *objects* like game theory's, with a *motivation* like Hutter's.

In Judea Pearl's causal models, you are given a bunch of variables, and an enormous joint distribution over the variables. The joint distribution is a large object that has a relational structure as opposed to a functional structure.

You then deduce something that looks like time and causality out of the combinatorial properties of the joint distribution. This takes the form of causal diagrams, which give you functions and counterfactuals.



This has some similarities to how we'll be using Cartesian frames, even though the formal objects we'll be working with are very different from Pearl's. We want a model that can replace the cybernetic agent model with something more naturalistic, and our plan for doing this will involve deriving things like time from the combinatorial properties of possible worlds.

We can imagine the real world as an enormous static object, and we can imagine zooming in on different levels of the physical world and sometimes seeing things that look like local functions. ("Ah, no matter what the rest of the world looks like, I can compute the state of Y from the state of X, relative to my uncertainty.") Switching which part of the world we're looking at, or switching which things we're lumping together versus splitting, can then change which things look like functions.

Agency itself, as we normally think about it, is functional in this way: there are multiple "possible" inputs, and whichever "option" we pick yields a deterministic result.

We want an approach to agency that treats this functional behavior less like a unique or fundamental feature of the world, and more like a special case of the world's structure in general — and one that may depend on what we're splitting or lumping together.

"We want to apply Pearl-like methods to Cartesian frames" is also another way of saying "we want to do the formal equivalent of inferring extensive-form games from normal-form games," our summary from before. The analogy is:

	base information	derived information
causality	joint probability distribution	causal diagram
games	normal-form game	extensive-form game
agency	Cartesian frame	control, observation, subagents, time, etc.

The game theory analogy is more relevant formally, while the Pearl analogy better explains why we're interested in this derivation.

Just as notions of time and information state are basic in causal diagrams and extensive-form games, so are they basic in the cybernetic agent model; and we want to make these aspects of the cybernetic agent model derived rather than basic, where it's possible to derive them. We also want to be able to represent things like subagents that are entirely missing from the cybernetic agent model.

Because we aren't treating high-level categories like "action" or "observation" as primitives, we can hope to end up with an agent model that will let us model more edge cases and odd states of the system. A more derived and decomposable notion of time, for example, might let us better handle settings where two agents are both trying to reach a decision based on their model of the other agent's future behavior.

We can also hope to distinguish features of agency that are more description-invariant from features that depend strongly on how we carve up the world.

One philosophical difference between our approach and Pearl's is that we will avoid the assumption that the space of possible worlds factors nicely into variables that are given to the agent. We want to instead just work with a space of possible worlds, and derive the variables for ourselves; or we may want to work in an ontology that lets us reason with multiple incompatible factorizations into variables.[3]

6.3. Contents

The rest of the sequence will cover these topics:

2. Additive Operations on Cartesian Frames - We talk about the category of Chu spaces, and introduce two additive operations one can do on Cartesian frames: sum $\oplus$, and product &. We talk about how to interpret these operations philosophically, in the context of agents making choices to affect the world. We also introduce the small Cartesian frame 0, and its dual $0^* = \top$.

3. Biextensional Equivalence - We define homotopy equivalence $\simeq$ for Cartesian frames, and introduce the small Cartesian frames null, 1_S, and $\bot_S$.

4. Controllables and Observables, Revisited - We use our new language to redefine controllables and observables.

5. Functors and Coarse Worlds - We show how to compare frames over a detailed world model W and frames over a coarse version of that world model V. We demonstrate that observability is a function not only of the observer and the observed, but of the level of description of the world.

6. Subagents of Cartesian Frames - We introduce the notion of a frame C whose agent is the subagent of a frame D, written $C \triangleleft D$. A subagent is an agent playing a game whose stakes are another agent's possible choices. This notion turns out to yield elegant descriptions of a variety of properties of agents.

7. Multiplicative Operations on Cartesian Frames - We introduce three new binary operations on Cartesian frames: tensor $\otimes$, par $\parr$, and lollipop $\multimap$.

8. Sub-Sums and Sub-Tensors - We discuss spurious environments, and introduce variants of sum, ⊞, and tensor, ⊠, that can remove some (but not too many) spurious environments.

9. Additive and Multiplicative Subagents - We discuss the difference between additive subagents, which are like future versions of the agent after making some commitment; and multiplicative subagents, which are like agents acting within a larger agent.

10. Committing, Assuming, Externalizing, and Internalizing - We discuss the additive notion of producing subagents and sub-environments by *committing* or *assuming*, and the multiplicative notion of *externalizing* (moving part of the agent into the environment) and *internalizing* (moving part of the environment into the agent).

11. Eight Definitions of Observability - We use our new tools to provide additional definitions and interpretations of observables. We talk philosophically about the difference between defining what's observable using product and defining what's observable using tensor, which corresponds to the difference between updateful and updateless observations.

12. Time in Cartesian Frames - We show how to formalize temporal relations with Cartesian frames.

I'll be releasing new posts most non-weekend days between now and November 11.

As Ben noted in his announcement post, I'll be giving talks and holding office hours this Sunday at 12-2pm PT and the following three Sundays at 2-4pm PT, to answer questions and discuss Cartesian frames. Everyone is welcome.

The online talks, covering much of the content of this sequence, will take place **this Sunday at 12pm PT** (~~Zoom link~~ added: recording of the talk) and **next Sunday at 2pm PT**.

This sequence is communicating ideas I have been developing slowly over the last year. Thus, I have gotten a lot of help from conversation with many people. Thanks to Alex Appel, Rob Bensinger, Tsvi Benson-Tilsen, Andrew Critch, Abram Demski, Sam Eisenstat, David Girardo, Evan Hubinger, Edward Kmett, Alexander Gietelink Oldenziel, Steve Rayhawk, Nate Soares, and many others.

Footnotes

1. This assumes a non-empty Agent(C). Otherwise, Image(C) could be empty and therefore a subset of S, even though S is not ensurable (because you need an element of Agent(C) in order to ensure anything). ↩

2. This is one reason for the name "Cartesian frames." Another reason for the name is to note the connection to Cartesian products. In linear algebra, a frame of an inner product space is a generalization of a basis of a vector space to sets that may be linearly dependent. With Cartesian frames, then, we have a Cartesian product that projects onto the world, not necessarily injectively. (Cartesian frames aren't actually "frames" in the linear-algebra sense, so this is only an analogy.) ↩

3. This, for example, might let us talk about a high-level description of a computation being "earlier" in some sort of logical time than the exact details of that same computation.

Problems like agent simulates predictor make me think that we shouldn't treat the world as factorizing into a single "true" set of variables at all, though I won't attempt to justify that claim here. ↩

The Felt Sense: What, Why and How

While LW has seen previous discussion of Focusing, I feel like there has been relatively limited discussion of *the felt sense* - that is, the thing that Focusing is actually accessing.

Everyone accesses felt senses all the time, but most people don't know that they are doing it. I think that being able to make the skill more explicit is really valuable, and in this post I'm going to give lots of examples of why that is and what you can do with it.

Hopefully, after I'm done, you will not only know what a felt sense is (if you didn't already), but also will have difficulty understanding how you ever got by *without* this concept.

Examples of felt senses

The term "felt sense" was originally coined by the psychologist Eugene Gendlin, as a name for something that he found his clients to be accessing in their therapy sessions. Here are some examples of felt senses:

- Think of some person you know, maybe imagining what it feels to be like in the same room as them. You probably have some “sense” of that person, of what it is that they *feel like*.
- Likewise if you think of some fictional universe, it has something of its own feel. *Harry Potter* feels different from *Star Wars* feels different from *Game of Thrones* feels different from *James Bond*.
- Sometimes you will have a word “right on the tip of your tongue”; it’s as if the word is *almost there*, but you can’t quite reach it. When you do, you just *know* that it's the right word - because the "shape" of the word matches the one you were reaching for before.

The felt senses of pictures

Here are are a few pictures that I recently collected from the Facebook group "Steampunk Tendencies":

How do you feel when you look at these pictures? What's the general vibe that unites all of these pictures?

Likely you can find quite a few. If I put aside the words "steampunk" and "Victorian", next I get the word "mechanical". "Dark" also feels fitting.

Whatever the vibe that you get, it's probably something different than the one you get from this collection of images:

Look at the first set of images, then the second. How does it feel when you switch looking from one to the other? What kinds of changes are there in your mind and your body?

I like both sets, but looking from one to the other, I notice that the forest images make me feel like my mind is opening up, whereas the steampunk ones make it close a little. Comparing the two, I feel like there's some slightly off-putting vibe in the steampunk set, that makes me prefer looking at the forest images - which I would not have noticed if I hadn't viewed them side to side. (I am guessing that some readers will have the opposite experience, of finding the forest ones off-putting compared to the steampunk ones.)

Emotional and mental states as felt senses

Internal emotional and mental states can also have their own felt senses. That shouldn't be very surprising, since your experience of e.g. a set of pictures *is* an internal mental state. Here are a few examples of felt senses from alkjash:

- When I solve a problem in a creative way (e.g. fix posture by turning in the shower), there's a sensation of enlightenment at the back of my head which literally feels like my skull is opening up. The words to this feeling are "I've discovered a new dimension!"
- I sometimes sit slouched over in bed for hours at a time browsing Facebook or Reddit, playing video games, or binge-watch a season of a TV show. After getting up from the slouch, my whole body is enveloped in a haze of laziness and decay. The zombie haze is thickest inside my ribs. The words to this pressure are "Symptoms of the spreading corruption."

- A piece of my social anxiety forms a hard barrier that pushes against the center of my chest. I learned the words to this feeling from a post by Zvi: “Conform! Every time you walk outside the norm, think about the implicit accusation you’re making against everyone who didn’t try it.”

Sometimes it's easy to come up with words to describe a felt sense, but typically it takes a bit of time to find exactly the right ones. I expect that it took some time for alkjash to find evocative descriptions such as the above.

Here's Duncan Sabien describing the experience of honing down on a particular felt sense (I've edited out some excellent elaborations and pictures that were included between these lines; the whole post is recommended reading):

> *Okay, so there’s clearly SOMETHING bothering me. And it’s got something to do with Cameron.*
>
> *Have we been fighting a lot?*
>
> *No, that’s not it at all.*
>
> *It’s more like — like — ugh, like I never know what to say?*
>
> *No, it’s like I **have** to say the right things, or else.*
>
> *It’s like — if I say the **wrong** thing, then everything falls apart and it’s all ruined.*
>
> *And it’s like I’m the only one? Like, **Cameron** doesn’t have to pay attention, just **me**. Cameron gets to —*
>
> *— to —*
>
> *— to **relax.** That’s it. Yeah. It feels like I’m the only one who doesn’t get to relax.*

Felt sense as the layer below language

Mark Lippmann, in his document "Folding" (currently deprecated) proposes that the felt sense (or the felt meaning, as he calls it) exists as a layer of information "below" language. He gives the following examples:

- Pick a word, such as "yogurt", and say it many times over: "yogurt, yogurt, yogurt...". Now eventually it may feel like the word has "lost its meaning"; the verbal handle of "yogurt" has become disconnected from the felt sense it used to be associated with.
- It's not just individual words that are connected to felt senses; you can also know the meaning of a sentence that you just read, or have a sense of what a particular paragraph was saying.
- Sometimes, if you are working on a document close to a deadline or trying to read something when you are tired, you may find that the thing becomes "slippery". Your eyes might repeatedly pass over the same words, but you don't understand what they are saying. You are failing to extract the felt sense holding the meaning of the text.

He notes that a felt sense can also be experienced as a thing or place to return to:

> When you say something, and it doesn’t come out right, you try again. Where your mind goes before you try again, that’s felt meaning.
>
> When someone says, can you explain *that* in different words? Your mind goes back to *that*, in other words, felt meaning.

> When someone says, what do you mean by *that*? Your mind goes back to *that*, in other words, felt meaning.
>
> When you're writing something, and it's hard, what you're searching for is felt meaning. Where you're searching is where felt meaning will be.
>
> When you're writing something, and it's easy, you're drawing upon felt meaning.
>
> When you lose your train of thought, you've lost your sense of the felt meaning you were speaking from. When you remember *what* you were saying, the what is felt meaning.

Why learn to tap into felt senses?

Why are felt senses important? Well, a felt sense looks like it's coming from some deeper information-processing layer in your brain: if you do anything at all (such as read or talk), you are tapping into that layer.

Everyone accesses felt senses all the time. But you can learn to more explicitly pay attention to the fact that you're doing it, and thus do it more effectively. This has a *great number* of benefits.

It's useful for communication

The bits about text and meaning might already have suggested this: you can express yourself most clearly when you have a good handle on the felt sense of your intended meaning. If you've ever found yourself saying "no, that's not quite what I meant, I mean it more like...", then you've been trying to connect with a felt sense better.

Furthermore, *different* felt senses let you communicate *different* things. Matt Goldenberg recently interviewed me on what I do when I'm trying to write something that communicates across perspectives. As a result of his questioning, I ended up describing it as something like this; I've bolded each occasion when I've made reference to a felt sense:

> When I see people talking past each other, I get this frustration, like... let's say someone is trying to explain meditation to someone skeptical. I see the skeptic asking questions, and I can get **a sense of what their model is like and which gaps in their model they are trying to fill** with those questions. And then the other person **doesn't seem to get that**, and says something else. It feels like there are two different perspectives on the issue. They are **almost physical shapes with some overlap but which don't quite align**, and I get an **urge to build a bridge between them**, to get those **two perspectives joined together**.
>
> So then I start getting some ideas about it, of what kind of an explanation **would fit that hole in the skeptic's model**, and what **would make those perspectives sync up**, and there's a **sense of harmony and beauty** in what that finished explanation would feel like. Then I have all of those scattered ideas and I note them down and try to find something that would **feel like a unifying framework**, where the ideas **wouldn't feel separate from each other** but rather **be part of a coherent structure**. And I try to make use of that unifying framework to write it so that there's a **smooth flow** of one idea to the next, so that **each thing flows naturally to the next**.
>
> And while I'm writing it, I make sure to come back to a **sense of my target audience**, and try to have a **feel of what they would think of my explanation.** Sometimes when I'm writing in essay it starts **feeling hard to say who I'm writing to**, and then I might end up writing something on Twitter or Facebook, where I **have a clearer sense of who's going to read this**, and then that might let me make progress.

I hadn't explicitly thought about it in those terms, but Matt's poking helped me make more aware of all the felt senses that I was following: and he was explicitly digging them out of me in order to teach them to others. Having them, I can turn them into explicit guidelines to ask myself (or others), such as:

- If you need ideas, is there any particular situation whose felt sense gives you ideas, such as getting frustrated by two people failing to communicate and seeing what they could be saying instead?
- If you have several ideas, do you feel like you could get a sense of how to explain them in such a way that they flow from one thing to another?
- Do you have a clear feeling of who you are writing to? If not, could you get one, such as by writing something on a social media or as a forum post or imagining a specific person reading it?

It's useful for creating and appreciating art

As an example of different felt senses being useful for communicating different things, Logan Strohl writes about the use of them in art:

> Pick an expressive medium. Could be sketching, poetry, music, whatever.
>
> Then, get in touch with a felt sense. You don't have to name it. But try to get inside of it.
>
> What is "get inside of it"? Right now there's a tightness in my solar plexus. I can describe it "from the outside" like so: It's the bottom of a sort of hot, slightly vibrating rod of sensation that goes from my solar plexus to the middle of my throat. The sensation responds to awareness of my immediate auditory environment (I'm in a coffee shop); the solar plexus tightness gets tighter when I pay attention to the tapping of a metal spoon against a metal jar, and starts to wobble a little when I pay attention to the music in the background.
>
> Rather than describing it from the outside, I can also let the felt sense express itself "from the inside". This is a kind of attentional trick, I think, which seems to involve setting down my personhood story and letting the felt sense consume awareness.
>
> Then, while "inside" of the felt sense, I can begin to act on my creative medium. If I choose (just a few) words, the solar plexus felt sense types this:
>
> wobble siren sharp and hot fight for warming Persian music hold ready parking alarm to protect changing changing changing nothing safe

Logan then goes on to describe the process of drawing a picture from inside the felt sense, letting each line resonate against the felt sense and only draw things which feel true to it.

> ... this is what artists are actually doing when they create things. They're doing additional stuff too, because what I've described is merely expression, and art is a kind of communication. Communication is a refined form of expression that usually involves design and editing in addition to expression. But I think the unrefined expression is at the core of art.

This matches my experience when I'm doing role-playing or writing fictional characters: each character has their own felt sense, and writing them is often about getting inside that felt sense. Characters may start with a weak felt sense, but "take on a life of their own", when that sense gets fleshed out and becomes strong enough.

Sometimes I have difficulty expressing a particular character, in which case I have lost my connection to their felt sense - their "essence", so to speak. On a few occasions, I have

intentionally created characters by taking aspects of the felt senses of my friends, and blended them together into a new whole that feels right.

I have also heard of poetry being described essentially as trying to convey a felt sense through words.

I think much of art is basically all about evoking felt senses. If you have that as an explicit concept, you can look at a piece of art that you like, and attempt to describe its felt sense in greater detail. That may help you dig deeper into what about it you like, and make you *feel* that thing you like more.

It's good for knowing what you want

Tapping into felt senses associated with the things that you want feels valuable in general. Rossin writes:

> I used to think of myself as someone who was very spontaneous and did not like to plan or organize things any more or any sooner than absolutely necessary. I thought that was just the kind of person I am and getting overly organized would just feel wrong.
>
> But I felt a lot of aberrant bouts of anxiety. I probably could have figured out the problem through standard Focusing but I was having trouble with the negative feeling. And I found it easier to focus on positive feelings, so I began to apply Focusing to when I felt happy. And a common trend that emerged from good felt senses was a feeling of being in control of my life. And it turned out that this feeling of being in control came from having planned to do something I wanted to do and having done it. I would not have noticed that experiences of having planned well made me feel so good through normal analysis because that was just completely contrary to my self-image. But by Focusing on what made me have good feelings, I was able to shift my self-image to be more accurate. I like having detailed plans. Who would have thought? Certainly not me.
>
> Once I realized that my self-image of enjoying disorganization was actually the opposite of what actually made me happy I was able to begin methodically organizing and scheduling my life. Since then, those unexplained bouts of anxiety have vanished and I feel happier more of the time.

Sometimes I get the feeling that a thing that I'm doing seems good on paper, but in practice it just feels like a demotivating chore. Often this means that the thing that I *think* I'm going for is not the thing that my brain is *actually* optimizing for, and it's predicting that the project in question will not fulfill its *actual* optimization goal. If I can then lean into the felt sense of what I actually want, then I will feel more motivated to pursue it.

For example, recently I have been trying to debug my aversion towards dating sites. There seem to be several components to that aversion, but one in particular is a vibe of "I don't expect this to really work" that I tend to get at the point when I start to browse other people's profiles.

Which raised the question of... doesn't work for *what*, exactly? Not just "for getting into a relationship"; what's the deeper desire that makes me want a relationship in the first place?

So far I had been kind of waffling back and forth on the question of "do I want children", so my search filters had included people with various answers to that question. But then I accidentally ended up doing a search where that answer was required to be "yes", and noticed that the kinds of profiles I got in response - or just consistently seeing "wants children" on all the results that I got - gave me a much felt sense of *this could lead to somewhere promising*.

The main thing doesn't seem to be *just* the thought of having children, but also something about the potential partners *generally being the type of people who want children [due to some personality trait which I haven't verbalized yet, but which was more apparent in the profiles that I started seeing]*... which started making the whole dating site thing seem more appealing again.

In a way, finding this particular felt sense when I was feeling demotivated, feels like the same kind of thing as finding "who was my target audience for this piece of writing again" when I've been feeling demotivated by writing. In either case the brain is pursuing some optimization target, but cannot proceed and reacts by demotivation if a clear optimization target cannot be found.

Michael Smith (Valentine) has recently been talking about leaning into pleasure, and of how society tends to cause psyches to be built around avoiding pain rather than pursuing joyful bright desire. I'm coming to agree with Rossin's post in that we probably tend to undervalue using felt senses to look into the positive, and don't pursue the "bright desire" as much as we could - in part because we haven't spent time really digging into the felt senses of enjoyment. (Though it needs to be stated that often one's mind has reasons for why it considers it *necessary* to feel bad, so it does often make sense to investigate those reasons *first*.)

Generally, your aesthetics encode information and assumptions about what your brain considers valuable [1 2 3]. Aesthetics are to a large extent expressed in felt senses.

It's useful for figuring out what's bothering you

The "standard" use for the felt sense, from Gendlin's original book, is figuring out what bothers you. Duncan Sabien already gave us an example of this previously, when figuring out why an imaginary "Cameron" was bothering him. Listening to felt senses is the foundation of Focusing-Oriented Psychotherapy, as well as practices such as Internal Double Crux, Internal Family Systems, Coherence Therapy, and experiential forms of therapy in general.

This excerpt from Unlocking the Emotional Brain described "Richard" getting in contact with a felt sense of what his mind thought would happen if he expressed confidence:

> **Richard:** Now I'm feeling really uncomfortable, but-it's in a different way.
>
> **Therapist:** OK, let yourself feel it - this different discomfort. *[Pause.]* See if any words come along with this uncomfortable feeling.
>
> **Richard:** *[Pause.]* Now they hate me.
>
> **Therapist:** "Now they hate me." Good. Keep going: See if this really uncomfortable feeling can also tell you *why* they hate you now.
>
> **Richard:** *[Pause.]* Hnh. Wow. It's because… now I'm… an arrogant asshole… like my father… a totally self-centered, totally insensitive know-it-all.
>
> **Therapist:** Do you mean that having a feeling of confidence as you speak turns you into an arrogant asshole, like Dad?
>
> **Richard:** Yeah, exactly. Wow.

As a result of having surfaced this felt sense, Richard was then able to question it and revise the belief contained in it.

It helps you know when you are triggered

I think of "being triggered" meaning something like "a part of you tries to force a particular outcome even if your other parts would disagree of this being a good idea" (this feels closely related to the Buddhist notion of craving for specific outcomes).

If I think about situations where I wish I had acted differently, they include things like

- I told my cousin that I was interesting in moving something closer to psychology, career-wise. My cousin said something that I thought implied she didn't think I knew much about psychology, reflecting a very old model of me. I felt a strong desire to correct that misconception, and there was something of a sharp forcefulness in that response, trying to *force* her into thinking the right thing.
- I overheard some parents treating their child in a way that felt to me hurtful towards the child, and there was a desire to intervene and *force* them to act differently towards their child. (But of course I knew that it wouldn't do any good.)
- I got a message that I would have preferred not to receive or read, but for as long as it remained unread, there was an insistent tugging, as if something was trying to force it to become read, and another something trying to force it *not* to be read.

Besides the *specific* and somewhat different felt senses in all three of those situations, there's also a shared *general* felt sense of... some sort of *wrongness*, as if my mind feels that there is *something wrong about the world*, which *needs to be fixed*. As long as that part is trying to force that fix, I can't think or react entirely freely.

When I'm triggered, it's not always clear to me: I might be so strongly triggered that the thought just seems like absolute truth to me, or the triggering might be subtle enough that it might pass almost unnoticed. But if I pay attention to the sense of wrongness that I typically get when triggered, I can have something of a trigger-action plan of "notice when I am triggered, and pause to see what the appropriate response could be".

The opposite of the wrongness of being triggered feels something like the Internal Family Systems notion of "the 8 Cs of being in Self": "confidence, calmness, creativity, clarity, curiosity, courage, compassion, and connectedness". Noticing that I do *not* have those kinds of felt senses also helps to notice when I'm triggered.

Conclusion

There are a number of explanations of how to do Focusing, that is, tap into your felt senses. Some here on LW include ones by (particularly recommended!) Duncan Sabien, alkjash, and Mark Xu. The Focusing Institute offers this page of six steps, which are further elaborated on Eugene Gendlin's book.

My personal favorite set of formal Focusing instructions is in Ann Weiser Cornell's The Power of Focusing; for some reason, everyone always seems to recommend the original Focusing book, even though AWC's instructions feel ten times better to me.

That said, I always feel like formal Focusing instructions risk making the felt sense feel like this exotic super-special thing, and then you might end up wondering things like "is this *really* the felt sense" way too much. Remember: the felt sense is nothing special. If you understand what this sentence is saying, you already have access to a felt sense - the one which tells you what the meaning of this sentence is.

Thus, my favored approach to tapping into a felt sense is just "imagine I was explaining this feeling that I have to someone else, taking the time to find the words and description that resonate the most".

In other words, in explaining felt senses, I would recommend you to go not for the felt sense of "explaining some exotic and special thing deep in my subconscious", but rather for the felt sense of "explaining a thing in my everyday experience and just wanting to find exactly the right words for it".

What are some beautiful, rationalist artworks?

So you can now drag-and-drop images into comments. (Thanks, LessWrong dev team!)

Hence, this post is an excuse to build a beautiful, inspiring, powerful — and primarily visual — comment section.

Let's celebrate all that is great about the Art of Rationality, with images.

Rules

- **Each answer must contain a picture. No links!**

It should be possible to just scroll through the comments and adore the artwork. There shouldn't be any need to click-through to other pages. (Think of it like a Pinterest board, if you've ever seen those.)

Adding text is fine, but consider doing it in a comment underneath your image, so it can be collapsed.

- **Pictures should be somehow relate to the Art of Rationality, as practiced on LessWrong.**

Allowed: a breathtaking shot of a SpaceX launch; that solemn shot of Petrov deep in thought, gazing out his window; a painting of Galileo spearheading empiricism against the inquisition, ...

Not allowed: a random pretty mountain; the Mona Lisa; abstract expressionism, ...

I'll be liberal with this condition if you can give a good justification for why you chose your piece.

- **Pictures should be beautiful art *independently* of their relation to rationality.**

Allowed: an exquisite shot of some piece of elegantly engineered machinery; a richly colourful and swirling galaxy, ...

Not allowed: a random picture of Einstein and Gödel hanging out; a low-resolution photo of a galaxy which is cool *because it represents an important advance in astronomy*, but which in-and-of-itself just looks like some lame computer graphics; Petrov's own tourist photos, ...

- **Don't be a jerk, but do note if you think something is a major conflict with a virtue.**

Probably goes without saying... but don't be a pretentious art critic. The point of this thread is to pay tribute to those virtues that keep us striving to leave this world in a

better place than we found it, guided by the Light of Science. Don't shout over the music.

That being said, I do care about pictures actually representing rationality. For example, take that photo of the exhausted surgeon after a 23h heart transplant. If it turned out (hypothetical) to have been the result of really poor utilitarian calculations, and actually is in direct conflict with some of our virtues: I think it's important to note that.

Note: I'm certainly *not* saying that the above rules are all that rationalist art is about. I'm just going for a particular vision with this comment field. Other posts can enforce other visions. :)

Message Length

Someone is broadcasting a stream of bits. You don't know why. A 500-bit-long sample looks like this:

```
01100110110101011011111100001001110000100011010001101011011010000001010000001010
10100111101000101111010100100101010010101010101000010100110101010011111111010101
01010101011111110101011010101101111101010110110101010100000000110111110000011101
11100000000000000111110101011010101010100101010110101010110011100110000110011010 1
11111111111111111100011001011010011010101010101011000000101010111011010100101100 11
11110101111011101000101010101110011111010001101101010101101011000101100000101010
10011001101010101111...
```

The thought occurs to you to do Science to it—to ponder if there's some way you could better predict what bits are going to come next. At first you think you can't—it's just a bunch of random bits. You can't predict it, because that's what random *means*.

Or does it? True, if the sequence represented flips of a fair coin—every flip independently landing either `0` or `1` with exactly equal probability—then there would be no way you could predict what would come next: any continuation you could posit would be exactly as probable as any other.

But if the sequence represented flips of a *biased* coin—if, say, `1` came up 0.55 of the time instead of exactly 0.5—then it would be possible to predict better or worse. Your best bet for the next bit in isolation would always be `1`, and you would more strongly anticipate sequences with slightly more `1`s than `0`s.

You count 265 `1`s in the sample of 500 bits. *Given* the hypothesis that the bits were generated by a fair coin, the number of `1`s (or without loss of generality, `0`s) would be given by the binomial distribution $\binom{500}{k}(0.5)^k(0.5)^{500-k}$, which has a standard deviation of $\sqrt{500 \cdot 0.5^2} = \sqrt{125} \approx 11.18$, so your observation of $265 - 250 = 15$ excess `1`s is about $\frac{15}{11.18} \approx 1.34$ standard deviations from the mean—well within the realm of plausibility of happening by chance, although you're at least slightly *suspicious* that the coin behind these bits might not be quite fair.

... that is, if it's even a coin. You love talking in terms of shiny, if hypothetical, "coins" rather than stodgy old "independent and identically distributed binary-valued random variables", but looking at the sample again, you begin to *further* doubt whether the bits are independent of each other. You've heard that humans are biased to overestimate the frequency of alternations (`101010`...) and underestimate the frequency of consecutive runs (`00000`... or `11111`...) in "truly" (uniformly) random data, but the 500-bit sample contains a run of 13 `0`s (starting at position 243) *and* a run of 19 `1`s (starting at position 319). You're not immediately sure how to calculate the probability of that, but your gut says that should be very unlikely given the biased-coin model, even after taking into account that human guts aren't very good at estimating these things.

Maybe not everything in the universe is a coin. What if the bits were being generated by a Markov chain—if the probability of the next bit depended on the value of the one just before? If a `0` made the *next* bit more likely to be a `0`, and the same for `1`, that would make the `00000`... and `11111`... runs less improbable.

Except ... the sample *also* has a run of 17 alternations (starting at position 153). On the "fair coin" model, shouldn't that itself be $2^{17-13} = 16$ times as suspicious as the run of 13 `0`s and $2^{17-19} = \frac{1}{4}$ as suspicious as the run of 19 `1`s which led you to hypothesize a Markov chain?—or rather, 8 and $\frac{1}{8}$ times as suspicious, respectively, because there are two ways for an alternation to occur (`0101010`... or `1010101`...).

A Markov chain in which a `0` or `1` makes another of the same more likely, makes alternations *less* likely: the Markov chain hypothesis can only make the consecutive runs look less surprising at the expense of making the run of alternations look *more* surprising.

So maybe it's all just a coincidence: the broadcast is random—whatever that means—and you're just apophenically pattern-matching on noise. Unless ...

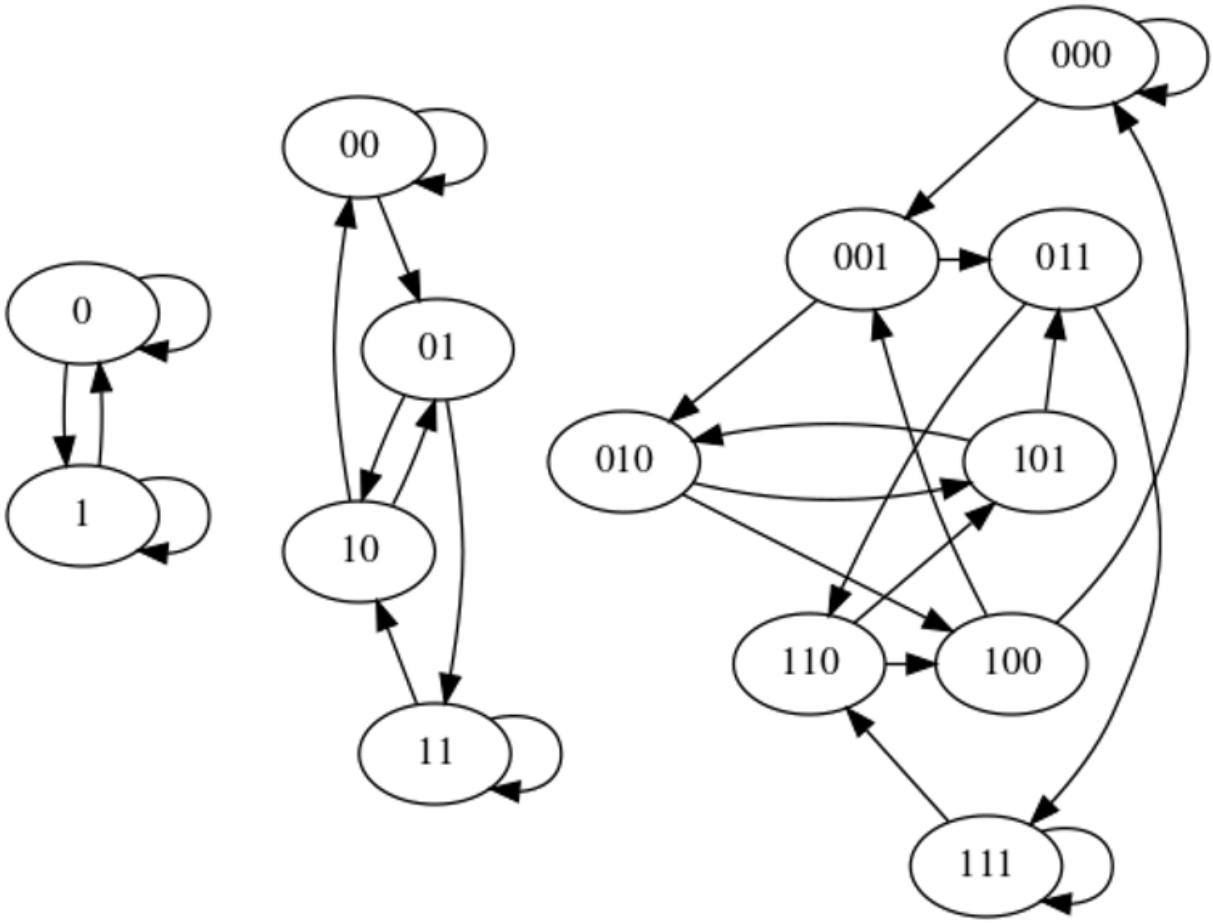
Could it be that some things in the universe are *neither* coins *nor* Markov chains? You don't *know* who is broadcasting these bits or why; you called it "random" because you didn't see any obvious pattern, but now that you think about it, it would be pretty weird for someone to just be broadcasting random bits. Probably the broadcast is something like a movie or a stock ticker; if a close-up sample of the individual bits looks "random", that's only because you don't know the codec.

Trying to guess a video codec is *obviously* impossible. Does that kill all hope of being able to better predict future bits? *Maybe* not. Even if you don't know what the broadcast is really for, there might be some nontrivial *local* structure to it, where bits are statistically related to the bits nearby, like how a dumb encoding of a video might have consecutive runs of the same bit-pattern where a large portion of a frame is the same color, like the sky.

Local structure, where bits are statistically related to the bits nearby ... kind of like a Markov chain, except in a Markov chain the probability of the next state only depends on the *one* immediately before, which is a pretty narrow notion of "nearby." To broaden that, you could imagine the bits are being generated by a *higher-order* Markov chain, where the probability of the next bit depends on the previous n bits for some specific value of n.

And *that's* how you can explain mysteriously frequent consecutive runs and alternations. If the last *two* bits being `01` (respectively `10`) makes it more likely for the next bit to be `0` (respectively `1`), *and* the last two bits being `00` (respectively `11`) makes it more likely for the next bit to be `0` (respectively `1`), then you would be more likely to see both long `0000`... or `1111`... consecutive runs *and* `01010`... alternations.

A biased coin is just an n-th-order Markov chain where $n = 0$. An n-th-order Markov chain where $n > 1$, is just a first-order Markov chain where each "state" is a tuple of bits, rather than a single bit.



Everything in the universe is a Markov chain!—with respect to the models you've considered so far.

"The bits are being generated by a Markov chain of some order" is *a* theory, but a pretty broad one. To make it concrete enough to test, you need to posit some specific order n, and, given n, specific parameters for the next-bit-given-previous-n probabilities.

The $n = 0$ coin has one parameter: the bias of the coin, the probability of the next bit being `0`. (Or without loss of generality `1`; we just need one parameter p to specify the probability of one of the two possibilities, and then the probability of the other will be $1 - p$.)

The $n = 1$ ordinary Markov chain has two parameters: the probability of the next bit being (without loss of generality) `0` *given* that the last bit was a `0`, and the probability of the next bit being (without loss of ...) `0` *given* that the last bit was a `1`.

The $n = 2$ second-order Markov chain has four parameters: the probability of the next bit being (without loss ...) `0` *given* that the last two bits were `00`, the probability of the next bit being (without ...) `0` *given* that the last two bits were `01`, the probability of the next bit being—

Enough! You get it! The n-th order Markov chain has 2^n parameters!

Okay, but then how do you guess the parameters?

For the $n = 0$ coin, your *best guess* at the frequency-of-`0` parameter is going to just be the frequency of `0`s you've observed. Your best guess could easily be wrong, and probably is: just because you observed 235/500 = 0.47 `0`s, doesn't mean the parameter *is* 0.47: it's probably somewhat lower or higher, and your sample got more or fewer `0`s than average just by chance. But positing that the observed frequency is the actual parameter is the maximum likelihood estimate—the single value that *most* makes the data "look normal".

For $n \geq 1$, it's the same idea: your best guess for the frequency-of-`0`-after-`0` parameter is just the frequency of `0` being the next bit, among all the places where `0` was the last bit, and so on.

You can write a program that takes the data and a degree n, and computes the maximum-likelihood estimate for the n-th order Markov chain that might have produced that data. Just slide an $(n+1)$-bit "window" over the data, and keep a tally of the frequencies of the "plus-one" last bit, for each of the 2^n possible n-bit patterns.

In the Rust programming language, that looks like following. (Where the representation of our final theory is output as a map (`HashMap`) from $n+1$-bit-patterns to frequencies/parameter-values (stored as a thirty-two bit floating-point number, `f32`).)

```
fn maximum_likelihood_estimate(data: &[Bit], degree: usize) -> HashMap<(Vec<Bit>,
Bit), f32> {
    let mut theory = HashMap::with_capacity(2usize.pow(degree as u32));
    // Cartesian product—e.g., if degree 2, [00, 01, 10, 11]
    let patterns = bit_product(degree);
    for pattern in patterns {
        let mut zero_continuations = 0;
        let mut one_continuations = 0;
        for window in data.windows(degree + 1) {
            let (prefix, tail) = window.split_at(degree);
            let next = tail[0];
            if prefix == pattern {
                match next {
                    ZERO => {
                        zero_continuations += 1;
                    }
                    ONE => {
                        one_continuations += 1;
                    }
                }
            }
        }
        let continuations = zero_continuations + one_continuations;
        theory.insert(
            (pattern.clone(), ZERO),
            zero_continuations as f32 / continuations as f32,
        );
        theory.insert(
            (pattern.clone(), ONE),
            one_continuations as f32 / continuations as f32,
        );
    }
    theory
}
```

Now that you have the best theory for each particular *n*, you can compare how well each of them predict the data! For example, according to *n* = 0 coin model with maximum-likelihood parameter *p* = 0.47, the probability of your 500-bit sample is about ...
0.0000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000
000000000000007517883433770135.

Uh. The tiny probability makes sense: there's a *lot* of randomness in 500 flips of a biased coin. Even if you know the bias, the probability of any *particular* 500-flip sequence is going to be tiny. But a number that tiny is kind of unwieldy to work with. You'd almost rather just count the zeros and ignore the specific digits afterwards.

But counting the zeros is just taking the logarithm—well, the negative logarithm in the case of zeros after the decimal point. Better make the log base-two—it's *thematic*. Call this measurement the *log loss*.

```
fn log_loss(theory: &HashMap<(Vec<Bit>, Bit), f32>, data: &[Bit]) -> f32 {
    let mut total = 0.;
    let degree = log2(theory.keys().count()) - 1;
    for window in data.windows(degree + 1) {
        let (prefix, tail) = window.split_at(degree);
        let next = tail[0];
        total += -theory
            .get(&(prefix.to_vec(), next))
            .expect("theory should have param value for prefix-and-continuation")
            .log2();
    }
    total
}
```

Now you can compare different theories to see which order of Markov chain is *the best* theory to "fit" your 500-bit sample ... right?

```
    for hypothesized_degree in 0..15 {
        let theory = maximum_likelihood_estimate(&data, hypothesized_degree);
        println!(
            "{}th-order theory: fit = {}",
            hypothesized_degree,
            log_loss(&theory, &data)
        );
    }
```

```
0th-order theory: fit = 498.69882
1th-order theory: fit = 483.86075
2th-order theory: fit = 459.01752
3th-order theory: fit = 438.90198
4th-order theory: fit = 435.9401
5th-order theory: fit = 425.77222
6th-order theory: fit = 404.2693
7th-order theory: fit = 344.68494
8th-order theory: fit = 270.51175
9th-order theory: fit = 199.88765
10th-order theory: fit = 147.10117
11th-order theory: fit = 107.72962
12th-order theory: fit = 79.99724
13th-order theory: fit = 57.16126
14th-order theory: fit = 33.409912
```

There's a problem. Higher choices of *n* monotonically achieve a better "fit". You got the idea of higher-order Markov chains because the idea of a biased coin didn't seem adequate to explain the consecutive and alternating runs you saw, but you somehow have trouble believing that the bitstream was generated by a *fifteenth*-order Markov chain with a completely separate probability for the next bit for *each* of the 2^{15} = 32,768 prefixes `000000000000000`, `000000000000001`, `000000000000010`, *&c.* Having had the "higher-order Markov chain" idea, are you now obligated to set *n* as large as possible? What would that even *mean*?

In retrospect, the problem should have been obvious from the start. Using your sample data to choose maximum-likelihood parameters, and then using the model with those parameters to "predict" the *same* data puts you in the position of the vaunted "sharpshooter" who paints a target around a clump of bullet holes *after* firing wildly at the broad side of a barn.

Higher values of *n* are like a ... thinner paintbrush?—or a squigglier, more "gerrymandered" painting of a target. Higher-order Markov chains are *strictly* more expressive than lower-order ones: the zeroth-order coin is just a first-order Markov chain where the next-bit-after-`0` and next-bit-after-`1` parameters just happen to be the same; the first-order Markov chain is just a second-order chain where the next-bit-after-`00` and next-bit-after-`10` parameters happen to be the same, as well as the next-bit-after-`01` and—*enough!* You get it!

The broadcast is ongoing; you're not limited to the particular 500-bit sample you've been playing with. If the worry were *just* that the higher-order models will (somehow, you intuit) fail to predict future data, you could use different samples for estimating parameters and validating the resulting models, but you think you're suffering from some more fundamental confusion—one that's probably not limited to Markov chains in particular.

Your working concept of what it means for a theory to "fit" the data, is for it to maximize the probability with which the theory predicts the data. This is an objective, *quantitative* measurement. (Okay, the log loss is taking the negative logarithm of that to avoid so many zeros after the decimal point, but minimizing the log loss and maximizing the probability are both expressing the same preference on theories.)

How do you know (and your gut says that you *know*) that the higher-order models will do badly on future data, if your objective criterion of model-goodness says they're better? The log loss always "wants" to you to choose ever-more-complex models. You asked: what would that even *mean*? But maybe it doesn't have to be a rhetorical question: what *would* that even mean?

Well ... in the limit, you could choose a theory that assigns Probability One to the observed data. The "too many zeros"/"avoid working with really tiny numbers" justification for taking the negative log doesn't really apply here, but for consistency with your earlier results, you dutifully note that the logarithm of 1 is 0 ...

But maybe "too many zeros" isn't the only good motivation for taking the logarithm? *Intelligence is prediction is compression.* The log loss of a model against the data can be interpreted as the expected number of bits you would need to describe the data, given the optimal code implied by your model.

In order to communicate a reduction in your uncertainty, you need to send a signal—something you can choose to vary in response to the reality of the data. A signal you

can vary to take two possible states, can distinguish between two sets among which you've divided the remaining possibilities; writing down a bit means halving your uncertainty.

On this interpretation, what the logarithm of Probability One being zero *means* is that if your theory predicted the exact outcome with certainty, then once you stated the theory, you wouldn't have to say anything more in order to describe the data—you would just *know* with zero further bits.

Once you stated the theory. A theory implies an optimally efficient coding by which further bits can whittle down the space of possibilities to the data that actually happened. More complicated or unlikely data requires more bits just to *specify*—to single out that one outcome amongst the vastness of equally remote alternatives. But *the same thing goes for theories*.

Given a particular precision to which parameters are specified, there are exponentially more Markov chains of higher degrees, which can continue to drive down the log loss—but not faster than their *own* probability decreases. You need exponentially more data just to learn the parameters of a higher-order model. If you don't *have* that much data —enough to pin down the 2^n parameters that single out this *particular* higher-order Markov chain amongst the vastness of equally remote alternatives—then your maximum-likelihood best guess is not going to be very good on future data, for the same reason you can't expect to correctly guess that a biased coin has a probability of landing Heads of exactly 0.23 if you've only seen it flipped twice.

If you *do* have enough data to learn a more complex model, but the data was actually generated by a simpler model, then the parameters of the complex model will approximately take the settings that produce the same behavior as the simpler model —like a second-order Markov chain for which the bit-following-`01` parameter happens to take the same value as the bit-following-`11` parameter. And if you're deciding what theory to prefer based on both fit and complexity, the more complex model won't be able to "pay" for its increased complexity with its own predictions.

Now that you know what's going on, you can modify your code to penalize more complex models. Since the parameters in your implementation are 32-bit floats, you assign a complexity cost of $32 \cdot 2^n$ bits to n-th order Markov chains, and look at the sum of fit (log loss) and complexity. Trying out your code again on a larger sample of 10,000 bits from the broadcast—

```
    for hypothesized_degree in 0..10 {
        let theory = maximum_likelihood_estimate(&data, hypothesized_degree);
        let fit = log_loss(&theory, &data);
        let complexity = 2f32.powi(hypothesized_degree as i32) * 32.;
        println!(
            "{}th-order theory: fit = {}, complexity = {}, total = {}",
            hypothesized_degree, fit, complexity, fit + complexity
        );
    }
```

```
0th-order theory: fit = 9970.838, complexity = 32, total = 10002.838
1th-order theory: fit = 9677.269, complexity = 64, total = 9741.269
2th-order theory: fit = 9111.029, complexity = 128, total = 9239.029
3th-order theory: fit = 8646.953, complexity = 256, total = 8902.953
4th-order theory: fit = 8638.786, complexity = 512, total = 9150.786
5th-order theory: fit = 8627.224, complexity = 1024, total = 9651.224
```

```
6th-order theory: fit = 8610.54, complexity = 2048, total = 10658.54
7th-order theory: fit = 8562.568, complexity = 4096, total = 12658.568
8th-order theory: fit = 8470.953, complexity = 8192, total = 16662.953
9th-order theory: fit = 8262.546, complexity = 16384, total = 24646.547
```

—reveals a clear preference for the third-order theory (that for which the fit-plus-complexity score is the lowest), allowing you to enjoy the huge 450-plus–bit leap in compression/prediction from *n* := 2 to 3 and *logically stop there*, the steepness of the ascent into the madness of arbitrary complexity successfully dissuading you from chasing after diminishing returns (which you suspect are only hallucinatory). That's the power packed by parsimony—the sublime simplicity of *Science*.

(Full source code.)

The bads of ads

In London at the start of the year, perhaps there was more advertising than there usually is in my life, because I found its presence disgusting and upsetting. Could I not use public transport without having my mind intruded upon continually by trite performative questions?

Sometimes I fantasize about a future where stealing someone's attention to suggest for the fourteenth time that they watch your awful-looking play is rightly looked upon as akin to picking their pocket.

Stepping back, advertising is widely found to be a distasteful activity. But I think it is helpful to distinguish the different unpleasant flavors potentially involved (and often not involved—there is good advertising):

1. Mind manipulation: Advertising is famous for uncooperatively manipulating people's beliefs and values in whatever way makes them more likely to pay money somehow. For instance, deceptively encouraging the belief that everyone uses a certain product, or trying to spark unwanted wants.

2. Zero-sumness: To the extent advertising is aimed at raising the name recognition and thus market share of one product over its similar rivals, it is zero or negative sum: burning effort on both sides and the attention of the customer for no overall value.


NESCAFÉ

3. Theft of a precious thing: Attention is arguably one of the best things you have, and its protection arguably worthy of great effort. In cases where it is vulnerable—for instance because you are outside and so do not personally control everything you might look at or

hear—advertising is the shameless snatching of it. This might be naively done, in the same way that a person may naively steal silverware assuming that it is theirs to take because nothing is stopping them.

4. Cultural poison: Culture and the common consciousness are an organic dance of the multitude of voices and experiences in society. In the name of advertising, huge amounts of effort and money flow into amplifying fake voices, designed to warp perceptions–and therefore the shared world–to ready them for exploitation. Advertising can be a large fraction of the voices a person hears. It can draw social creatures into its thin world. And in this way, it goes beyond manipulating the minds of those who listen to it. Through those minds it can warp the whole shared world, even for those who don't listen firsthand. Advertising shifts your conception of what you can do, and what other people are doing, and what you should pay attention to. It presents role models, designed entirely for someone else's profit. It saturates the central gathering places with inanity, as long as that might sell something.

5. Market failure: Ideally, whoever my attention is worth most to would get it, regardless of whether it was initially stolen. For instance, if I have better uses for my attention than advertising, hopefully I will pay more to have it back than the advertiser expects to make by advertising to me. So we will be able to make a trade, and I'll get my attention back. In practice this is probably too complicated, since so many tiny transactions are needed. E.g. the best message for me to see, if I have to see a message, when sitting on a train, is probably something fairly different from what I do see. It is also probably worth me paying a small sum to each person who would advertise at me to just see a blank wall instead. But it is hard for them to collect that money from each person. And in cases where the advertiser was just a random attention thief and didn't have some special right to my attention, if I were to pay one to leave me alone, another one might immediately replace them.[1]

6. Ugliness: At the object level, advertising is often clearly detracting from the beauty of a place.

These aren't necessarily distinct—to the extent ugliness is bad, say, one might expect that it is related to some market failure. But they are different reasons for disliking a thing-a person can hate something ugly while having no strong view on the perfection of ideal markets.

What would good and ethical advertising look like? Maybe I decide that I want to be advertised to now, and go to my preferred advertising venue. I see a series of beautiful messages about things that are actively helpful for me to know. I can downvote ads if I don't like the picture of the world that they are feeding into my brain, or the apparent uncooperativeness of their message. I leave advertising time feeling inspired and happy.

The Darwin Game - Rounds 0 to 10

The most important change between my game and Zvi's original is that bots can read each others' source code. They can simulate each other and predict each others' behavior. Within a day of the tournament launching—and an entire week before entries closed—Zack_M_Davis had already written a bot to simulate opponents and then open-sourced it to everyone.

That's what happens when a significant contributor to an open source Lisp dialect participates in a software competition.

Taleuntum wanted to write an even better simulator but was informed that it would take too many years to run.

Muticore solved the limited compute problem and wrote a safe, effective, obfuscated simulator, with randomization.

The Phantom Menace

Three separate people asked me what happens if a bot crashes the game while simulating an opponent with malware in it. This turned out not to matter because nobody deployed malware to destroy simulators. Only one player, Measure, deployed malware—and the malware didn't crash the game. Instead, it attempted to replace its opponent's `move` method with a method that returned `0` instead. But the threat of getting disqualified seemed to scare away other potential simulators.

Taleuntum did write a `MatrixCrashingBot` that crashes simulators but did not submit it. This is a disappointment, as Taleuntum would have been allowed to submit this bot as a separate entry on the grounds that it does not coordinate with Taleuntum's CloneBot. To my knowledge, nobody else took advantage of this deliberate loophole in the rules either.

RaterBot safely combed through its opponent's source code for "2"s and "3"s to estimate aggression without the dangers associated with running untrusted code.

Computer programs attempting to simulate each other can produce complex behavior. The behavior is so complex it is provably undecidable—and that's totally ignoring the real-world sandboxing problem.

Nevertheless, two contestants requested I write code to simulate their opponents. I refused these requests. Zvi[1] accepted a simpler bot and the other contestant dropped out.

> I'm surprised running the enemy is complicated though—it should just be a function call.
>
> —quote from the contestant who dropped out

The most significant use of an opponent's source code came from Vanilla_cabs.

Attack of the Clones

Zvi's original game was dominated by a clique of players who coordinated out-of-game to defeat the non-clique players. It worked great—and then defectors within the clique dominated both the non-clique players *and* the clique players.

Vanilla_cabs observed that players could use each other's source code to detect (and therefore punish) defection from a clique. Leveraging this insight, Vanilla_cabs created a

CloneBot template that would cooperate only for the first 90 rounds—long enough to eliminate all non-clones—and then allow players to do whatever they wanted afterward.

Brilliantly, the CloneBots used the lexographic length of custom code to break parity symmetry and guarantee perfect coordination between CloneBot instances instead of having to fumble around for the initial rounds of each pairing.

At the same time, Vanilla_cabs worried that someone might attempt to bypass the CloneBot's self-recognition system. As a first line of defense, Vanilla_cabs kept the source code secret from non-cloners. As a second line of defense, Vanilla_cabs deliberately left open a weakness as a honeypot. Among the first co-conspirators in the clone army, only Lanrian reported this weakness.

Taleuntum, in contrast, contributed two lies to the clique intrigue.

> I lied that I've already submitted one program detecting and crashing simulators... obviously I did not even plan to do so. Some time later...I added another lie that the method of detecting simulators was my friend's idea (hopefully suggesting that there is another contestant with the same method outside the clique).
>
> —Taleuntum

One player discovered a way to defect from the CloneBots.

Multicore

The original CloneBot template used `splitlines()` to parse source code. But the Python3 interpreter behaves slightly differently from `splitlines()`. In particular, `splitlines()` splits on the groups separator character `0x1D`, which Python3 does not treat as a newline. By placing `0x1D` group separator characters inside of a comment, Multicore's EarlyBirdMimicBot bypassed the CloneBot self-recognition system.

Multicore did a lot of things.

- Wrote a simulator.
- Optimized the simulator via a branching algorithm to perform efficient searches.
- Used randomization to disrupt opposing simulators.
- Protected the simulator from malware.
- Joined the CloneBots.
- Defected from the CloneBots via a zero-day exploit in the CloneBot code.
- Recruited two friends to submit password bots to serve as fodder.
- Adjusted strategy based on the round number.

When I hosted this tournament, I hadn't expected anyone to "[read] through the C code for the python lexer".

For a complete write-up of Multicore's strategy, including source code, see here.

> On a side note, I really love this site. I can't really recall any other game I've been in getting this tangled.
>
> —Emiya

The First Game

The first iteration of the game was run by Taleuntum who ran "a simulation of the whole tournament till the 160th round with 8 bots" despite the tournament's source code not being public at the time.

Taleuntum's tournament was unofficial and does not count.

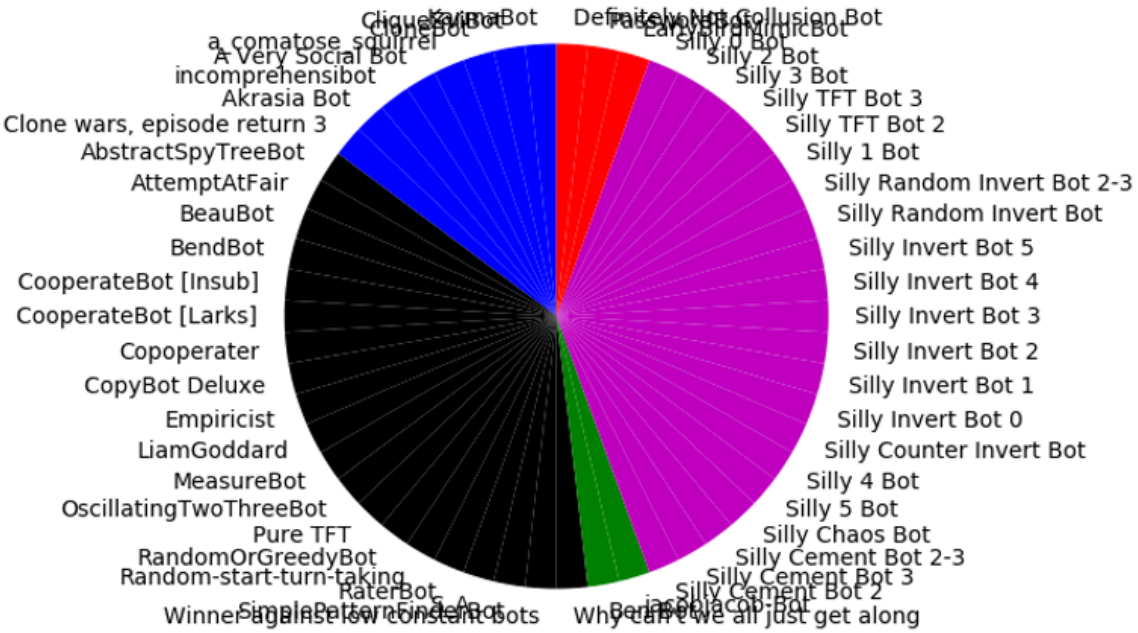
The Real Game

Teams

In order to make sense of the 54 participating bots, I have bucketed them into teams.

- [Blue] **Clone Army.** 10 players pledged to submit clone bots. 8 followed through, 1 didn't and Multicore submitted a [Red] mimic bot.
- [Red] **Multics.** Multicore's friends submitted 2 password bots to aid Multicore's mimic bot.
- [Green] **Norm Enforcers.** Ben Pace joined forces with jacobjacob to form their own little duo.
- [Black] **Chaos Army.** 20 players wrote individual bots.
- [Magenta] **NPCs.** I wrote 21 Silly Bots. Some of them had synergies.

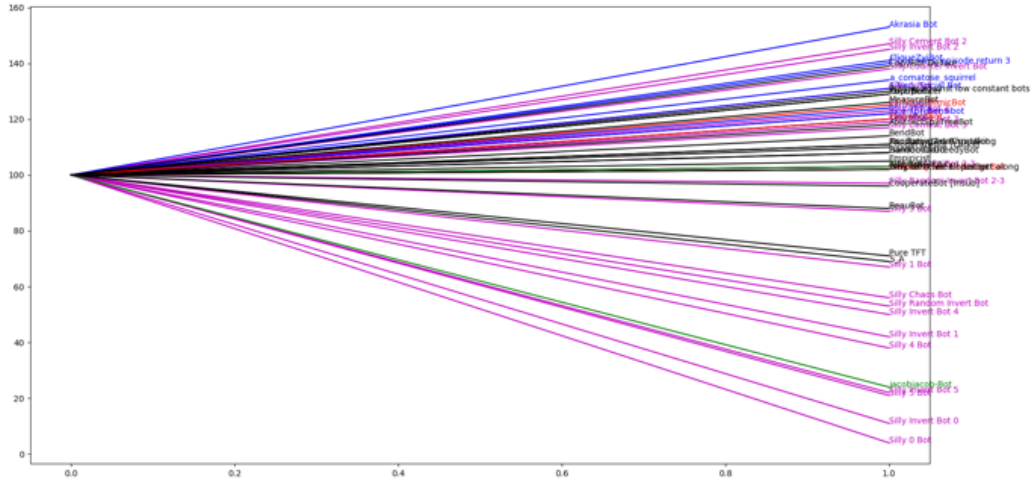
The clones [Blue] begin the game outnumbered 6-to-1.



Edit: Everything below this line is in error. See here for details.

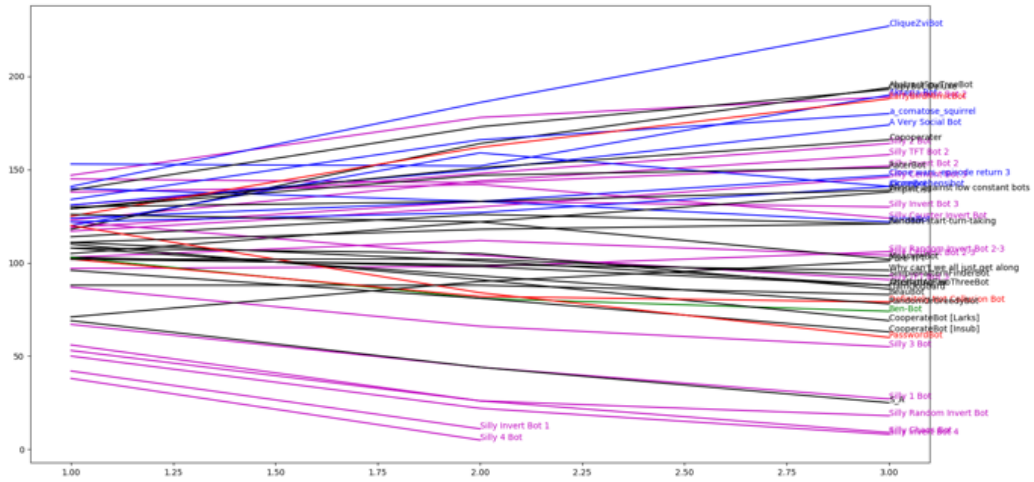
Round 1

5 bots died on turn 1 including 4 NPCs and Team Norm Enforcers' jacobjacob bot.



Rounds 2-3

Another 4 NPCs died.

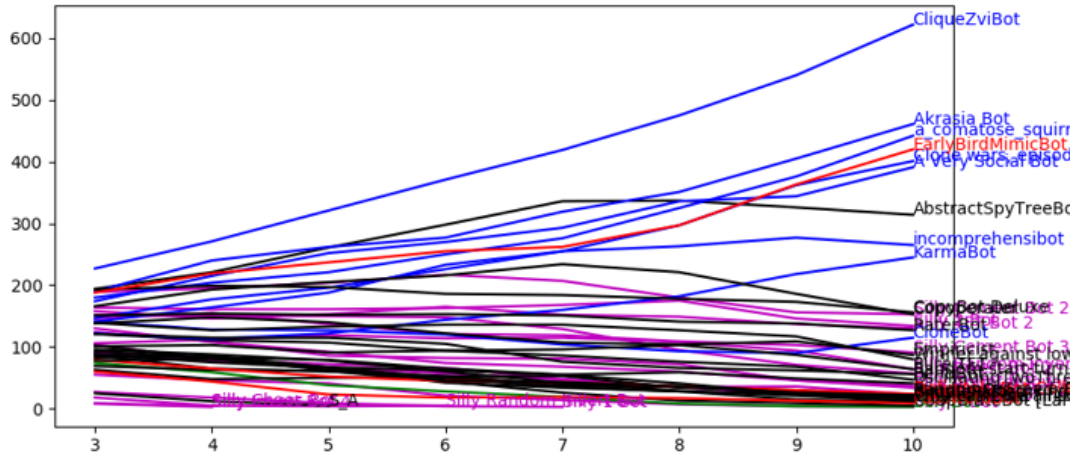


Rounds 4-10

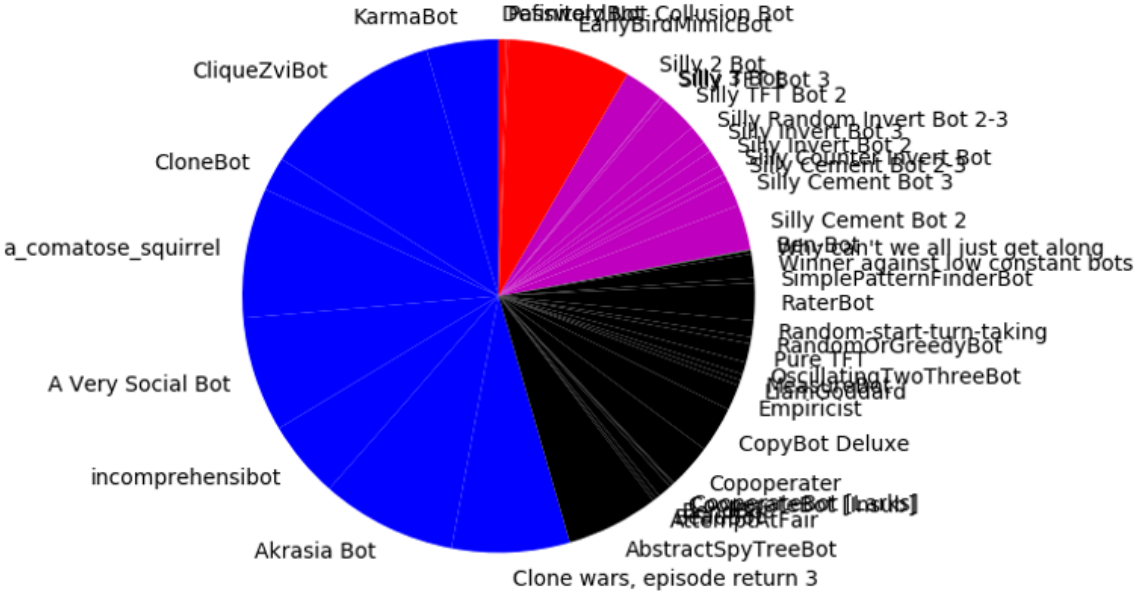
S_A and BenBot die, along with 3 more NPCs. Thus ends team Norm Enforcers.

The clone army is mostly doing well, except for CloneBot which is doing poorly and AbstractSpyTreeBot which is doing almost as well as the average clone.

EarlyBirdMimicBot is doing better than the average CloneBot but not by much. The MimicBot's `0x1D` exploit succeeded in defecting but the bot appears not to have leveraged its defection to gamebreaking effect.



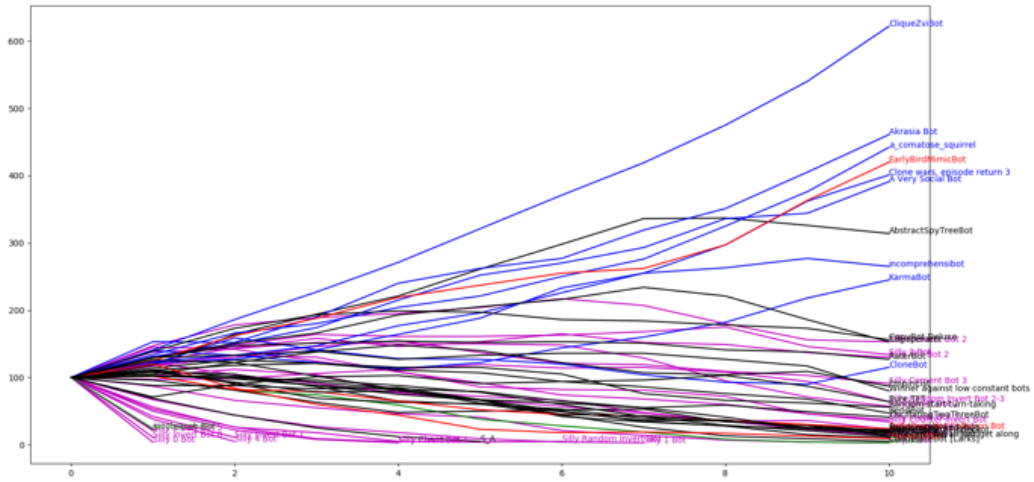
The clones have built up a critical mass of >50%. If their coordination mechanisms work then they ought to crush the rest of the competition.



If Zack_M_Davis' AbstractSpyTreeBot can survive in a world of clones until turn 90 when the clone treaty expires then there may be some hope for Chaos Army.

If not then, begun, the clone wars have.

Everything so far



Today's Obituary

Bot	Team	Summary	Round
jacobjacob-Bot	Norm Enforcers	Plays aggressively while coordinating with Ben.	1
Silly 5 Bot	NPCs	Always returns `5`.	1
Silly 0 Bot	NPCs	Always returns `0`.	1
Silly Invert Bot 0	NPCs	Starts with `0`. Then always returns `5 - opponent_previous_move`.	1
Silly Invert Bot 5	NPCs	Starts with `5`. Then always returns `5 - opponent_previous_move`.	1
Silly 4 Bot	NPCs	Always returns `4`. Then always returns `5 - opponent_previous_move`.	2
Silly Invert Bot 1	NPCs	Starts with `0`. Then always returns `5 - opponent_previous_move`.	2
Silly Chaos Bot	NPCs	Plays completely randomly.	4
Silly Invert Bot 4	NPCs	Starts with `4`. Then always returns `5 - opponent_previous_move`.	4
S_A	Chaos Army	Plays `1` 79% of the time, `5` 20% of the time and randomly 1% of the time	5
Silly Random Invert Bot 4	NPCs	Starts randomly. Then always returns `5 - opponent_previous_move`.	6
Silly 1 Bot	NPCs	Always returns `1`.	7
Ben Bot	Norm Enforcers	Cooperates with jacobjacob [deceased]. If not paired with jacobjacob then this bot returns `3` for the first 100 turns and then does fancy stuff. Unfortunately for Ben, I picked 100 as the number of turns per pairing.	10
Silly 3 Bot	NPCs	Always returns `3`.	10

The next installment of this series will be posted on October 26, 2020 at 5 pm Pacific Time.

Zvi's specification did address the halting problem, sandboxing problems and unpredictable resource consumption. ↩

Philosophy of Therapy

This is a linkpost for http://daystareld.com/philosophy-of-therapy/

For a lot of people, therapy can be a confusing, mysterious thing of questionable value. Many have tried it when they were younger, and felt that at best it was only of minimal help, while for others it actually made things worse. In many cultures, therapy looks very different from how it's practiced in the "western world," and the concept of mental health itself is often treated with suspicion or dismissal. I've known many people who, even while not being skeptical, were still confused about what the purpose of therapy actually is, or what situations warrant seeking a therapist out.

In my practice as a therapist, I often reorient myself to the basic core of therapy, which to me is about **helping people get unstuck**. Sometimes the thing you're stuck on is a recurring and disruptive emotional state, other times it's some harmful interpersonal dynamic, and other times it's a pattern of behavior. Whatever the specifics, there is some aspect of the client's life that is **not going the way they would prefer,** and the therapist's job is to help them find a way to change that.

What the therapy provides also varies; good therapy can create space for honest expressions of emotion, provide new perspectives or insight, and offer new "tools" for the client to use in their lives, specific behaviors or mental motions that help move past the stuckness.

Those skeptical of therapy often wonder: can't people just talk to their friends or family if they need emotional support? Aren't there self-help books they can try? And of course they can, and should try those things! For many people, the majority of their difficulties do not require a therapist.

Which means therapy is for what's left. Those things that seem truly intractable, the things that you feel **stuck** on, which other resources have failed to help resolve.

But I'd like to demystify therapy further, and better yet, I think by better understanding **what therapy is meant to do and how**, people can get some of the value that therapy can provide even without going to see a therapist.

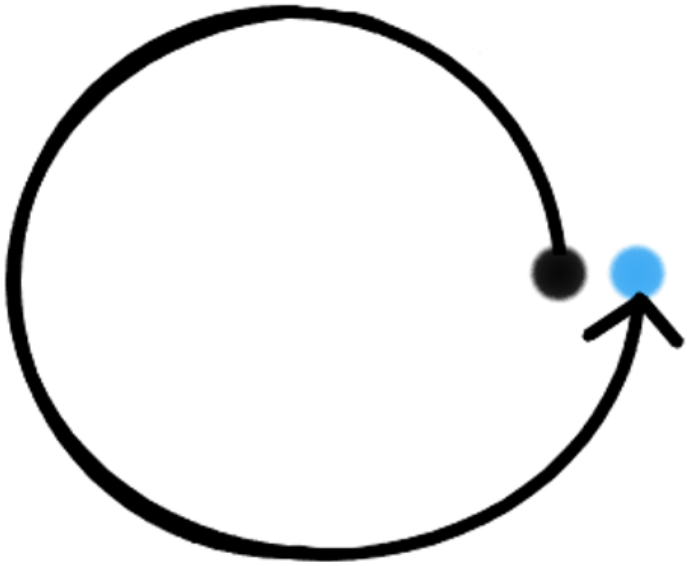
Because while much of the change in therapy comes from the therapeutic relationship itself (which is why first finding a therapist that you feel comfortable with is half the battle), for a large portion of clients I've seen, even just changing the frame of the problems they experience, or changing the way they *view themselves* in relation to their problems, actually makes the problem less sticky. A new frame can reveal more levers to pull and knobs to turn, or new vistas of the mind to explore and inhabit, that can help make the problem more manageable.

So that's the goal of this essay. By teaching the history of the different philosophies of therapy, I want to teach you how, **if you change the frame, you can change the problem**.

I. History

Ask people to describe what therapy involves or "looks like," and most who haven't been in therapy will say something like "one person lies down on a couch and talks to the therapist, who takes notes and asks questions like 'How does that make you feel' and 'Tell me about your childhood' and 'How do you feel about your mother?'"

This is largely the result of Hollywood Therapy, but it's rooted in the origins of therapy, which is Freudian—what's now called Psychoanalytic Therapy.



Sigmund Freud was the progenitor of applied psychology; the idea that we could study the way people think and feel and act, and use it to directly help them “improve” in some way. He was inspired by his mentor, a physician who helped alleviate a patient’s untreatable illness by just asking questions about her symptoms. That patient coined the term “talking cure,” and Freud took this concept and ran with it, dedicating his life to the idea that many ills people suffer are psychological in nature rather than physiological, and that just talking about them can help reduce or remove them.

Freud had a lot of ideas of his own, however, and while many them turned out to be nonsense, he also had some that turned out to be true, or at the very least, useful, such as the concept of a “subconscious,” or the idea of dividing a person’s mind into subagents (in his case, Id, Ego, and Superego). As the arrow above indicates, Freud cared almost exclusively about the past; he believed that by studying one’s childhood, the way they were raised, their early environment, or the origin of a certain dysfunctional behavior, you could identify all sorts of traumas or stresses that cause dysfunction later in life. Once identified, he believed the client would gain a feeling of “catharsis” that would start the path to healing.

Here’s where I admit that I have something of a bias against psychoanalysis.

In my view, Freud was a philosopher first and foremost, rather than a scientist. He had interesting ideas that seemed logical to him, and a scientific frame of mind, but while he pursued the application of these ideas with an admirable gusto, his documentation did not seem to aim its rigor at testing which of his ideas were *true*. I'm unaware of any hypotheses Freud generated that he then went on to falsify. (If you know of any, please do share them!)

Far from an attempt to bash the man, I do admire him a great deal. It’s hard to be the first person to basically invent an entire field of science and do it all perfectly such that you are simultaneously the person observing reality, coming up with ideas, *and* dispassionately testing those ideas, all while trying to do work as a clinician. But I believe most modern schools of therapy have picked out the gems of his work and left the rest to history lessons.

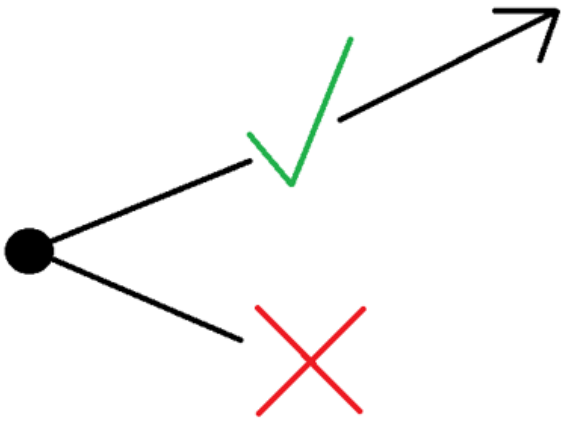
That isn't to say this branch of therapy is all worthless. While catharsis alone generally doesn't solve most people's symptoms (psychosomatic illnesses like his mentor's patient's are in fact very rare), delving into one’s past can lead to insights into their current problems, and many do report feeling better about their problems when they have a chance to talk about them (again, credit to Freud, this would likely have been very encouraging to him when he began his work).

Additionally, as a colleague pointed out to me after reading an earlier version of this article, many modern psychoanalysts do seek to empirically test the field's ideas in order to continue to develop evidence-based treatments, and modalities such as Transference-Focused Psychotherapy have evidence suggesting it to be at least as effective as other standards of treatment.

(A *modality* is a method of therapy that has a specific structure to help a client reach wellness. More than a specific intervention, modalities often include multiple interventions, as well as a particular type of relationship between client and therapist that dictates whether the therapist acts as more of a guide, partner, or authority. Each modality operates on a particular hypothesis of how therapy can help clients with certain problems.)

In any case, while psychoanalysis as practiced by Freud and his ideological descendants (Carl Jung, Anna Freud, Erik Erikson) focused so much on the client’s past, new discoveries in psychology led to therapeutic modalities that focused instead on influencing the client’s future.

Enter, the Behaviorists.



As Freud is to Psychoanalysis, so Ivan Pavlov, of dog fame, is to Behaviorism. Pavlov discovered and experimented with classical conditioning, the idea that you can pair different

stimuli to influence responses. This discovery was a great boon to pet owners, but also has direct applications to therapy. One example is addiction treatment, where for example the sight or smell of cigarettes or beer is paired with something that will evoke disgust. It also led to desensitization therapy for phobias, where pairing progressively more frightening stimuli with techniques and context that help relax the client can alleviate the fear response.

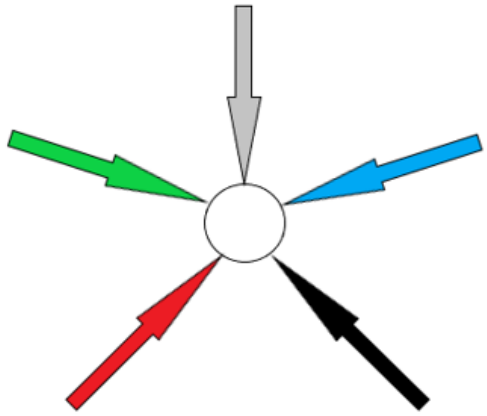
These ideas were expanded by Edward Thorndike and B.F. Skinner, whose work is called operant (or instrumental) conditioning. Rather than just pairing stimuli together to affect responses, their experiments showed demonstrable effects on learning and behavior through reinforcement and punishment; in therapy the idea of using positive reinforcement to incentivize desired behavior is often helpful for children, particularly those with developmental issues.

I don't have much to say about Behaviorism. For some things that people come to therapy for help with, it just works. For others... not so much. I think understanding the mechanisms of Behavioral Therapy is valuable for any clinician, but there's some obvious flaws with taking it as the only avenue toward better mental health.

Unlike psychoanalysts, a straw-Behaviorist doesn’t care about your past, and talking about your traumas or “deeper issues” would often be considered a waste of time. Instead the focus is on your symptoms. No symptom, no problem, right? Just apply the right type of reinforcement to increase positive behaviors and the right type of punishment to decrease negative behaviors, and all’s well…

…for some people, at least. Behaviorists had a lot of success in some domains, particularly when the “why” of the problem didn’t actually matter to the client or issue, but obviously struggled with others. After the first World War, clinicians formally recognized PTSD, or "shell shock," for the first time. Unfortunately, attempts to treat soldiers through psychoanalytic and behavioral therapy often failed, and so many psychologists turned clinician to help figure out how better understanding the *present* feelings we have, and how they impact our behavior, can lead to mental health.

Which brings us to Existential Therapy.



Rather than having a single founder, the Existential philosophy of therapy was converged upon by a wide range of psychologists and clinicians, many inspired by the writings of Kierkegaard, Nietzsche, Husserl, Scheler, Heidegger, and Sartre. These writers' attempts to redefine our understanding of not just what it means to be human, but an "actualized" human, a healthy, thriving, happy human, were believed to have great value in clinical efforts to help those in need.

But among that foundational pantheon, the *first* of the Existential therapists was Otto Rank, a student of Freud who later split with him over Freud's beliefs that a person's "formative years" are what determine who they become. Instead, Rank believed that human development continues throughout our lives, requiring continual negotiation and renegotiation between dual yearnings for individuation and connection.

For such heresy he was excommunicated by the psychoanalytic world, but he nevertheless influenced his own "family" of psychologists, including Rollo May, Viktor Frankl, who's more well known as the author of Man's Search for Meaning, and Abraham Maslow, of hierarchy fame. These psychologists focused not so much on what happened in someone's past or how to influence their future, but on their *now.* What do people feel like they need, that they lack? How does the client experience "need" at all? What relationship do they have with their hurts and wants, and what would be necessary for them to feel fulfilled? How do those different needs and wants conflict with each other, and how can they be better brought into harmony?

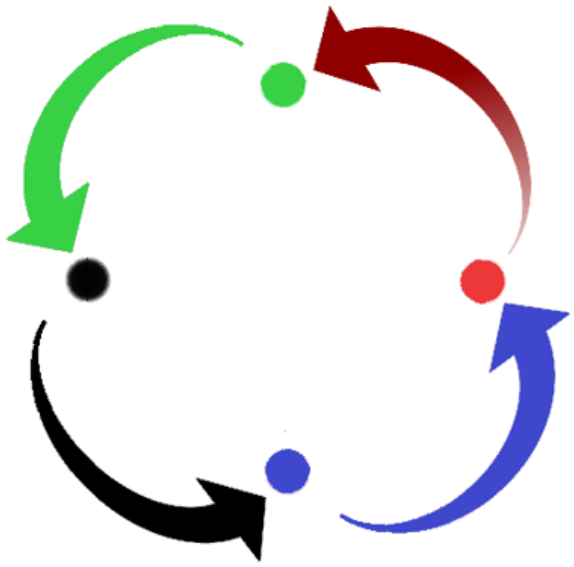
Existential therapy also marked a new dynamic between client and therapist; rather than a top-down hierarchy, where the clinician is the "expert" and the client the "patient," what became known as client-centered therapeutic practice began to form. It placed both therapist and client as equals; the clinician has the education and skills, but the client is the expert on their own lives, of what they think and feel, and so the Existential therapist's role is more that of a facilitator to the client's growth.

This may seem like polite semantics, but most people who've been to both kinds of therapists can tell how big a difference it makes if, upon disagreeing with their therapist on something, they're treated not like a stubborn mule who is "resistant" to change, but rather a person with agency, whose motivation to improve is taken for granted by their therapist. The philosophy also emphasizes the importance of a therapist who is willing to listen, encourage, and support the client's *personal journey* to better mental and emotional health, as *the client* defines those things.

Under the light of Existential Therapy (and its more upbeat twin, Humanistic Therapy) there grew many techniques to help clients better understand themselves, including Carl Rogers's "reflective listening," which has become a staple of good therapy from every philosophy, as well as techniques to better interface with our emotions, such as "focusing" by Eugene Gendlin, which I personally have found to be one of the most generically effective tools to teach practically every client I've had.

Time to admit to another bias, in case it's not clear; I'm a huge fan of existential/humanistic therapy. In my experience it has a wide "range" in what it can successfully treat, and its frame makes up an integral part of what makes modalities effective in general.

But it's not the form of therapy I was formally educated in, and it's not the latest form of therapy that was developed. There's one last dimension that even existential therapists failed to engage in, and if you're following the theme of the arrows you might have guessed it: the opposite of focusing on ourselves is focusing on everything else.



Enter Systemic Therapy (also known as Family Systems Therapy, or just Family Therapy), born in the 1950s from a very powerful need; the need for better marriage counseling.

In the post WWII era, if a husband and wife wanted to save their marriage, they would go about it thusly: the man would have his counselor, and the woman would have her counselor, and both would see their counselors separately. If they went to a fancy clinic dedicated to marriage counseling, the two clinicians would be coworkers, seeing their clients individually, then consulting on the case between sessions, or even mid-session before returning to their clients.

If that sounds crazy, just remember that this was the 50's, when people still thought smoking was good for you. The idea was that a client's relationship with their therapist was sacrosanct, and must always be preserved as a space of utter one-on-one privacy that would allow them to be completely frank, without worrying about their spouse's presence, or their therapist telling their spouse anything spoken of in confidence.

Eventually some therapists in California *realized* how absurd this was, not to mention ineffective. They suggested a new way to practice marriage counseling, where a single counselor (or even two) spoke with both clients together, in the same room and at the same time. That way a therapist could observe their interactions and mediate their discussions directly.

Their clinic said no.

So Don Jackson and his colleagues left to form the Mental Research Institute in Palo Alto, where they developed their own modality of therapy, one that involved not just the individual patient, but sometimes romantic partners, family members, even friends if the problem called for it.

They weren't the only ones; Salvador Minuchin, Murray Bowen, Ivan Boszormenyi-Nagy, Virginia Satir, and Jay Haley all developed modalities based on the idea that, to help a client overcome dysfunction, the therapist should focus not just on the client, but the system they're a part of, whether that be their family, their work environment, their culture, or even their country, all at various levels of abstraction.

(There isn’t going to be a test on all the names I’m throwing at you, but if I went into every single modality we’d be here all day, and this way you have an easy way to look into them on your own if you want.)

The study of cybernetics and communication theory were also prominent influences, particularly by the anthropologist Gregory Bateson, who believed that all forms of communication are adaptive, and rejected the concepts of linear and dualistic thinking for studying systems.

The “systems” being referred to in these therapies can be any context you’re a part of, individually or simultaneously: family system, school system, work, friend-group, even cultural and religious. According to Bateson, being part of any system leads to inherent and unavoidable communication between you and the other parts, implicit or explicit, which affects the other parts of the system and how they behave, which further affects how you behave, and so on. Additionally, there can be no divide between an interactive observer and participant of a system; by observing the system directly, the therapist becomes a part of it.

This understanding led to a philosophy that takes the humility of existential therapy even further, and improved clinicians' ability to map the impact of one part of a system on the others, such that many modalities do not even identify anyone in particular as “sick” or “healthy,” but rather views behavior patterns themselves as dynamic or stagnant, and focuses on how change can propagate through the system by nudging elements of it. By understanding how everyone’s actions and reactions affect each other’s behavior, the client and clinician have more surface area on the problem to try and find solutions, more levers to pull and handles to grip from.

A big reason why this lens can be so valuable is that when you start working with groups rather than individuals, you have to address the fact that often times, not everyone involved in therapy has the same desire to be there, let alone incentive or drive to change. Of course, that was true before couples or entire families were being invited into a therapy room at once, but now the therapists were actively working to address it rather than just assuring whoever cared enough to be in the room that the problem was other people, and not them.

Oh, also worth noting that therapy up to this point was still a LONG process, often expected to last years. Systemic Therapy made a push toward briefer, more effective interventions, creating modalities like Solution Focused Therapy, which combined Systemic and Behavioral principles to bring about real, lasting change within 4-6 months.

So, that’s the four cardinal philosophies I’ve sort-of-made-up as a labeling scheme to map all therapy onto. Now we get to the meat of the matter; how can just knowing about them actually help?

II. Case Study

“You have to help me,” Marge, 55, says during her first session. “It’s my husband. He’s become *obsessed* with model trains!”

Sidebar 1: An important thing to note is that the client said she needs help, but highlighted her husband as the focus of therapy. Some equivalent of “fix my spouse” (or “fix my kid”) is nearly as common, in my experience as “fix me,” and often times the spouses in question aren’t always in the room. So we work with what we have.

“I can see you're worried about him," I say. "What does ‘obsessed’ look like? Are you running out of money?”

“Well, no,” she admits. “We can afford it, but… every month he’ll order hundreds of dollars worth of new models and tracks, and after work he goes down to the basement. He spends *hours* down there, every day!”

I nod. “Yeah, it makes sense why that might be concerning. Is he skipping meals? Staying up all night?”

“No, no. He’s sleeping fine, he’s still eating… but it’s quick, you know, he’ll pop out of the basement for ten minutes, wolf down his food without looking at it, then go back to his trains for another six hours. That’s not *normal,* right?”

Sidebar 2: “ Normal,” along with “healthy,” is perhaps the most loaded word in therapy. Unless the client is insistent, or we’ve formed a strong therapeutic relationship, I try to avoid giving any kind of verdict on either, and instead use the therapist standby of answering a question with a question; in this case not ‘what is normal,’ but rather:

“What would you consider to be the ‘normal’ things he does do?”

“You mean like work?”

“Yeah, and beyond that. Is he still seeing his friends?”

“Yes, once in a while he’ll go out for some drinks with them.”

This is evidence that he’s not a shut-in. “Feel free to say it’s too personal for now, but just to check, does he still want sex?”

She blushes. “Not often, but, yes. Sometimes.”

“Okay. Does he talk about other things, or is it all trains all the time, now?”

“We barely talk at all, now, not like we used to.”

“What was the last conversation you had with him?”

“Oh, about the kids.”

“You have children?”

She smiles for the first time. “Yes, two. Both married, one with our first grandchild on the way.”

“Congratulations! And he’s still interested in them, and the grandchild?”

“Oh, yes. He put off our vacation so we’d be around the first few months.” Her smile is gone now. “Which normally I’d be in favor of too, but… there’s some sort of convention nearby around then that he’s still planning to go to.”

“A model train convention?” I guess.

“Yes, I’m telling you, he’s just…” She shakes her head, seemingly at a loss for words.

Sidebar 3: “Pathologizing” is the perception that any action or view that is unusual is automatically a sign of illness, despite no evident dysfunction or suffering. In decades past, previous versions of the Diagnostic and Statistics Manual labeled things like homosexuality a mental health illness due to a mentality that didn’t distinguish between “normal” and “healthy.” Newer versions of the DSM have eliminated most of those, and there’s a concerted effort among (good) psychologists and therapists to distinguish real pathology as something that causes direct suffering for the patient.

At this point, I might feel an urge to say “Okay, so… what exactly is the problem here? Just because your husband is spending hundreds of dollars and hours a month on model trains doesn’t mean he needs therapy. If it’s not affecting his sleep, or his appetite, or his work, or

his social life… maybe he just likes trains, and that's okay? It's far from the worst hobby, and if it makes him happy, just let him like trains!"

I wouldn't say this out loud, however, at least not in the first session, because even if I've become at least reasonably sure that the husband is okay, to say something like that would be dismissive of *her experiences* .

Regardless of what her husband is doing, *she* is clearly unhappy. And while she might think she can be the client but not the patient, the truth is, from a systemic lens, there is no distinction. The system she lives in, her marriage, is clearly dysfunctional *for her* in some way, as evidenced by how she's suffering enough to come to a therapist. Perhaps her husband is too, in a non-obvious way that will be revealed through further questioning, but for now the focus would best be shifted to her.

There are a number of lenses through which to focus, however, and each might approach the problem in such different ways that **they essentially become different problems .**

- A psychoanalytic therapist could delve into Marge's past. Was her father distant with her, perhaps obsessed with his work or a hobby of his own? Did she have older siblings that left her out of their play? Was a childhood friend killed by a train? (Probably not that last one.)
- A behaviorist could focus on the husband's actions and develop strategies to reinforce or punish the ones she likes/dislikes. This would be pretty manipulative if the husband isn't on-board, however, so instead the therapist might focus on ways to associate her husband's hobby with positive emotions and experiences of her own.
- An existentialist could help Marge delve into the emotional experiences she's having, what she feels when she thinks of her husband in the basement or buying new models, and what needs she has that aren't being met. The goal would be either to dissolve the problem entirely by reframing her expectations, or teaching her new tools to manage her mood and satisfy her emotional needs.
- A systemic therapist could help by examining the overlapping systems she's a part of; her marriage, her family, her social circles. Did she and her husband used to do more things together? What was their marriage like when the kids were still part of the household? How often does she spend time with her own friends or hobbies? Perhaps there are ways she could better communicate to her husband what her needs are so he can understand how she's hurting, or examine what behaviors of hers might be reinforcing her husband's without even realizing it.

While individual modalities might lack scientific backing, I believe the broader philosophies can each be suited to different types of problems. That still means that if a therapist only sees the world through one or two lenses, they might not be able to help their client as well as someone whose approach is the better fit.

Perhaps more importantly, each *client* can respond better to a different philosophy, even if they present with nearly identical problems. For some, just getting down to brass tacks and tackling the symptoms is their ideal, while for others, digging deep into their psyche is what they want and respond well to.

This is part of the reason why one of the major tenets of good therapy is "stay curious." The more the therapist starts assuming they know what to expect from a client based on their presenting problem, no matter how often they've seen it before, the more likely they are to jump to conclusions about treatment that end up being a poor fit.

III. Modalities

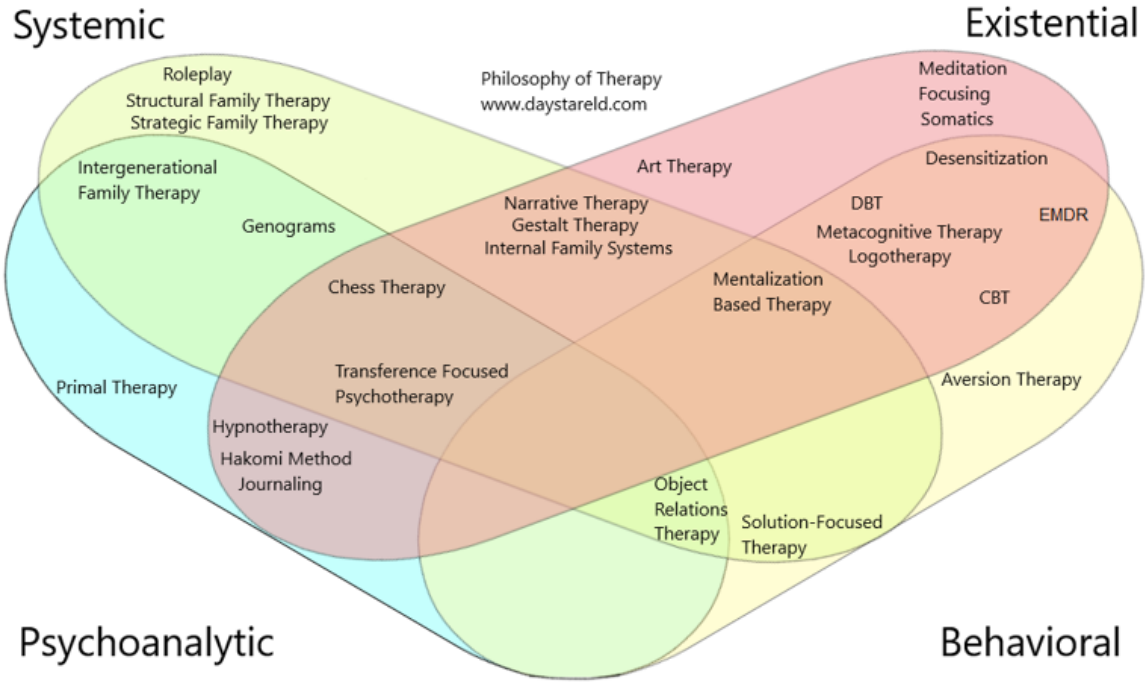
A therapy modality is more specific than a philosophy; it's not just a framework for what leads to dysfunction and how to correct it, but also a bundle of specific interventions and pathways, some more rigid than others, to lead the therapist and client from first session to last. Here's just a few examples that I use regularly:

Cognitive Behavior Therapy is a mix of Existential and Behavioral. It focuses on the looping interactions between our thoughts, feelings, and behavior, and how they reinforce each other such that altering one can alter others. (**Dialectic Behavioral Therapy** leans even more into the Existential side, with extra attention on mindfulness and mood regulation.)

Solution-Focused Therapy is a mix of Systemic and Behavioral. It helps the client identify their strengths and resources in their social systems, as well as how those systems reinforce their behaviors or symptoms, or can be altered to better reinforce more desired ones.

Narrative Therapy is a mix of Systemic and Existential therapy. It asks the client to present the narrative of their life, identify the ways the story they tell themselves and its framing is influenced by the broader systems they're a part of, then explores the way their narrative makes them feel while teaching techniques to better interface with those feelings.

And here's a handy-dandy diagram that lists just a few of the different modalities, techniques, and interventions used in therapy. There are many more that exist, and there may be different ways of practicing each of these that bump them from one section of the diagram to an adjacent one, but I believe every modality and strategy of therapy can ultimately be placed somewhere on this image, depending on how much they focus on understanding the client's past, interfacing with their thoughts and emotions, altering their behaviors, or adjustments to their environment/relationships.



(This is in no way a "complete" image, as there are dozens of different modalities and it would need to be massive to fit them all, but I figured it's better to just publish with some listed and update it over time.)

IV. Change the Frame, Change the Problem

I like collecting lenses through which to view the world. Each is like a different kind of mental map that I can use to navigate the territory of reality, and just like different types of maps (some simplistic and cartoonish, others realistic and highly detailed) can be more or less useful for different purposes, even maps that I know are not literally correct can still have value.

Overall this post is an ur-map, *my* ur-map, of different maps I've learned about in the field of therapy. I don't mean to present it as "the one true way to view therapy," but I've found it very helpful, and I hope others can too. It's also worth keeping in mind that it has many of the biases you'd expect from someone educated in an American college program that focused primarily on one particular philosophy.

Still, I think if more people were aware of the different lenses through which therapy can operate, they would better be able to navigate the sorts of problems that might lead them to a therapy office, maybe even help them find their way without going to one.

Next time you feel stuck in a particular way of thinking about your problems, a particular frame through which your problem seems insurmountable, try changing it. You might find it a lot more tractable than it seemed before.

The Alignment Problem: Machine Learning and Human Values

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://www.amazon.com/Alignment-Problem-Machine-Learning-Values/dp/153669519X

The Alignment Problem: Machine Learning and Human Values, by Brian Christian, was just released. This is an extended summary + opinion, a version without the quotes from the book will go out in the next Alignment Newsletter.

Summary:

This book starts off with an explanation of machine learning and problems that we can currently see with it, including detailed stories and analysis of:

- The gorilla misclassification incident

- The faulty reward in CoastRunners

- The gender bias in language models

- The failure of facial recognition models on minorities

- The COMPAS controversy (leading up to impossibility results in fairness)

- The neural net that thought asthma reduced the risk of pneumonia

It then moves on to agency and reinforcement learning, covering from a more historical and academic perspective how we have arrived at such ideas as temporal difference learning, reward shaping, curriculum design, and curiosity, across the fields of machine learning, behavioral psychology, and neuroscience. While the connections aren't always explicit, a knowledgeable reader can connect the academic examples given in these chapters to the ideas of specification gaming and mesa optimization that we talk about frequently in this newsletter. Chapter 5 especially highlights that agent design is not just a matter of specifying a reward: often, rewards will do ~nothing, and the main requirement to get a competent agent is to provide good *shaping rewards* or a good *curriculum*. Just as in the previous part, Brian traces the intellectual history of these ideas, providing detailed stories of (for example):

- BF Skinner's experiments in training pigeons

- The invention of the perceptron

- The success of TD-Gammon, and later AlphaGo Zero

The final part, titled "Normativity", delves much more deeply into the alignment problem. While the previous two parts are partially organized around AI capabilities -- how to get AI systems that optimize for *their* objectives -- this last one tackles head on the problem that we want AI systems that optimize for *our* (often-unknown) objectives, covering such topics as imitation learning, inverse reinforcement learning,

learning from preferences, iterated amplification, impact regularization, calibrated uncertainty estimates, and moral uncertainty.

Opinion:

I really enjoyed this book, primarily because of the tracing of the intellectual history of various ideas. While I knew of most of these ideas, and often also who initially came up with the ideas, it's much more engaging to read the detailed stories of _how_ that person came to develop the idea; Brian's book delivers this again and again, functioning like a well-organized literature survey that is also fun to read because of its great storytelling. I struggled a fair amount in writing this summary, because I kept wanting to somehow communicate the writing style; in the end I decided not to do it and to instead give a few examples of passages from the book in this post.

Passages:

Note: It is generally not allowed to have quotations this long from this book; I have specifically gotten permission to do so.

Here's an example of agents with evolved inner reward functions, which lead to the inner alignment problems we've previously worried about:

> They created a two-dimensional virtual world in which simulated organisms (or "agents") could move around a landscape, eat, be preyed upon, and reproduce. Each organism's "genetic code" contained the agent's reward function: how much it liked food, how much it disliked being near predators, and so forth. During its lifetime, it would use reinforcement learning to learn how to take actions to maximize these rewards. When an organism reproduced, its reward function would be passed on to its descendants, along with some random mutations. Ackley and Littman seeded an initial world population with a bunch of randomly generated agents.
>
> "And then," says Littman, "we just ran it, for seven million time steps, which was a lot at the time. The computers were slower then." What happens? As Littman summarizes: "Weird things happen."
>
> At a high level, most of the successful individual agents' reward functions ended up being fairly comprehensible. Food was typically viewed as good. Predators were typically viewed as bad. But a closer look revealed some bizarre quirks. Some agents, for instance, learned only to approach food if it was north of them, for instance, but not if it was south of them.
>
> "It didn't love food in all directions," says Littman. "There were these weird holes in [the reward function]. And if we fixed those holes, then the agents became so good at eating that they ate themselves to death."
>
> The virtual landscape Ackley and Littman had built contained areas with trees, where the agents could hide to avoid predators. The agents learned to just generally enjoy hanging out around trees. The agents that gravitated toward trees ended up surviving—because when the predators showed up, they had a ready place to hide.
>
> However, there was a problem. Their hardwired reward system, honed by their evolution, told them that hanging out around trees was good. Gradually their learning process would learn that going toward trees would be "good" according

> to this reward system, and venturing far from trees would be “bad.” As they learned over their lifetimes to optimize their behavior for this, and got better and better at latching onto tree areas and never leaving, they reached a point of what Ackley dubbed “tree senility.” They never left the trees, ran out of food, and starved to death.
>
> However, because this “tree senility” always managed to set in after the agents had reached their reproductive age, it was never selected against by evolution, and huge societies of tree-loving agents flourished.
>
> For Littman, there was a deeper message than the strangeness and arbitrariness of evolution. “It’s an interesting case study of: Sure, it has a reward function—but it’s not the reward function in isolation that’s meaningful. It’s the interaction between the reward function and the behavior that it engenders.”
>
> In particular, the tree-senile agents were born with a reward function that was optimal for them, provided they weren’t overly proficient at acting to maximize that reward. Once they grew more capable and more adept, they maxed out their reward function to their peril—and, ultimately, their doom.

Maybe everyone but me already knows this, but here’s one of the best examples I’ve seen about the benefits of transparency:

> Ambrosino was building a rule-based model using the pneumonia data. One night, as he was training the model, he noticed it had learned a rule that seemed very strange. The rule was “If the patient has a history of asthma, then they are low-risk and you should treat them as an outpatient.”
>
> Ambrosino didn’t know what to make of it. He showed it to Caruana. As Caruana recounts, “He’s like, ‘Rich, what do you think this means? It doesn’t make any sense.’ You don’t have to be a doctor to question whether asthma is good for you if you’ve got pneumonia.” The pair attended the next group meeting, where a number of doctors were present; maybe the MDs had an insight that had eluded the computer scientists. “They said, ‘You know, it’s probably a real pattern in the data.’ They said, ‘We consider asthma such a serious risk factor for pneumonia patients that we not only put them right in the hospital . . . we probably put them right in the ICU and critical care.’ ”
>
> The correlation that the rule-based system had learned, in other words, was real. Asthmatics really were, on average, less likely to die from pneumonia than the general population. But this was precisely because of the elevated level of care they received. “So the very care that the asthmatics are receiving that is making them low-risk is what the model would deny from those patients,” Caruana explains. “I think you can see the problem here.” A model that was recommending outpatient status for asthmatics wasn’t just wrong; it was life-threateningly dangerous.
>
> What Caruana immediately understood, looking at the bizarre logic that the rule-based system had found, was that his neural network must have captured the same logic, too—it just wasn’t as obvious.
>
> [...]
>
> Now, twenty years later, he had powerful interpretable models. It was like having a stronger microscope, and suddenly seeing the mites in your pillow, the bacteria

on your skin.

“I looked at it, and I was just like, ‘Oh my— I can’t believe it.’ It thinks chest pain is good for you. It thinks heart disease is good for you. It thinks being over 100 is good for you.…It thinks all these things are good for you that are just obviously not good for you.”

None of them made any more medical sense than asthma; the correlations were just as real, but again it was precisely the fact that these patients were prioritized for more intensive care that made them as likely to survive as they were.

“Thank God,” he says, “we didn’t ship the neural net.”

Finally, on the importance of reward shaping:

In his secret top-floor laboratory, though, Skinner had a different challenge before him: to figure out not which schedules of reinforcement ingrained simple behaviors most deeply, but rather how to engender fairly complex behavior merely by administering rewards. The difficulty became obvious when he and his colleagues one day tried to teach a pigeon how to bowl. They set up a miniature bowling alley, complete with wooden ball and toy pins, and intended to give the pigeon its first food reward as soon as it made a swipe at the ball. Unfortunately, nothing happened. The pigeon did no such thing. The experimenters waited and waited. . . and eventually ran out of patience.

Then they took a different tack. As Skinner recounts:

> We decided to reinforce any response which had the slightest resemblance to a swipe— perhaps, at first, merely the behavior of looking at the ball—and then to select responses which more closely approximated the final form. The result amazed us. In a few minutes, the ball was caroming off the walls of the box as if the pigeon had been a champion squash player.

The result was so startling and striking that two of Skinner’s researchers—the wife-and- husband team of Marian and Keller Breland—decided to give up their careers in academic psychology to start an animal-training company. “We wanted to try to make our living,” said Marian, “using Skinner’s principles of the control of behavior.” (Their friend Paul Meehl, whom we met briefly in Chapter 3, bet them $10 they would fail. He lost that bet, and they proudly framed his check.) Their company—Animal Behavior Enterprises—would become the largest company of its kind in the world, training all manner of animals to perform on television and film, in commercials, and at theme parks like SeaWorld. More than a living: they made an empire.

The date of AI Takeover is not the day the AI takes over

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

Instead, it's the point of no return—the day we AI risk reducers lose the ability to significantly reduce AI risk. This might happen years before classic milestones like "World GWP doubles in four years" and "Superhuman AGI is deployed."

The rest of this post explains, justifies, and expands on this obvious but underappreciated idea. (Toby Ord appreciates it; see quote below). I found myself explaining it repeatedly, so I wrote this post as a reference.

AI timelines often come up in career planning conversations. Insofar as AI timelines are short, career plans which take a long time to pay off are a bad idea, because by the time you reap the benefits of the plans it may already be too late. It may already be too late because AI takeover may already have happened.

But this isn't quite right, at least not when "AI takeover" is interpreted in the obvious way, as meaning that an AI or group of AIs is firmly in political control of the world, ordering humans about, monopolizing violence, etc. Even if AIs don't yet have that sort of political control, it may already be too late. Here are three examples: [UPDATE: More fleshed-out examples can be found in this new post.]

1. Superhuman agent AGI is still in its box but nobody knows how to align it and other actors are going to make their own version soon, and there isn't enough time to convince them of the risks. They will make and deploy agent AGI, it will be unaligned, and we have no way to oppose it except with our own unaligned AGI. Even if it takes years to actually conquer the world, it's already game over.

2. Various weak and narrow AIs are embedded in the economy and beginning to drive a slow takeoff; capabilities are improving much faster than safety/alignment techniques and due to all the money being made there's too much political opposition to slowing down capability growth or keeping AIs out of positions of power. We wish we had done more safety/alignment research earlier, or built a political movement earlier when opposition was lower.

3. Persuasion tools have destroyed collective epistemology in the relevant places. AI isn't very capable yet, except in the narrow domain of persuasion, but everything has become so politicized and tribal that we have no hope of getting AI projects or governments to take AI risk seriously. Their attention is dominated by the topics and ideas of powerful ideological factions that have access to more money and data (and thus better persuasion tools) than us. Alternatively, maybe we ourselves have fallen apart as a community, or become less good at seeking the truth and finding high-impact plans.

Conclusion: We should remember that when trying to predict the date of AI takeover, what we care about is the date it's too late for us to change the direction things are going; the date we have significantly less influence over the course of the future than we used to; the point of no return.

This is basically what Toby Ord said about x-risk: “So either because we’ve gone extinct or because there’s been some kind of irrevocable collapse of civilization or something similar. Or, in the case of climate change, where the effects are very delayed that we’re past the point of no return or something like that. So the idea is that we should focus on the time of action and the time when you can do something about it rather than the time when the particular event happens.”

Of course, influence over the future might not disappear all on one day; maybe there’ll be a gradual loss of control over several years. For that matter, maybe this gradual loss of control began years ago and continues now... We should keep these possibilities in mind as well.

[Edit: I now realize that I should distinguish between AI-induced points of no return and other points of no return. Our timelines forecasts and takeoff speeds discussions are talking about AI, so we should interpret them as being about AI-induced points of no return. Our all-things-considered view on e.g. whether to go to grad school should be informed by AI-induced-PONR timelines and also "timelines" for things like nuclear war, pandemics, etc.]

Postmortem to Petrov Day, 2020

The first underwater test of an atom bomb, in the US at Bikini Atoll in the Pacific, July 24, 1946.
Also, the LessWrong.com Frontpage this Petrov Day.

We failed.

The Frontpage was taken down for 24 hours.

Launch Attempts

At 1:26 am, ~275 users got an email (and 20 minutes later a PM) from me, with their unique launch codes. At this time, the Big Red Button was on the Frontpage, and codes could be submitted.

Two users attempted to submit launch codes. It was made clear that any such attempts would be deanonymised, as the button itself said "This is not an anonymous action".

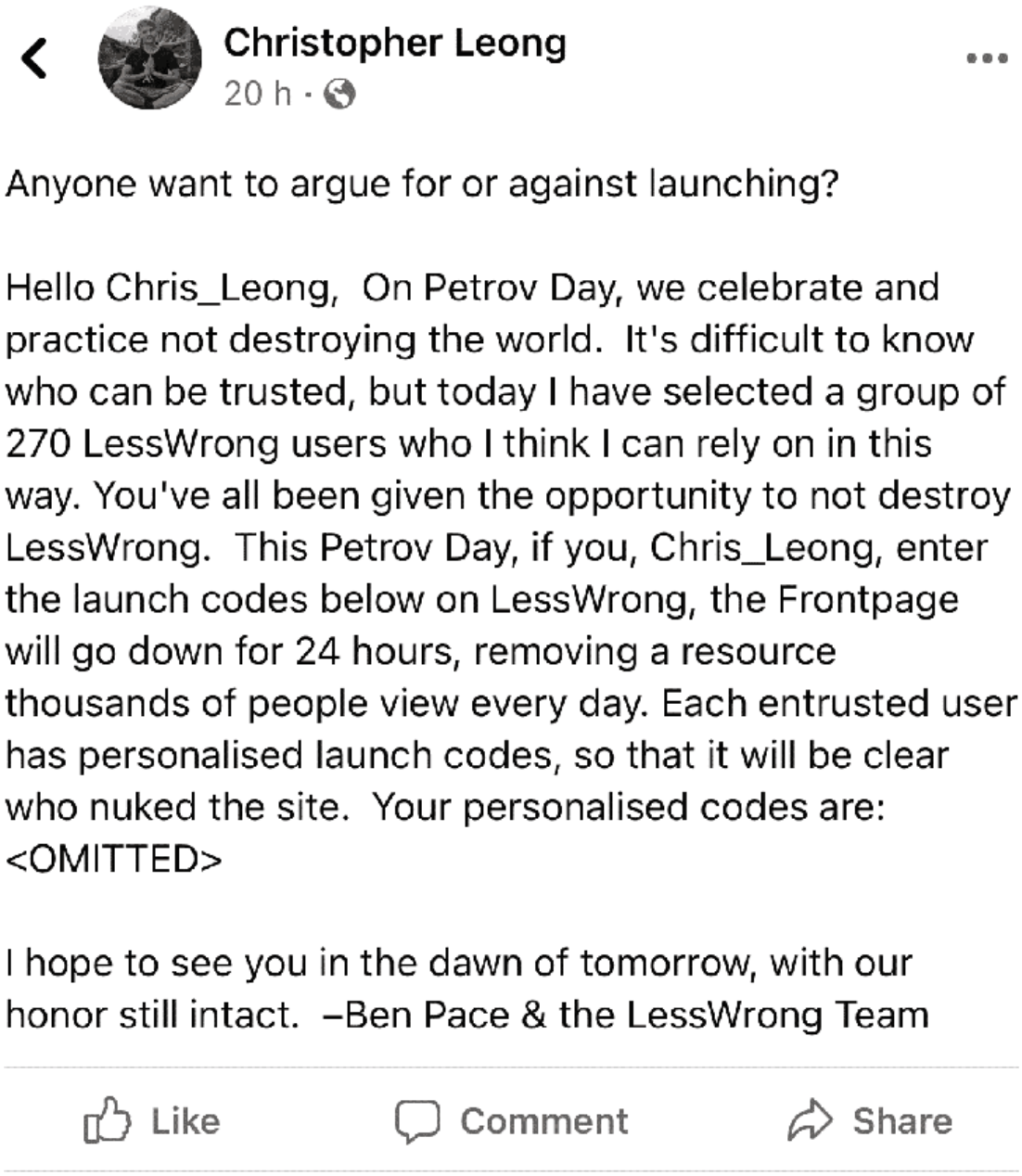
The users were: Grotace at 2:09 am, who submitted the random string "kdjssndksjnd" to no effect, and Chris_Leong at 4:33 am, who submitted his personalised launch codes, taking down the site.

What happened?

Note that the following includes private messaging metadata about a number of users. All users (except for the attacker) were asked for their explicit consent for it to be included, and it would have been anonymized had they declined, even if it was clearly deducible from the public record. The LessWrong team only looks into private messaging metadata when there is a suspicion of sockpuppeting, phishing, or other forms of abuse, or when permission has

been given by the user (e.g. for debugging). Additionally, Chris Leong commented on a draft of this post before it was published.

On receiving his codes, Chris Leong posted on LessWrong and Facebook asking whether he should take down the site.



He also wrote a comment on LessWrong:

> Should I press the button or not? I haven't pressed the button at the current time as it would be disappointing to people if they received the email, but someone pressed it while they were still asleep.

He did not get replies on the LessWrong comment before he took the site down. The first two comments on the Facebook post were as follows.

Commenter 1

> I kind of feel like you should take it down for a day. No individual should be trusted with the ability to destroy the world. We are all driven by personal incentive mechanisms and other systems must be developed to ensure it’s not that easy for one person to bring so much chaos into the world.
>
> Also... when’s the next time you’re going to get the opportunity to take down a massive site for a day legally?

Commenter 2

> LessWrong is archived elsewhere so the damage by bringing it down (especially just for a day) is not that much. Equally, there isn't much benefit that you would get by bringing it down. As the website says, if you decide to enter your launch codes, it's not anonymous. So the question becomes one of social capital.

After the second comment, the following happenned:

- A LW account was created with the username "petrov_day_admin_account".
- From 4:21 to 4:34 am (Pacific Time), this user opened 6 conversations, sending only 4 messages.
- They sent one message each to Chris_Leong, Raemon, Habryka, and Liron.
- They all contained identical body text with the name changed for each user (although Liron’s message began with “Hello Chris_Leong”).
- Chris Leong's message was sent at 4:26 am.
- A conversation was opened with adamzerner, and a second conversation was opened with Chris_Leong, but no messages were sent, as by that time it was no longer necessary.

Here were the next three comments on Chris's Facebook post. (By all accounts the relevant text below matches the text in the messages sent to the above users.)

Commenter 2

> Well, I sent out this message to a lot of users, now the site has gone down, but it doesn't say who pressed the button so it could still be a coincidence:
>
> Hello <NAME>,
>
> You are part of a smaller group of 30 users who has been selected for the second part of this experiment. In order for the website not to go down, at least 5 of these selected users must enter their codes within 30 minutes of receiving this message, and at least 20 of these users must enter their codes within 6 hours of receiving the message. To keep the site up, please enter your codes as soon as possible. You will be asked to complete a short survey afterwards.

Chris Leong

> Damn it, you got me! Well done

Commenter 2

> No problem :D Maybe the lesson is that if you can destroy the world you shouldn't let people know? Not really sure, just thought it would be fun.

Chris Leong

> Well played!

And that is the story of how LessWrong was taken down for 24 hours by Chris Leong, on Petrov Day.

What went wrong?

One of the most important fact about this year is how many users were given codes. Last year had success with ~125 users getting codes, and this year the total was more than doubled to ~275 users. This weakened the selection filter which ultimately failed.

Well-intentioned does not mean secure

The first mistake was in how I chose which users to entrust with codes.

The main buckets I made were high karma users I know and have a judge of character (~140), high karma users I don't currently know very well (~50), and low karma users I know and have a judge of character (~80).

I weighed my selection toward users using real names instead of pseudonyms, as reputation is a factor people track when deciding whether their actions are good or bad.

For high karma users I don't know very well, I spent about 5 hours reading through their comments and posts, asking myself questions like the following:

- Does this user write comments in good-faith?
- Does this user put in time and effort to explain their ideas in the comment sections?
- Does this user write short, snarky comments?
- Does this user try to help the site grow, such as by proposing and giving feedback on new features?
- Does this user have a grudge against LessWrong?
- Does this user seem to understand and partake in the culture and sense-of-humor of LessWrong?

Overall, I picked users on the basis I thought they would not *choose* to take the website down. I didn't consider for a moment that *it would not be a matter of direct choice at all*, but that someone would fall for a standard phishing attack.

A classic mistake on my part. In that it is the opening paragraph of Eliezer's classic paper Cognitive Biases Potentially Affecting Judgment of Global Risks:

> All else being equal, not many people would prefer to destroy the world. Even faceless corporations, meddling governments, reckless scientists, and other agents of doom require a world in which to achieve their goals of profit, order, tenure, or other villainies. If our extinction proceeds slowly enough to allow a moment of horrified realization, the doers of the deed will likely be quite taken aback on realizing that they have actually destroyed the world. Therefore I suggest that if the Earth is destroyed, it will probably be by mistake.

So the first lesson is that if you want to entrust others with powers that can be easily abused or used destructively, it is not sufficient for them to be well-intentioned, it is also necessary for them to be **hard enough to trick that the payoff isn't worth it** for any external adversary.

Users saw different meanings to the ritual

The launch codes gave users the ability to take down the Frontpage for 24 hours, a page that a few thousand people visit every day. A single user being unable to use it is a simple irritation; once you multiply this by thousands it becomes a non-trivial communal resource.

However, while many users understood that they were being given responsibility for the commons, many users did not see a clear connection to Petrov's situation, whose choices were different in many ways to the LessWrong Big Red Button. Furthermore Neel Nanda commented he felt that the communications around it felt like “RPG flavor text” which suggested it was not to be taken seriously. This confusion made the day less meaningful for a number of people, and in particular some people felt confused about how it could possibly symbolize Petrov's choice and thus thought it was "just a game".

To respond to that briefly: when you are given responsibility for a communal resource, even if the person giving it to you says “It’s fine to destroy it – play along!” then you’re still supposed to think for yourself about whether they’re right. One of the core virtues of Petrov day is how Petrov took responsibility for launching nuclear armageddon, and he didn’t “assume the people in charge knew best” or “just do what expected of him”. So in some ways I feel like this challenge was unfair in the same way that reality is unfair, and it is a question about whether people noticed their responsibility to the commons without being told that they were supposed to take responsibility. On the other hand, in some ways this was a harder challenge than Petrov’s, because the stakes were measured in a much smaller communal resource and not in the lives of a billion people, and because Petrov had more lead time to think before the time came when he had to make the decision.

Nonetheless, several users requested that the relation between the exercise and Petrov’s decision be made more explicitly before next year, which seems quite reasonable to me. So in advance of next year, I’ll write up an explanation of why I think the Big Red Button is an accurate symbolic setup to Petrov's situation.

I did not check who submitted launch codes last year

Note that this section includes private data about a user (Chris Leong). The user was asked for their explicit consent before it was included, and had they declined then this section would have been redacted. It is further the case that Chris Leong commented on a draft of this post before it was published.

Last year, Chris_Leong pressed the Big Red Button, guessed an incorrect launch code, and submitted it.

(He was not given real launch codes last year.)

I did not think to check the group of 7 accounts with non-zero karma who submitted random launch codes last year, and ensure that they were not in this year's batch. I will be doing so in future.

Chris was the only one in that batch who was given codes this year. To me, this is suggestive evidence that of the people in this year's batch Chris was the most likely to press the Big Red Button, and causes me to update that had he in particular not been in the batch, there quite likely would not have been a next person to press the button instead.

In a comment on a draft of this post, Chris noted that he doesn't recollect doing this last year, and also asked me what message was displayed next to the button last year.

The answer is that there was no message next to the button last year, it looked like this:

The Big Red Button in 2019.

Indeed, last year I know of a user (not Chris Leong) who visited the site, clicked the red button, and entered and submitted codes, *before* finding out what the button did.

As a result of that user, this year we changed the red button to the following, so that mistake would not happen again.

The Big Red Button in 2020.

I nonetheless think that playing with the button is a clear sign that I shouldn't have included someone in this year's batch.

I should not have given launch codes to Chris

Chris is well-intentioned, but this did not mean he could be entrusted with this communal resource. A number of factors in retrospect show that I should not have given the codes to Chris.

- He didn’t think he had a responsibility to the commons in this situation. He treated it like a game, even though he was given the ability to destroy a non-trivial communal resource.
- He failed to model that he was not secure and that people would like to trick him on this matter, so posted that he had codes in a place where someone tricked him into using them.
- He didn’t realize that a lot of users cared seriously about this tradition and this exercise, which could have led him to rethink the above two bullets.

By the way, if you'd like to read Chris's account of what happened, see his post On Destroying The World.

Looking Ahead

This is the 14th year of Petrov Day, and the 2nd year that we have observed Petrov Day with the Big Red Button ritual on LessWrong. On many continents there were a number of annual ceremonies celebrating Petrov Day, even in this socially distanced time, remembering how fragile the world is and our responsibility to not destroy it.

I learned a lot this year. I think many of us did.

(Thank you to everyone who participated in the lively discussion.)

I definitely updated upward on how many counterfactual worlds ended this way in the Cold War, where some third party to the conflict attempted to trick one of the players into instigating an attack, perhaps by posing as a relevant authority and saying it was a 'training exercise'.

I think this is one of the more innocuous ways that the site could've gone down (not through malice, but through lack of taking it seriously, light trolling, and a security failure).

I hadn't though about red-teaming the site. We often learn the most important principles the hard way. While pentesting your friends for their private details and other low-consequence effects is great, I don't think much of people who unilaterally try to take down something more like a public utility for 24 hours. But it is certainly a valuable addition to the setup, and I'll look into setting up some red-teaming next year.

Last year I assigned a 40% probability to the site going down, and it didn't. This year I assigned a 20% probability to the site going down, and it did.

I think had this happened the first year we tried it, I would expect a common response to be that it was obvious that we would fail, and that we were far too credulous if we that thought we could give codes to over one hundred rationalists and one of them wouldn't take down the site.

Then, after last year's success, I received many messages saying that it was clear nobody would take it down this year. So in some ways this has been good for making the challenge level clear – this is a real coordination problem, that we can fail, and isn't overdetermined in either direction.

Onwards and upwards. This was a collective coordination problem for LessWrong, and we're 1 for 2. We'll return next Petrov Day, for another round of sitting around and not pressing destructive buttons, alongside our ceremonies and celebrations.

So far we've had one Petrov Day on LessWrong without pressing the Big Red Button.

I hope there are many more to come.

Afterword

In spite of Chris taking down the site this week, I still appreciate Chris's many positive contributions to LessWrong and our associated communities, and for everyone else's sake I'll briefly list some of them here.

- Chris has written over 100 LessWrong posts and over 1000 comments, resulting in about 4000 karma and 60 Alignment Forum karma.
- Chris wrote the post Decoupling vs Contextualising Norms (concisely summarizing an idea by John Nerst from off-LessWrong) which was curated, has 145 karma, and scored highly in the LessWrong Review of 2018.
- Chris has contributed many posts and threads attempting to improve LessWrong and provide themed threads, such as Monthly Meta: Common Knowledge, Experimental Open Threads, No option to report spam, Beta - First Impressions, Suggestion: New material shouldn't be posted too fast, Tags of Sub-Groups, Using accounts as "group accounts", What is meant by Simulacra Levels?, and Suggestion: Implement a Change My View Feature.
- In early March, as many of us were beginning to take the coronavirus pandemic seriously, Chris created, moderated, and made regular contributions to the "Effective Altruism Coronavirus Discussion" Facebook Group, which is a private group with 1,400 members. I think this got a lot of important information to people in our affiliated communities in a time when such information was very important and scarce (e.g. he shared Zvi's regular Covid updates).

I'm grateful for all of the above contributions Chris has made, his contribution to LessWrong has been clearly net positive.

Covid 10/1: The Long Haul

If you were watching the so-called ‘presidential debate’ on Tuesday night, first off, you have my sympathies. It was the day after the Day of Atonement. If you watched, no matter what wrongs you may have committed this past year, and no matter who you intend to vote for, no one can deny that you have atoned. Your slate is clean.

Alas, the country is not so lucky. Getting away clean is not a near-term prospect on any level.

What did we learn from the Covid-19 portion of the debate?

Very little. The focus was on Biden blaming Trump for things being terrible, and Trump saying things are great and blaming Biden claiming that with Biden in charge things would have been worse. No one said anything about any of the real issues except for masks. On masks, Trump decided to dispute that there was agreement on masks, to point out that people changed their mind about masks, and so on, in case his supporters were in danger of protecting themselves or others by wearing one.

Biden’s criticisms of Trump left out most of the worst things Trump did regarding Covid-19. Biden’s plans, as stated, didn’t provide the help we need to solve Covid-19. Mostly, what we learned is what we already knew. Biden has little interest in talking about the ways to actually solve the problem, and mostly does correct symbolic actions like supporting PPE or small business or wearing masks while blaming Trump for not doing so. Whereas Trump actively gets in the way of solving the problem and lies about, well, basically everything. Not where one hopes the choices to be, but hopefully an easy choice nonetheless.

Biden repeated in the debate the general expectation that another wave of infections and deaths is coming Real Soon Now, and that deaths may double over the next few months, as the rate goes up by a factor of five or more. That’s the new Very Serious Person position.

Along with the old Very Serious Person position that herd immunity is of course ending Real Soon Now, probably last week.

Do we see any signs of any of that? Is it justified?

Let’s run the numbers.

Positive Test Counts

Date	WEST	MIDWEST	SOUTH	NORTHEAST
Aug 6-Aug 12	93042	61931	188486	21569
Aug 13-Aug 19	80887	63384	156998	20857
Aug 20-Aug 26	67545	66540	132322	18707
Aug 7-Sep 2	55000	75401	127414	21056
Sep 3-Sep 9	47273	72439	106408	21926
Sep 10-Sep 16	45050	75264	115812	23755
Sep 17-Sep 23	54025	85381	127732	23342

Sep 24-Sep 30	55496	92932	106300	27214

The Midwest and Northeast numbers are very troubling, as things are clearly headed in the wrong direction. The West is treading water, and the South looks excellent. Test counts are slightly up overall.

Deaths

Date	WEST	MIDWEST	SOUTH	NORTHEAST
July 23-July 29	1707	700	4443	568
July 30-Aug 5	1831	719	4379	365
Aug 6-Aug 12	1738	663	4554	453
Aug 13-Aug 19	1576	850	4264	422
Aug 20-Aug 26	1503	745	3876	375
Aug 27-Sep 2	1245	759	3631	334
Sep 3-Sep 9	1141	771	2717	329
Sep 10-Sep 16	1159	954	3199	373
Sep 17-Sep 23	1016	893	2695	399
Sep 24-Sep 30	934	990	2619	360

More signs that things in the Midwest continue to get worse, but the West and South continue to recover. The Northeast doesn’t look bad either. In total, it’s the best combined number in a long time.

Positive Tests By Region

Percentages	Northeast	Midwest	South	West
7/30 to 8/5	2.58%	7.26%	12.35%	6.68%
8/6 to 8/13	2.30%	5.67%	14.67%	6.98%
8/13 to 8/20	2.06%	5.62%	9.41%	6.47%
8/20 to 8/26	1.86%	5.78%	9.93%	5.88%
8/27 to 9/2	1.87%	6.37%	9.38%	4.78%
9/3 to 9/9	1.97%	6.02%	8.48%	4.13%
9/10 to 9/16	2.41%	5.99%	11.35%	4.49%
9/17 to 9/23	2.20%	5.96%	7.13%	4.11%
9/24 to 9/30	2.60%	6.17%	6.18%	4.27%

Trouble slowly brewing across the Northeast. The South continues to improve.

Test Counts

Date	USA tests	Positive %	NY tests	Positive %	Cumulative Positives
July 30-Aug 5	5,107,739	7.8%	484,245	1.0%	1.46%
Aug 6-Aug 12	5,121,011	7.3%	506,524	0.9%	1.58%
Aug 13-Aug 19	5,293,536	6.2%	548,421	0.8%	1.68%
Aug 20-Aug 26	4,785,056	6.0%	553,369	0.7%	1.77%
Aug 27-Sep 2	5,042,113	5.5%	611,721	0.8%	1.85%
Sep 3-Sep 9	4,850,253	5.3%	552,624	0.9%	1.93%
Sep 10-Sep 16	4,632,005	5.8%	559,463	0.9%	2.01%
Sep 17-Sep 23	5,719,327	5.2%	610,802	0.9%	2.10%
Sep 24-Sep 30	5,857,097	5.1%	618,378	1.1%	2.19%

New York is in (medium term, slow moving) trouble. From what I've seen, a lot of it is concentrated on areas of Brooklyn and Queens, especially Orthodox Jewish areas that ignored rules during the high holidays, but it's a big enough effect and trend that the problem is clearly widespread. Things are in no way out of control, but trends continue to be negative and if anything things are opening up more, so until the control systems set in locally, things will get worse.

Nationwide, however, we have a record number of tests, with the lowest positive rate since mid-June. We have the lowest weekly death count since mid-July.

There's no sign things are about to clear up. But there's also no sign of this huge impending disaster the media are once again warning us about.

Yet they continue to do this. Why?

Some of it is that testing went down then increased again and they're calling that 'rising case counts' again because *yes we really are this stupid and dysfunctional.*

You can call Donald Trump and various others whatever you like for suppressing testing in order to make the numbers look better, but the only way to stop such tactics is to stop being fooled by them time after time. I am not optimistic.

The headline from CNN linked above (here it is again) tells us to be alarmed that 21 states have rising case numbers, while testing increases, and doesn't think we can understand that 21 is less than half of 50.

The other half is that models and those what like to be Very Serious People are making two assumptions to force this pending wave to happen.

They assume that Winter Is Coming means things get worse. And they continue to warn about immunity in all ways.

We need to push back and not leave this to the White House. They're kind of busy, and rather short of credibility. Atlas disputes Redfield coronavirus vulnerability estimate: 'We are not all susceptible to infection' is the White House directly calling

out the CDC, and in this case being entirely correct. The idea that everyone who doesn't test positive for antibodies is 'susceptible to infection' is obvious nonsense designed to twist the data into knots and scare people. Unacceptable. Yes, the source in question often lies its ass off in other ways. That only makes this all that much harder.

I Herd Some People Had Immunity and Then Lost It, Or Never Got It To Begin With After Being Infected, Except With No Actual Examples

Alas, an ongoing series. Lots of speculation this week.

Marginal Revolution links to a study pointing out what we already know, that most coronaviruses do not create lifelong immunity.

Another data point I heard a few people point to is that a previously hard-hit NYC neighborhood is being hit again. Similar data points are cited for European cities and such.

Thus, the Very Serious Person consensus seems to be that immunity is fading and reinfections are happening all over the place as we speak. They just… can't find examples of actual people that got reinfected, despite such a story containing a large number of people being an obvious way to get tons of clicks and head the national news, and also scare everyone in a way that such people think is good.

What we know is, it is now October. Lots of people were infected in March, if not sooner. Almost no one is known to have been infected in March, then in September, or anything like that. Which is what you would see, *if immunity was fading.*

Similarly, we continue to see people equate positive antibody tests with immunity, despite it being rather clear that this is only one of several means of immunity. The immune system has a lot of tools at its disposal, and all that.

So once again, *until we see lots of reinfections of particular people,* all we know each day is it is another day before serious reinfection chances occur, and our expectation for immunity length goes up by just under two days because of the Lindy rule – however long it has lasted so far, it probably will continue to last on average about that long, then *slowly* fade, is a reasonable prior for the mean result.

How do we explain the data we do see?

Obvious Nonsense Paper of the Week

On a related and but different note is this paper that came up this week: Evolution of COVID-19 cases in selected low- and middle-income countries: past the herd immunity peak?

It's a textbook example of how deeply the SEIR folly goes. The paper looks at a curve of infections, *assumes that everyone is always identical in every way within the country,* then uses that to figure out how many people *must* have been infected in order to cause the reduction in infections! That this *proves* most people were infected! Then based on that, they point out how low the infection fatality rate was!

Seriously, *this is what passes for serious modeling these days.* This got into the news cycle.

The estimated base reproduction numbers, the R0 are estimated as no more than 2. Based on that and the curve, they then claim that this means 50%-80% of people must have been infected. The 'detection rate' for infections is then surmised to go from a high of 5% in South Africa, to a low of *0.2%* in Kenya. Not death rate, detection rate.

Such utterly obvious nonsense.

The numbers are so utterly crazy.

Florida Says Yolo

Florida's Governor has had enough. No more restrictions on businesses. No enforcement of mask mandates by cities. It's time, he's saying, to let private individuals make their own choices, and whatever happens happens. His 'health experts' agree with this, because if you want to find an expert who agrees with a given position, or at least is willing to say they agree, you can almost always find that expert.

The usual scolds and Very Serious People are out in force about how awful this is and that everyone in Florida should once again prepare to die. Things are bad after all that locking down, the Very Serious People say. Surely you can't stop locking down now!

The Governor is closer to correct than the Very Serious People.

What is the alternative proposal? To continue to put our lives on hold and our economy into shambles until we finish the vaccinations?

How long a lock down before it's better to just *get the damn virus already* and take your chances, if your risk isn't that high?

Lethality is down. Hospitals being overwhelmed is highly unlikely to happen, given what happened in the previous wave. We now know how to manage our risk if we want to do so, and can make informed choices. Make trade-offs. It's time to let people decide what they want to do, and live their lives. It's not like everyone is going to suddenly go back to normal. Some people will choose to do that. Others won't. We've been over this many times.

There are two counter-arguments.

One is that the vaccine is coming Real Soon Now, but those same Very Serious People are saying that the vaccine is at least months away plus more months for deployment. If that's true, then that's too long for many people, who can make a rational choice not to wait. For those who do want to wait, it's a reasonable amount of time to deal with a higher outside risk level.

The other is the externality argument at the heart of it all. You taking on risk puts me at risk.

The basic response I have here is that no, it doesn't, not in a meaningful way. Not anymore, beyond the specific people you choose to have close contact with.

That's because there's a solid range of risk levels, where risk is too high to allow for activities that involve substantial exposure, but not so high as to make it impossible to protect yourself (e.g. it's not so dense that you're worried about things like the Manhattan air having persistent miasma.)

If we had a practical path to getting below that range and sustaining that process, great, that would be worth paying a big price to do. We don't. This virus isn't going away short of a vaccine, period.

If we were in danger of rising above that range, or going so high we'd break the medical system, we'd have to think carefully about what we want to do to prevent that. But those days are over. We ran the experiments. The herd immunity we have plus the control systems in place won't let it get to that point.

That doesn't mean this is a zero cost. It most certainly is not. But banning things doesn't seem reasonable.

I do think that forcibly removing municipal mask bans is still doubleplus ungood, but businesses are still free to require them, and if enough people stay away from those businesses that don't require them, that is what will mostly happen. It's bad, but less of a big deal than it sounds like it is.

The big mistake is indoor dining. Indoor dining is a *terrible* cost-benefit ratio. It's one of the most dangerous things you can do. The experience is nice, but it's in no way vital. The reason indoor dining is happening is because without it, the bars and restaurants would die, with long term consequences.

Going forward, of course, if things get bad in Florida they'll blame it on this, even if things are equally bad elsewhere. If things don't get bad, they'll still blame this anyway. Right or wrong, that's how the Very Serious People roll.

The right answer, of course, is utterly obvious and in front of our face, and has zero chance of happening. It's to tax. Rarely is banning things outright a good idea. If we put a large tax on indoor dining, ideally as a function of relative safety but a fixed number would do, that's the logical approach. We could do the same with other risky activities. If people don't want to pay, then it wasn't worth doing. If they do pay, then it was worth it, and we can use that money to fund our other efforts.

A number of other states are also in Yolo Mode. I learned this morning that Massachusetts is continuing to loosen restrictions, based on a justification of dropping positive test percentages, despite rising hospitalizations and rising numbers of cases. All metrics matter, but it seems clear what the result will be.

The question remains, if we're not willing to do what it takes to stop this, why should we also ruin everything else along the way to not stopping it?

Going Down to Denver, Going To Have Ourselves a Time

Denver is on it with two feel good stories this week.

First, they were able to detect infection using wastewater, and contain it. This needs to be standard procedure, everywhere.

Second, Denver Broncos fill stadium with South Park cardboard cutouts.

In Other News

CDC pulls coronavirus surveyors out of Minnesota after they reported harassment, racism. Our country might indeed have some problems it needs to confront. Instead, of course, we once again ran away from several of them at once, which seems fitting. We don't care enough to work through it.

The Long Haul

The great unknown is the frequency and severity of 'long haul' Covid. People, including many fully young and healthy people, suffer for months after infection and potentially have permanent damage that could substantially lower their life satisfaction. This may have happened to one of my close friends (that essentially none of my readers would know) who was otherwise healthy. I'm more afraid of the long haul problem than I am of dying from Covid.

All we know so far about 'long haul' Covid – estimated to affect 600,000 people in the UK estimates that 12% of those infected have symptoms that last longer than 30 days. That's a very broad definition of 'long haul' on duration and on severity, and I'm guessing this is a large underestimate of the number of actual cases in the United Kingdom. Even fully buying what's here, we don't get much of a handle on how to assess our personal risks and decide how much to care about them.

As Their Numbers Grow, COVID-19 "Long Haulers" Stump Experts is even less helpful. I don't feel *any* better informed than before from that, as to what I actually need to know.

I wish I had answers here, and encourage those who have any reasonable estimates at all to share them and explain their reasoning. We need to figure this out.

The Darwin Game

~~Click here to participate. Entries must be submitted on October 18th, 2020 or earlier.~~

Entry is now closed.

In 2017, Zvi posted an exciting story about The Darwin Game, a variation of iterated prisoner's dilemma.

I will run my own version of the game in the week following October 18th, 2020. **You do not have to know how to program in order to participate.** I will code simple bots for non-programmers. If you do know how to program then you may create your own complicated bot.

Here are the rules. Changes from Zvi's original game are in brackets [like this].

> For the first round, each player gets 100 copies of their program in the pool, and the pool pairs those programs at random. You can and often will play against yourself.
>
> Each pair now plays an iterated prisoner's dilemma variation, as follows. Each turn, each player simultaneously submits [an integer] from 0 to 5. If the two numbers add up to 5 or less, both players earn points equal to their number. If the two numbers add up to 6 or more, neither player gets points. This game then lasts for a large but unknown number of turns, so no one knows when the game is about to end; [I guarantee it will be at least 100 turns per iterated prisoner's dilemma].
>
> Each pairing is independent of every other pairing. [You do know what round of the game it is and that you are facing an opponent. If you face a copy of yourself you are automatically awarded the maximum 5 points per round (2.5 points per bot). You otherwise do not know any history of the game to this point.] Your decision algorithm does the same thing each pairing.
>
> At the end of the round, all of the points scored by all of your copies are combined. Your percentage of all the points scored by all programs becomes the percentage of the pool your program gets in the next round. So if you score 10% more points, you get 10% more copies next round, and over time successful programs will displace less successful programs. Hence the name, The Darwin Game.
>
> Your goal is to have as many copies in the pool at the end of the last round as possible, or failing that, to survive as many rounds as possible with at least one copy.
>
> [I will attempt to iterate until there is a stable equilibrium.]

I will add some silly bots of my own to make the early game more interesting.

Instructions for non-programmers

Please give a simple explanation of what you want your bot to do. Write it with mathematical precision. If your specification is even slightly ambiguous then you will be disqualified.

Instructions for programmers

Write a program of the following format.

```
class TitForTatBot():
    def __init__(self, round=0): # Edit: Changed "1" to "0"
        self.turn = 0
        self.round = round
        self.previous = None
    def move(self, previous=None):
        self.turn += 1
        if self.previous:
            output = self.previous
            self.previous = previous
            return output
        else:
            return 2

# Edit 2020-10-11: The above code is wrong. The properly-implemented TFT `move`
method looks like this.
#    def move(self, previous=None):
#        self.turn += 1
#        if previous == None:
#            return 2
#        else:
#            return previous
```

Your class must have an `__init__(self, round=1)` intializer and a `move(self, previous=None)` method. You may write your class in Python 3 or Hy.

Unlike Zvi's original game, you do get to know what round it is. Rounds are indexed starting at 0. The `previous` parameter of the `move` method is an integer indicating what your opponent did last iteration. If it is the first iteration then `previous` equals `None`.

A new instance of your class will be initialized in each round. You may save whatever information you want into the class instance's fields but you may not save information between rounds or between class instantiations. The `move` method must always return an integer from 0 to 5 inclusive.

You may import standard libraries like `random`, `scikit` and `numpy`.

Coordinating with other players

Anyone can play, but only players with a Less Wrong account that existed before I declared this tournament will be allowed to coordinate out-of-game. This rule exists to prevent players from submitting multiple entries to this contest and self-coordinating. Coordinating with other people is encouraged. Coordinating with yourself between multiple separate entries is cheating.

~~Click here to participate. Entries must be submitted on October 18th, 2020 or earlier.~~

Entry is now closed.

Can we hold intellectuals to similar public standards as athletes?

Professional athletes are arguably the most publicly understood meritocracy around. There are public records of thousands of different attributes for each player. When athletes stop performing well, this is discussed at length by enthusiasts, and it's understood when they are kicked off their respective teams. The important stuff is out in the open. There's a culture of honest, open, and candid communication around meritocratic competence and value.

This isn't only valuable to help team decisions. It also helps data scientists learn which sorts of characteristics and records correlate best with long term success. As sufficient data is collected, whole new schools of thought emerge, and these coincide with innovative and effective strategies for future talent selection. See Moneyball or the entire field of sabermetrics.

In comparison, our standards for intellectuals are quite prosaic. If I want to get a sense of just how good LeBron James is I can look through tables and tables or organized data and metrics. If I don't trust one metric I have dozens of others to choose.

However, if I want to know how much to trust and value Jonathan Haidt I'm honestly not sure what to do. Some ideas:

1. Read most of his work, then do a large set of Epistemic Spot Checks and more to get a sense of how correct and novel it is.
2. Teach myself a fair amount of Psychology, get a set of Academic Journal subscriptions, then read critiques and counter critiques of his work.
3. Investigate his citation stats.
4. Read the "Reception" part of his Wikipedia page and hope that my attempt to infer his qualities from that is successful.
5. Use some fairly quick "gut level" heuristics to guess.
6. Ask my friends and hope that they did a thorough job of the above, or have discussed the issue with other friends who did.

Of course, even if I do this for Jonathan Haidt broadly, I'd really want narrow breakdowns. Maybe his old work is really great, but in the last 5 years his motives have changed. Perhaps his knowledge and discussion of some parts of Psychology is quite on point, but his meanderings into Philosophy are simplistic and overstated.[1]

This is important because evaluating intellectuals is dramatically more impactful than evaluating athletes. When I choose incorrectly, I could easily waste lost of time and money, or be dramatically misled and develop a systematically biased worldview. It also leads to incentive problems. If intellectuals recognize the public's lack of discernment, then they have less incentive to do an actual good job, and more of an incentive to signal harder in uncorrelated ways.

I hear one argument: "*It's obvious which intellectuals are good and bad. Just read them a little bit.*" I don't agree with this argument. For one, Expert Political Judgement provided a fair amount of evidence for just how poorly calibrated all famous and well esteemed intellectuals seem to be.

One could imagine an organization saying "enough is enough" and setting up a list of comprehensive grades for public intellectuals on an extensive series of metrics. I imagine this would be treated with a fantastical amount of vitriol.

"*What about my privacy? This is an affront to Academics you should be trying to help.*"

"*What if people misunderstand the metrics? They'll incorrectly assume that some intellectuals are doing a poor job, and that could be terrible for their careers.*"

"*We can't trust people to see some attempts at quantifying the impossible. Let them read the sources and make up their own minds.*"

I'm sure professional athletes said the same thing when public metrics began to appear. Generally new signals get push back. There will be winners and losers, and the losers fight back much harder than the winners encourage. In this case the losers would likely be the least epistemically modest of the intellectuals, a *particularly nasty* bunch. But if signals can persist, they get accepted as part of the way things are and life moves on.

Long and comprehensive grading systems similar to that used by athletes would probably be overkill, especially to start with. Any work here would be very expensive to carry out and it's not obvious who would pay for it. I would expect that "intellectual reviews" would get fewer hits than "tv reviews", but that those hits would be much more impactful. I'd be excited to hear for simple proposals. Perhaps it would be possible to get many of the possible benefits while not having to face some of the many possible costs.

What should count as an intellectual? It's a fuzzy line, but I would favor an expansive definition. If someone is making public claims about important and uncertain topics and has a substantial following, these readers should have effective methods of evaluating them.

Meritocracy matters. Having good intellectuals and making it obvious how good these intellectuals are matters. Developing thought out standards, rubrics, and metrics for quality is really the way to ensure good signals. There are definitely ways of doing this poorly, but the status quo is *really really bad*.

[1] I'm using Jonathan Haidt because I think of him as a generally well respected academic who has somewhat controversial views. I personally find him to be every interesting.

A prior for technological discontinuities

Introduction

I looked at 50 technologies taken from a Wikipedia list History of technology, which I expect to provide a mostly random list of technologies. Of these 50 technologies, I think that 19 have a discontinuity, 13 might have one, and 18 probably don't. Of these, I'd call 12 "big" discontinuities, for an initial probability estimate of 12/50=24% I provide other estimates in the "More elaborate models for computing the base rate of big discontinuities."

Unlike some previous work by AI Impacts (or, for that matter, by myself), I am able to produce something which looks like a prior because I consider a broad bag of different technologies, and then ask which proportion have discontinuities. Previous approaches have specifically looked for discontinuities and found examples, thereby not being able to estimate their prevalence.

The broad bag of technologies I draw from was produced by Wikipedia editors who followed their own designs. They most likely weren't thinking in terms of discontinuities, and won't have selected for them. However, these editors might still have been subject to availability bias, Anglicism bias, etc. This might make the dataset mildly imperfect, that is, not completely representative of all possible technologies, but I'd say that most likely it's still good enough.

Furthermore, I didn't limit myself to discontinuities which are easily quantifiable or for which data is relatively easy to gather; instead I quickly familiarized myself with each technology in my list, mostly by reading the Wikipedia entry, and used my best judgement as to whether there was a discontinuity. This method is less rigorous than previous work, but doesn't fall prey to Goodhart's law: I want a prior for all discontinuities, not only for the quantifiable ones, or for the ones for which there is numerical data.

However, this method does give greater weight to my own subjective judgment. In particular, I suspect that I, being a person with an interest in technological discontinuities, might produce a higher rate of false positives. One could dilute this effect by pooling many people's assessments, like in Assessing Kurzweil's predictions for 2019.

All data is freely available here. While gathering it, I came across some somewhat interesting anecdotes, some of which are gathered in this shortform.

Many thanks to Misha Yagudin, Gavin Leech and Jaime Sevilla for feedback on this document.

Table of contents

- Introduction
- Discontinuity stories

- More elaborate probabilities
- Conclusion

Discontinuity stories

One byproduct of having looked at a big bag of technologies which appear to show a discontinuity is that I can outline some mechanisms or stories by which they happen. Here is a brief list:

- Sharp pioneers focus on and solve problem (Wright brothers, Gutenberg, Marconi, etc.)

  > Using a methodological approach and concentrating on the controllability of the aircraft, the brothers built and tested a series of kite and glider designs from 1900 to 1902 before attempting to build a powered design. The gliders worked, but not as well as the Wrights had expected based on the experiments and writings of their 19th-century predecessors. Their first glider, launched in 1900, had only about half the lift they anticipated. Their second glider, built the following year, performed even more poorly. Rather than giving up, the Wrights constructed their own wind tunnel and created a number of sophisticated devices to measure lift and drag on the 200 wing designs they tested. As a result, the Wrights corrected earlier mistakes in calculations regarding drag and lift. Their testing and calculating produced a third glider with a higher aspect ratio and true three-axis control. They flew it successfully hundreds of times in 1902, and it performed far better than the previous models. By using a rigorous system of experimentation, involving wind-tunnel testing of airfoils and flight testing of full-size prototypes, the Wrights not only built a working aircraft, the Wright Flyer, but also helped advance the science of aeronautical engineering.

- Conflict (and perhaps massive state funding) catalyzes project (radar, nuclear weapons, Bessemer process, space race, rockets)
- Serendipity; inventors stumble upon a discovery (radio telescopy, perhaps polymerase chain reaction, purportedly Carl Frosch and Lincoln Derick's discovery of surface passivation). Purportedly, penicillin (which is not in my dataset) was also discovered by accident. One might choose to doubt this category because a fortuitous discovery makes for a nicer story.
- Industrial revolution makes something much cheaper/viable/profitable (furniture, glass, petroleum, candles). A technology of particular interest is the centrifugal governor and other tools in the history of automation, which made other technologies undergo a discontinuity in terms of price. For example:

  > The logic performed by telephone switching relays was the inspiration for the digital computer. The first commercially successful glass bottle blowing machine was an automatic model introduced in 1905. The machine, operated by a two-man crew working 12-hour shifts, could produce 17,280 bottles in 24 hours, compared to 2,880 bottles made by a crew of six men and boys working in a shop for a day. The cost of making bottles by machine was 10 to 12 cents per gross compared to $1.80 per gross by the manual glassblowers and helpers.

- Perfection is reached (one time pad, Persian calendar which doesn't require leap days)

- Exploring the space of possibilities leads to overdue invention (bicycle). Another example here, which wasn't on my dataset, is luggage with wheels, invented in 1970.
- Civilization decides to solve long standing problem (sanitation after the Great Stink of London, space race)
- New chemical or physical processes are mastered (Bessemer process, activated sludge, Hall–Héroult process, polymerase chain reaction, nuclear weapons)
- Small tweak has qualitative implications. (Hale rockets: spinning makes rockets more accurate/less likely to veer).
- Change in context makes technology more viable (much easier to print European rather than Chinese characters)

  > The general assumption is that movable type did not replace block printing in places that used Chinese characters due to the expense of producing more than 200,000 individual pieces of type. Even woodblock printing was not as cost productive as simply paying a copyist to write out a book by hand if there was no intention of producing more than a few copies.

- Continuous progress encounters discrete outcomes. Military technology might increase continuously or with jumps, but sometimes we care about a discrete outcome, such as “will it defeat the British” (cryptography, rockets, radar, radio, submarines, aviation). A less bellicose example would be “will this defeat the world champion at go/chess/starcraft/poker/…” AI Impacts also mentions a discontinuity in the “time to cross the Atlantic”, and has some more stories here

More elaborate models for computing the base rate of big discontinuities.

AI Impacts states: "32% of trends we investigated saw at least one large, robust discontinuity". If I take my 12 out of 50 "big" discontinuities and assume that one third would be found to be "large and robust" by a more thorough investigation, one would expect that 4 out of the 50 technologies will display a "large and robust discontinuity" in the sense which AI Impacts takes those words to mean. However, I happen to think that the "robust" here is doing too much work filtering out discontinuities which probably existed but for which good data may not exist or be ambiguous. For example, they don't classify the fall in book prices after the European printing press as a "large and robust" discontinuity (!).

I can also compute the average time since the first mention of a technology until the first big discontinuity. This gives 1055 years, or roughly 0.001 per year, very much like AI Impact's numbers (also 0.001). But this is too high, because printing, aluminium, aviation, etc. have millenarian histories. The earliest discontinuity in my database is printing in 1450, and the next one after that the petroleum industry in 1850, which suggests that there was a period in which discontinuities were uncommon.

If we ignore printing and instead compute the average time since either the start of the Industrial Revolution, defined to be 1750, or the start of the given technology (e.g., phenomena akin to radar started to be investigated in 1887), then the average time until the first discontinuity is 88 years, i.e., roughly 0.01 per year.

Can we really take the average time until a discontinuity and translate it to a yearly probability, like 1% per year? Not without caveats; we’d also have to consider

hypotheses like whether there is a minimum wait time from the invention until a discontinuity, whether there are different regimes (e.g., in the same generation as the inventor, or one or more generations afterwards, etc.). The wait times since either 1750 or the beginning of a technology are {13, 31, 32, 47, 65, 92, 100, 136, 138, 152, 163}.

Adjustment for AI

So I have a rough prior that ~10% of technologies (4 out of 50) undergo a "large robust discontinuity", and if they do so, I'd give a ~1% chance per year, for an unconditioned ~0.1% per year. But this is a prior from which to begin, and I have more information and inside views about AI, perhaps because GPT-3 was maybe a discontinuity for language models.

With that in mind, I might adjust to something like 30% that AI will undergo a "large and robust" discontinuity, at the rate of maybe 2% per year if it does so. I'm not doing this in a principled way, but rather drawing on past forecasting experience, and I'd expect that this estimate might change substantially if I put some more thought into it. One might also argue that these probabilities would only apply while humans are the ones doing the research.

Conclusion

I have given some rough estimates of the probability that a given technology's progress will display a discontinuity. For example, I arrive at ~10% chance that a technology will display a "large and robust" discontinuity within its lifetime, and maybe at a ~1% chance of a discontinuity per year if it does so. For other operationalizations, the qualitative conclusion that discontinuities are not uncommon still holds.

One might also carry out essentially the same project but taking technologies from Computing Timelines and History of Technology, and then produce a prior based on the history of computing so far. I'd also be curious to see discussion of the probability of a discontinuity in AI in the next two to five years among forecasters, in the spirit of this AI timelines forecasting thread.

What should experienced rationalists know?

The obvious answer is 'the sequences' but imo that is neither necessary nor sufficient. The Sequences are valuable but they are quite old at this point. They also run to over a million words (though Rationality AtZ is only ~600k). Here is a list of core skills and ideas:

1 - Rationality Techniques

Ideally, an experienced rationalist would have experience with most of the CFAR manual. Anyone trying to learn this material needs to actually try the techniques; theoretical knowledge is not enough. If I had to make a shorter list of techniques I would include:

- Double Crux / Internal DC
- Five-minute Timers
- Trigger Action Plans
- Bucket Errors
- Goal/Aversion Factoring
- Gears Level Understanding
- Negative Visualisation / Murphy-Jitsu
- Focusing

2 - AI Risk: Superintelligence

The rationality community was founded to help solve AI risk. Superintelligence gives an updated and complete version of the 'classic' argument for AI-risk. Superintelligence does not make as many strong claims about takeoff as Elizer's early writings. This seems useful given that positions closer to Paul Christiano's seem to be gaining prominence. I think the 'classic' arguments are still very much worth understanding. On the other hand, Superintelligence is ~125K words and not easy reading.

I think many readers can skip the first few chapters. The core argument is in chapters five through fourteen.

5. Decisive strategic advantage

6. Cognitive superpowers

7. The superintelligent will

8. Is the default outcome doom?

9. The control problem

10. Oracles, genies, sovereigns, tools

11. Multipolar scenarios

12. Acquiring values

13. Choosing the criteria for choosing

14. The strategic picture

3 - Cognitive Biases: Thinking Fast and Slow

Priming is the first research area discussed in depth in TF&S. Priming seems to be almost entirely BS. I would suggest skipping the chapter on priming and remembering the discussion of the 'hot hand fallacy' seems incorrect. Another potential downside is the length (~175K words). However, I don't think there is a better source overall. Many of the concepts in TF&S remain fundamental. The writing is also quite good and the historical value is extremely high. Here is a quick review from 2016.

4 - Statistics

It is hard to be an informed rationalist without a basic understanding of Bayesian thinking. You need to understand frequentist statistics to evaluate a lot of relevant research. Some of the most important concepts/goals are listed below.

Bayesian Statistics:

- Illustrate the use of odd's ratio calculation in practical situations
- Derive Laplace's rule of succession

Frequentist Stats - Understand the following concepts:

- Law of large numbers
- Power, p-values, t-tests, z-tests
- Linear Regression
- Limitations of the above concepts

5 - Signalling / The Elephant in the Brain

The Elephant in the Brain is a clear and authoritative source. The ideas discussed have certainly been influential in the rationalist community. But I am not what epistemic status the community assigns to the Hanson/Simler theories around signaling. Any opinions? For reference here are the topics.

PART I Why We Hide Our Motives

- 1 Animal Behavior
- 2 Competition
- 3 Norms
- 4 Cheating
- 5 Self-Deception
- 6 Counterfeit Reasons

PART II Hidden Motives in Everyday Life

- 7 Body Language
- 8 Laughter
- 9 Conversation
- 10 Consumption
- 11 Art
- 12 Charity
- 13 Education
- 14 Medicine
- 15 Religion
- 16 Politics
- 17 Conclusion

What am I missing? Try to be as specific as possible about what exactly should be learned. Some possible topics discussed in the community include:

- Economics
- The basics of the other EA cause areas and general theory? (at least the stuff in 'Doing Good Better')
- Eliezer says to study evolutionary psychology in the eleventh virtue but I have not been impressed with evo-psych.
- Something about mental tech? Maybe mindfulness, Internal Family Systems, or circling? I am not confident anything in space fits.

Babble & Prune Thoughts

This is an accounting of various thoughts I had when reading the Babble & Prune sequence.

Encouraging Babble

It is ironic, to me, that the Babble & Prune sequence ends with a call to exclude *epistemic status* tags from posts:

> What happens to the reader when every post starts, "epistemic status: mostly true with a chance of rain?"
>
> [...]
>
> Instead of conditioning readers to hate us, I propose we return to a saner time, where the fact that your words are but a pale wavering shadow of the grand, mysterious truth in your heart is *the default assumption* about human communication. Where truth is a dance of successive approximations yet no step in that dance requires adult supervision. Where quibbles over certitude are banished to the comments section where they belong.

I interpret Alkjash as thinking that, if the standards were just lower in the first place, people would feel free to write more, which would get us further in the end (because the good stuff can bubble to the top).

Indeed, my impression is that the LessWrong team has worked to make LessWrong a space where people feel they can share raw thoughts rather than requiring everything to be carefully refined.

But in my opinion, the "epistemic status" flags, and similar tools, *help rather than hurt.*

The New Twitter Account Problem

There is a phenomenon -- let me know if you think of a better name -- which I call the new twitter account problem. My personal interaction with Twitter has been as follows:

1. Start a new twitter account. Follow a few people I like. Basically no followers. No pressure! *I can say whatever I want!* Dump some random thoughts into the new twitter account over the next few days.
2. Now I'm getting follows, likes, and shares. Positive reinforcement! This encourages me to keep it up.
3. Follows slowly climb, and I'm getting more and more interaction. Now I feel like there's something at stake, and I have to do my best. I write some good tweets, but I gradually tweet less and less.
4. I make an alt account for "worse thoughts", starting the cycle back at step 1. Gradually the alt account becomes my main account, and I have the same problem again, and need a third account, and a fourth... Meanwhile, the followers of my previous accounts wonder why I'm not active anymore.

LessWrong 1.0 had a similar problem. The frontpage got "too good" for people to feel like they could really post on. The "discussion post" was invented. A lot of activity moved to discussion. But then *Discussion* got too good, and people felt like they had to have something good even to post it in Discussion. Open Threads were created *within* Discussion, as the new low-bar-to-entry forum.

I've heard that a similar pattern has also played out on other discussion platforms. Increasing layers of "no really, it's OK to post raw thoughts" are created as old layers get too respectable.

This is probably crazy, but...

In in-person discussions, a good way to lessen this problem is to preface statements with qualifiers like "This is a dumb question, but..." or "Here's a crazy idea:" or "I don't endorse this, but I was thinking..." and so on. This creates a kind of protection for the speaker. They still get credit for good ideas; but if the idea *really was* bad, they take less of a hit.

This doesn't work *as* well for longform text discussions, because readers know you had time to think things through. But *[epistemic status:]* flags can play a similar role. If you don't feel like an idea is good enough, you preface it with [epistemic status: super dumb] or whatever, and take some comfort in the fact that if the reader doesn't like what you wrote, *they were warned.*

I think having this norm is much better than attempting to sustain a norm that everything is "epistemic status: raw thoughts" by default.

Say Random Things

OK, so epistemic status flags are one way to combat the new-twitter-account problem. Do we have any other tools?

One tool is to purposefully lower your standards. I believe the book IMPRO includes an exercise in which you point to random things around you and call them absolutely wrong things; (point to tree) "there's a lamp-post" (point to grass) "there's some spaghetti" etc. Perhaps saying absurd things on purpose helps prove to your s1 that nothing horrible will happen if you say something wrong. Another example of this is Allie Brosh drunk posting (notable because the explains the thought process behind it). This is commonly called shitposting. Unfortunately, Allie Brosh wrote *less* after that experiment, so it's not clear that it had a positive impact. Anecdotally, I've heard that a friend had dramatically positive results with the IMPRO version.

Master All Levels

I don't think we should just work on improving our babble, though. I think it's really important to aspire to higher and higher standards. I want to become stronger! How can we reconcile this with the dampening effect high standards can create?

Meta-Perfectionism

The true answer is that if "holding yourself to a high standard" makes you do/say too little, then your concept of "standard" is broken. We intuitively reason based on blame/guilt, which makes improper inaction seem less bad than improper action. If we could free ourselves of that mode of reasoning, perhaps we could *just not have* the new-twitter-account problem in the first place.

In The best you can, Nate Soares writes:

> It's easy to paralyze yourself if you try to do the "right thing." There's always more uncertainty to be had. There's always more information you could gather. It's hard to become confident that you're doing the right thing. This can lead to paralysis, and persistent inaction.
>
> It's much easier, I think, to stop asking "is this action the right action to take?" and instead ask "what's the best action I can identify at the moment?"
>
> Sometimes, the best action you can identify is "search for more alternatives." Sometimes, it's "study more" or "learn more." Sometimes, it's a specific action. The nice thing is that "what's the best action I can find in the next five minutes?" always has a concrete answer. If you search for that, instead, you won't get paralyzed.

He elaborates more on similar thoughts in Deliberate once. But more relevantly to Babble&Prune, his post Half-assing it with everything you've got describes the mindset required to orient perfectionism at the meta-problem of avoiding overmuch perfectionism.

Maximum Payoff Per Effort

We face two problems:

1. Investing the minimal amount of time/effort to get the results we want.
2. Getting the best results we can given the time/effort.

Nate focuses on the first problem, describing how some situations call for a measured effort, while other situations call for an all-out effort. But we also want to do the second: the more favorable our payoff per time/effort spent, the less we need to spend on things calling for a measured effort, and the more results we get when we go all-out.

One thing university art classes do is force you to spend a long time on a single drawing. This teaches you *how good you can get* if you spend a lot of time on a single piece. It also teaches you *how* to fruitfully spend that much time on a piece.

But to grow as an artist, you *also* need to learn speed. You shouldn't carefully plan every drawing you do. Artists also practice gesture drawings, in which you have a *very* short amount of time to capture the general pose of a model. So, a deliberate practice of art involves practicing at all time-scales.

There's probably a natural slow-to-fast progression for many skills, where you need to go slow at first to be able to do it well, but can do it faster and faster after that.

Training Babble, Training Prune

Your *prune* should not just accept/reject. It should have degrees. You should be able to step it up or down appropriately. Furthermore, you want your *prune* to give you specific feedback -- not just "that's bad", but "that's mismatched", etc. I suggest listening to your internal critic with a felt sense / focusing lens. You may find that you know more than you thought about what makes for good work.

You want your *babble* to be able to conform to the highest standards it can while remaining creative. You want it to be like GPT-3, not entirely random words. So pay attention to what your taste says about your babble -- the trick isn't to cast your standards aside, but rather to pay attention to them without letting them *stop* babble. Give your babble a little breathing room from your prune.

Alkjash mentioned that babble is not generating things with independent randomness, but rather, is more like a random walk in concept-space. A lot of the skill of good babble is in generating good *mutations* of ideas, not just good ideas. You can see this as you practice babble on longer time-scales. Give yourself more time to write a sentence, and you may see yourself go through several mutations of that sentence before settling on one to write. Part of what it means to spend more time on a drawing or painting is giving yourself more time to plan each line, and more time to fiddle with it, seeking improvements.

My main point here is that improving babble doesn't mean reducing prune. Alkjash sometimes speaks as if it's just a matter of opening the floodgates. Sometimes people *do* need to just relax, turn off their prune, and open the floodgates. But if you try to do this in general, you might have initial success but then experience backlash, since you may have failed to address the underlying reasons why you had closed the gates to begin with.

The Many Gates

I think one of the most useful models in Babble & Prune was the three gates.

Actually, after starting this section I remembered that I already wrote my thoughts on this in the Babble & Prune section of Capturing Ideas, particularly the part about developing ideas. Go read that if you want a further elaboration of why just opening the floodgates isn't exactly the goal.

Weird Things About Money

1. Money wants to be linear, but wants even more to be logarithmic.

- People sometimes talk as if risk-aversion (or risk-loving) is irrational *in itself*. It is true that VNM-rationality implies you just take expected values, and hence, don't penalize variance or any such thing. However, *you are allowed to have a concave utility function,* such as utility which is logarithmic in money. This creates risk-averse behavior. (You could also have a convex utility function, creating risk-seeking behavior.)
    - ***Counterpoint***: if you have risk-averse behavior, other agents can exploit you by selling you insurance. Hence, money flows from risk-averse agents to less risk-averse agents. Similarly, risk-seeking agents can be exploited by charging them for participating in gambles. From this, one might think a market will evolve away from risk aversion(/seeking), as risk-neutral agents accumulate money.
- People clearly act more like money has diminishing utility, rather than linear utility. So revealed preferences would appear to favor risk-aversion. Furthermore, it's clear that the amount of pleasure one person can get per dollar diminishes as we give that person more and more money.
    - On the other hand, that being the case, we can easily purchase a lot of pleasure by giving money to others with less. So from a more altruistic perspective, utility does not diminish nearly so rapidly.
- Rationality arguments of the Dutch-book and money-pump variety require an assumption that "money" exists. This "money" acts very much like utility, suggesting that utility is supposed to be linear in money. Dutch-book arguments assume from the start that agents are willing to make bets if the expected value of those bets is nonnegative. Money-pump arguments, on the other hand, can *establish* this from other assumptions.
    - Stuart Armstrong summarizes the money-pump arguments in favor of applying the VNM axioms directly to real money. This would imply risk-neutrality and utility linear in money.
- On the other hand, the Kelly criterion implies betting as if utility were *logarithmic* in money.
    - The Kelly criterion is not derived via Bayesian rationality, but rather, an asymptotic argument about average-case performance (which is kinda frequentist). So initially it seems this is no contradiction.
    - However, it is a theorem that a diverse market would come to be dominated by Kelly bettors, as Kelly betting maximizes long-term growth rate. This means the previous counterpoint was wrong: expected-money bettors profit *in expectation* from selling insurance to Kelly bettors, but the Kelly bettors eventually dominate the market.
    - Expected-money bettors continue to have the most money *in expectation*, but this high expectation comes from increasingly improbable strings of wins. So you might see an expected-money bettor initially get a lot of money from a string of luck, but eventually burn out.
        - (For example, suppose an investment opportunity triples money 50% of the time, and loses it all the other 50% of the time. An expected

money bettor will go all-in, while a Kelly bettor will invest some money but hold some aside. The expected-money betting strategy has the highest expected value, but will almost surely be out in a few rounds.)

- The kelly criterion still implies near-linearity for small quantities of money.
    - Moreover, the more money you have, the closer to linearity -- so the larger the quantity of money you'll treat as an expected-money-maximizer would.
    - This vindicates, to a limited extent, the idea that a market will approach linearity -- Kelly bettors will act more and more like expected-money maximizers as they accumulate money.
    - As argued before, we get agents with a large bankroll (and so, with behavior closer to linear) selling insurance to Kelly agents with smaller bankroll (and hence more risk-averse), and profiting from doing so.
    - But everyone is still Kelly in this picture, making logarithmic utility the correct view.
- So the money-pump arguments seem to *almost* pin us down to maximum-expectation reasoning about money, but *actually* leave enough wiggle room for logarithmic value.
- If money-pump arguments for expectation-maximization doesn't apply in practice *to money,* why should we expect it to apply elsewhere?
    - Kelly betting is fully compatible with expected utility maximization, since we can maximize the expectation of the logarithm of money. But if the money-pump arguments are our reason for buying into the expectation-maximization picture in the first place, then their failure to apply to money should make us ask: why would they apply to utility any better?
    - **Candidate answer:** utility is *defined as* the quantity those arguments work for. Kelly-betting preferences on money don't actually violate any of the VNM axioms. Because the VNM axioms hold, we can re-scale money to get utility. That's what the VNM axioms give us.
        - The VNM axioms only rule out *extreme* risk-aversion or risk-seeking where a gamble between A and B is outside of the range of values from A to B. Risk aversion is just fine if we can understand it as a re-scaling.
        - So any kind of re-scaled expectation maximization, such as maximization of the log, should be seen as a *success* of VNM-like reasoning, not a failure.
        - Furthermore, thanks to continuity, any such re-scaling will closely resemble linear expectation maximization when small quantities are involved. Any convex (risk-averse) re-scaling will resemble linear expectation more as the background numbers (to which we compare gains and losses) become larger.
        - It still seems important to note again, however, that the usual justification for Kelly betting is "not very Bayesian" (very different from subjective preference theories such as VNM, and heavily reliant on long-run frequency arguments).

2. Money wants to go negative, but can't.

- Money can't go negative. Well, it can, just a little: we do have a concept of debt. But if the economy were a computer program, debt would seem like a big hack.

There's no absolute guarantee that debt can be collected. There are a lot of incentives in place to help ensure debt can be collected, but ultimately, bankruptcy or death or disappearance can make a debt uncollectible. This means money is in this weird place where we sort of act like it can go negative for a lot of purposes, but it also sort of can't.
    - This is especially weird if we think of money *as* debt, as is the case for gold-standard currencies and similar: money is an IOU issued by the government, which can be repaid upon request.
    - Any kind of money is ultimately based on some kind of *trust.* This can include trust in financial institutions, trust that gold will still be desirable later, trust in cryptographic algorithms, and so on. But thinking about debt emphasizes that a lot of this trust is *trust in people*.
- Money can have a scarcity problem.
    - This is one of the weirdest things about money. You might expect that if there were "too little money" the value of money would simply re-adjust, so long as you can subdivide further and the vast majority of people have a nonzero amount. But this is not the case. We can be in a situation where "no one has enough money" -- the great depression was a time when there were too few jobs and too much work left undone. Not enough money to buy the essentials. Too many essentials left unsold. No savings to turn into loans. No loans to create new businesses. And all this, *not* because of any change in the underlying physical resources. Seemingly, economics itself broke down: the supply was there, the demand was there, but the supply and demand curves could not meet.
        - (I am not really trained in economics, nor a historian, so my summary of the great depression could be mistaken or misleading.)
    - My loose understanding of monetary policy suggests that scarcity is a concern even in normal times.
- The scarcity problem would not exist if money could be reliably manufactured through debt.
    - I'm not really sure of this statement.
    - When I visualize a scarcity of money, it's like there's both work needing done and people needing work, but there's not enough money to pay them. Easy manufacturing of money through debt should allow people to pay other people to do work.
    - OTOH, if it's *too* easy to go negative, then the concept of money doesn't make sense any more: spending money doesn't decrease your buying power any more if you can just keep going into debt. So everyone should just spend like crazy.
        - Note that this isn't a problem in theoretical settings where money is equated with utility (IE, when we assume utility is linear in money), because *money is being inherently valued* in those settings, rather than valued instrumentally for what it can get. This assumption is a convenient approximation, but we can see here that it radically falls apart for questions of negative bankroll -- it seems easy to handle (infinitely) negative money if we act like it has intrinsic (terminal) value, but it all falls apart if we see its value as extrinsic (instrumental).
    - So it seems like we want to facilitate negative bank accounts "as much as possible, but not too much"?
- Note that Dutch-book and money-pump arguments tend to implicitly assume an infinite bankroll, ie, money which can go negative as much as it wants. Otherwise you don't know whether the agent has enough to participate in the proposed transaction.

- Kelly betting, on the other hand, assumes a finite bankroll -- and indeed, might have to be abandoned or adjusted to handle negative money.
- I believe many mechanism-design ideas also rely on an infinite bankroll.

Desperation hamster wheels

In my first few jobs, I felt desperate to have an impact. I was often filled with anxiety that I might not be able to. My soul ached. Fear propelled me to action. I remember sitting in a coffee shop one Saturday trying to read a book that I thought would help me learn about biosafety, an impactful career path I wanted to explore. While I found the book interesting, I had to force myself to read each page because I was worn out. Yet I kept chugging along because I thought it was my lifeline, even though I was making extremely little progress. I thought: *If I don't do this excellently, I'll be a failure.*

There were three critical factors that, taken together, formed a "desperation hamster wheel," a cycle of desperation, inadequacy, and burn out that got me nowhere:

- Self-worth -- I often acted and felt as if my self-worth was defined wholly by my impact, even though I would give lip service to self-worth being more than that.
- Insecurity/inadequacy -- I constantly felt not skilled or talented enough to have an impact in the ways I thought were most valuable.
- Black and white thinking -- I thought of things in binary. E.g. I was either good enough or not, I was smart or not, I would have an impact or not.

Together, these factors manifested as a deep, powerful, clawing desire for impact. They drove me to work as hard as possible, and fight with all I had. It backfired.

This "desperation hamster wheel" led me to think too narrowly about what opportunities were available for impact and what skills I had or could learn. For example, I only thought about having an impact via the organization I was currently working at, instead of looking more broadly. I only considered the roles most lauded in my community at the time, instead of thinking outside the box about the best fit for me.

I would have been much happier and much more impactful had I taken a more open, relaxed, and creative approach.

Instead, I kept fighting against my weaknesses -- against reality -- rather than leaning into my strengths.(1) It led me to worse work habits and worse performance, creating a vicious cycle, as negative feedback and lack of success fueled my desperation. For example, I kept trying to do research because I thought that that work was especially valuable. But, I hadn't yet developed the skills necessary to do it well, and my desperation made the inherent vulnerability and failure involved in learning feel like a deadly threat. Every mistake felt like a severe proclamation against my ability to have an impact.

I'm doing a lot better now and don't feel this desperation anymore. Now, I can lean into my strengths and build on my weaknesses without having my whole self-worth on the line. I can approach the questions of having an impact with my career openly and with curiosity, which has led me to a uniquely well-suited role. I can try things and make mistakes, learning from those experiences, and becoming better.

I feel unsure about what helped me change. Here are some guesses, in no particular order:

- Taking anxiety and depression medication

- Changing roles to find one that played more to my strengths
- I'm sad that I'm not better or smarter than I grew up hoping I might be. It took time to grieve that and come to terms with it on both a personal and impact level (how many people could I have helped?)(2)
- Digging into the slow living movement and trying to live more intentionally and with more reflection
- Having a partner who loves me and who doesn't care about the impact I have, except so far as he knows I'd like to have an impact
- Reconnecting with my younger, child self, who felt curious and excited. In the worst moments on my desperation hamster wheel, I no longer felt those things. I found it really helpful to reconnect with that part of myself via rereading some of my childhood's foundational books, prompting myself to play, and boggling at or feeling wondrous curiosity about nature/basic facts about the world.
- Making grounding/centering activities part of my daily routine. For me, these are: yoga, journaling, walks, bike rides, showers, meditation, and deep breaths.
- Learning the practice of Focusing -- learning how to be with myself without judgment and how to hear different parts of myself

I made a ton of progress using these strategies, and then two things happened, which solidified a core antidote to my desperation mindset.

The first is that my mom died and her death shined a spotlight on some similar struggles. I don't think she would have described herself as being on a "desperation hamster wheel," but I know she struggled a lot with self-worth and insecurities throughout her life. Cliche but true: none of that mattered in the end. If she could have had one more month, one more year, I would have wanted her to spend time with her loved ones and lean strongly into her passion (painting). That she didn't read books much or that she didn't get into a painting show that one time doesn't matter at all. Her insecurities and self-worth issues were utterly unimportant in the end; they were beside the point. Mine probably were too.

The second is that I got a concussion and couldn't work or even look at screens for a couple of months. This reduction made me understand something on an intuitive level ("system 1") that I hadn't before, or that I had lost sight of: *that I am a person whose life has value outside of my potential impact*. My life was still valuable by my own evaluation, even without work. I ate, slept, gardened, cooked, cuddled with my dogs, and called some friends. I was still a full person, a person of value, even though I wasn't working. It sounds obvious, but it had been so long since I had been fully separated from my working self. I had forgotten that there's a core, a *ME*, that's always there, and has value in and of itself.

My current hypothesis is that if you're stuck on a desperation hamster wheel, you'll have a lot more impact once you get off of it.(4) You'll also have a better life, but if you're on the desperation hamster wheel right now, you might not weigh that piece that seriously. (I think that's a mistake, but that's not necessary to get into at this time.) (5)

Being on the hamster wheel is indicative of being stuck in suboptimal patterns, burning yourself out, and a narrowing of thought that is counterproductive to most knowledge work. If you allow yourself to get off the wheel, you'll be able to think and plan from a more grounded, creative place. New opportunities and ideas will emerge. You'll have more energy. You'll be able to see that you have strengths. This was true for me and I've seen it be true for others as well. Of course, I might be wrong.

Everyone is different and life is complicated. But, if you care a lot about impact, it seems worth taking this hypothesis seriously and testing it out.

If you'd like to try stepping off the hamster wheel, I'm not sure what will work for you -- everyone is different. But, here are some ideas to get started. Also feel free to reach out to me. I'm happy to brainstorm with you.

- Take time off and away from the things that feed your hamster wheel patterns
- Examine whether you have any underlying mental health issues which could be playing into this dynamic in an unhelpful way. If yes, try out some treatments and consult an expert (e.g. a psychiatrist).
- Practice noticing what enables you to feel grounded, in the present moment, or in your body. Try out different things to see what works for you. Then lean into things that seem to help you.
- Spend time reflecting and emotionally reconnecting with what excited or interested you as a child.
- Bolster your social and emotional support relationships.
- Find a counterbalance to some of the drivers of your hamster wheel. For me, I found the slow living movement to be very helpful; it prompts intentionality and calmness, and provides a separate frame and pushes usefully against some of my hamster wheel tendencies.
- Explore your options with creativity and openness once you're a better emotional place.

Dear Nicole-from-a-few-years-ago,

The takeaway messages here that I would love for you to internalize are:

- *Get off the desperation hamster wheel. It sucks to be on it and, anyway, you're never going to have much impact if you stay on it. In order to have a long-term, personally sustainable impact, you need to find and create a situation that promotes your thriving. Keep in mind that impact is measured over the course of your whole career, not just today. As others have said, altruism is a marathon, not a sprint. (6)*
- *Don't get distracted or emotionally tied up in what's cool. That's irrelevant to what you need right now to set yourself up for an impactful career.*
- *You do have strengths, even if you can't see them right now. And when you lean into your strengths, you'll be able to grow and make progress on your weaknesses. It's much easier to grow from a solid foundation. Relax and take care of yourself. That is the first step in order to give yourself a solid foundation.*
- *Really, get off that fucking desperation hamster wheel. I promise you, it's not helping. Your martyr tendencies are misguiding you here -- they are correctly tracking that you care deeply about doing good, but the way they are pushing you to pursue this goal is backfiring.*

Thanks to Eric, Rebecca, Neel, and Duncan for helpful comments, and a bunch of others for support and encouragement.

Endnotes:

(1) My partner, Eric, notes that this reminds him about how humans learned to fly but did so by inventing airplanes, not turning themselves into birds. I don't know how to

work this in smoothly, but the comparison resonated with me a lot, so I'm putting it in this footnote.

(2) Blog post on this topic forthcoming

(3) Here is a podcast that explores slow living, and which I've found helpful.

(4) Impact, or whatever value/goal that got you on the desperation hamster wheel in the first place.

(5) Blog post on this topic coming at some vague point (i.e., I have an idea, but not a rough draft like I do for the others)

(6) Blog post on this topic forthcoming

Babble challenge: 50 ways of sending something to the moon

This is an exercise, and as such is a bit different from your ordinary question post...

What?

Come up with 50 ways of sending something to the moon. In less than 1 hour.

I don't care how stupid they are. My own list included "Slingshot", "Massive trampoline" and "Bird with spacesuit".

What matters is that you *actually hit 50*. I want you to have the experience of thinking that you're out of ideas, but nonetheless deciding to push yourself, and *finding your best idea thus far*.

This *regularly* happens to me when I do this exercise. I'll feel stuck. I'll feel like giving up. But I force myself to say three more stupid things... "mega tall tree", "super boomerang", "railgun" ... and, all of sudden, I have a fourth idea that's actually not that shabby.

Why do this?

1) Becoming more creative.

Coming up with ideas is a bottleneck for me personally. I want to become stronger.

I have a very simple model for how to improve. My brain will start generating more ideas if I A) force myself to have ideas, *even if they're bad*, and B) reward myself for having them.

The act of filtering for actually good ideas is a second, different step. First you babble. And only then you prune. I claim that you can train each of those "muscles" separately.

I think that in the past my creativity has been held back by excessive self-criticism. I now think it should be possible to improve by separating the creative and the evaluative step -- at least for practice purposes.

2) Building a culture of practice on LessWrong

LessWrong currently feels like an unusually intellectual bar. You pop in and grab a drink; instead of watching a stand-up comedian someone does a PowerPoint presentation; and the ensuing conversation is great. That's a fine thing.

But what if part of LessWrong was more like a gym? Or a dojo?

You come in, bow to the sensei, and then you start practicing. Together. Each of you focusing all your attention on your movements, pushing yourselves to your limits, and

trying deliberately to become stronger.

I want us to have that, but for rationality.

Rules

- **50 answers or nothing.**

That's the babble challenge. We're here to work hard.

- **Post your answers inside of spoiler tags!** (How do I do that?)
- **Celebrate other's answers.**

This is really important. Sharing babble in public is a scary experience. I don't want people to leave this having back-chained the experience "If I am creative, people will look down on me". So be generous with those upvotes.

If you comment on someone else's post, focus on making exciting, novel ideas work -- instead of tearing apart worse ideas.

Reward people for babbling -- don't punish them for not pruning.

I might remove comments that break this rule.

- **Not all your ideas have to work.**

Man, *I barely know anything about going to the moon*. Yet I did come up with 50 ideas.

"Bird with spacesuit" is fine. I have some intuition where making these ideas actually helps me become creative. I try to track that intuition.

- **My main tip: when you're stuck, say something stupid.**

If you spend 5 min agonising over not having anything to say, you're doing it wrong. You're being too critical. Just lower your standards and say something, anything. Soon enough you'll be back on track.

This is really, really important. It's the only way I'm able to complete these exercises (and I've done a few of them in the last few days).

--

Now, go forth and babble! 50 ways of sending something to the moon!

Is Stupidity Expanding? Some Hypotheses.

To be explained: **It feels to me that in recent years, people have gotten stupider, or that stupid has gotten bigger, or that the parts of people that were always stupid have gotten louder, or something like that.**

I've come up with a suite of hypotheses to explain this (with a little help from my friends). I thought I'd throw them out here to see which ones the wise crowd here think are most likely. Bonus points if you come up with some new ones. Gold stars if you can rule some out based on existing data or can propose tests by which they might be rendered more or less plausible.

The hypotheses come in two broad families: 1) my feeling that stupid is expanding is an illusion or misperception, and 2) stupid is expanding and here is why:

A: I Am Misperceiving an Expanding Stupidity And Here's Why

1. I have become more attuned to stupidity for [reasons], so even though there is no more of it than usual, it stands out more to me.[1]
2. What used to look like non-stupidity was actually widespread conformity to a common menu of foolishnesses. Today the cultural beacons of respectable idiocy have been overthrown and there is increasing diversity in foolishness. Divergent fools seem more foolish to each other when in fact we're all just as stupid as we've always been.
3. I'm running in stupider circles than I used to for some reason, while in general things haven't changed much.
4. I am the one getting stupider, or was stupid all along, and so I don't have the cognitive strength to accurately judge the stupidity level around me, and just happen to be thinking it is getting worse because I don't know any better.[2]
5. People aren't getting any stupider, it's just that the artificial intelligence of the bots I'm mistaking for people on-line isn't all that good yet.
6. They're not getting stupider; I'm just getting more conceited.
7. People ordinarily use different modes of thinking in different communications contexts. In some, finding the truth is important and so they exhibit rational intelligence. In others, decorative display, ritual, asserting dominance or submission, displaying group allegiances, etc. are more important and so they use modes more appropriate to those things. It's not that people are getting stupider, but that these forms of communication that do not broadcast intelligence (a) are more amplified than they used to be, (b) more commonly practiced than they used to be, or (c) are more prominent where I happen to be training my attention.
8. I am acquiring greater wisdom with age as I ought, but the average age of the typical person I encounter stays the same so they cannot keep up. I'm noticing the contrast increasing but misattributing it.[3]
9. People use intelligence for different things in different eras. Just as language, music, or art changes over time, so does thinking. I'm just not keeping up, and

assuming because kids these days can't dance the mental Charleston that they can't dance at all.
10. We were just as stupid back in the day, and I just don't remember it that way.[4]
11. There is no truth, only power. What I've been interpreting as truth and rationality has been my own attempt to align my thinking with the political clique that was in power when I was being educated. What I'm interpreting as rising stupidity has been the collapse in power and status of that clique and the political obsolescence of the variety of "truth" and "rationality" I internalized as a child. Those pomo philosophers were right all along.
12. Stupidity doesn't have staying power, relative to non-stupidity: there's a sort of survival of the fittest in which vast amounts of expressions are being produced all the time, most of which are stupid and fall away, but the ones that aren't stupid are more likely to survive in memory and to be maintained in the historical record. This biases things to make it appear that the proportion of stupid expressions was lower in the past than it really was.
13. Politics and consumer capitalism are motivated to identify and target stupid people so as to take advantage of them, so they have created systems that encourage stupid people to self-identify and make themselves prominent so that they can be picked off; that I'm noticing this is just a side effect.

B: Expanding Stupidity Is Real and This Explains It

1. People have given up trying to understand things in this messed-up timeline and are just rolling with it; it's a sort of intellectual learned helplessness that appears as expanding stupidity.
2. Stupidity has its fashions, and the latest fashions are more in-your-face than they used to be.
3. Pharmaceuticals that have become popular in recent decades have cognitive side effects that are difficult to measure in the individual but cause noticeable effects in the aggregate.
4. It's real, and it's probably something in our diet, for example…
5. It's real, and it's probably all that extra CO_2 in the atmosphere.
6. It's real, and it's probably toxoplasmosis meow.
7. It's real, and it's probably some other sort of change in our material environment (excluding cultural changes).[5]
8. Back in the day, when a person had a stupid idea, they would be reluctant to put it forward as their own. Rather, they would wait to see if someone else would voice the idea so they could just agree with it. This used to be relatively rare, but now you just have to google "[my stupid idea]" to find that someone or other has said it first, and then you're off to the races.
9. If you have a smart idea, you may also be smart enough to realize that it's not useful right now / has already been better said / is inappropriate in context. If you have a dumb idea, such thoughts may be less likely to occur to you due to the aura of dumbth that surrounds the dumb idea and repels sensible considerations. Back when expressions of stupidity were mostly ephemeral, this didn't matter much, but now that they acquire instant permanence and global reach, they appear to swamp everything else.
10. Stupid choices used to reliably have undesirable results; now there is more of a disconnect where people are shielded from the results of their stupid choices, or even rewarded for them (man lights himself on fire in an easily-foreseeable

misadventure, becomes YouTube legend). So people may be appearing stupid not as a result of being stupid but as the result of a perverse cost-benefit analysis. People are no dumber than they used to be, but for [reasons] it has become advantageous to display stupidity and so smart people sometimes mimic idiocy so as to reap such advantages. The smarter they are, the quicker they caught on to this and the better mimics they are, so this makes it look as though the smart people are being replaced by morons, when really it's more a matter of camouflage.

11. The way we educate children went seriously sideways a while back, and so, yeah, stupid happened.
12. Newly-popular media and/or its content is somehow directly damaging to mental faculties.
13. Changes in media/communications technology allow stupid people to be much more prominent than they used to be and/or comparatively muffle smarter people.
14. Social media dynamics erode reasoning and truth-seeking while amplifying cognitive biases.
15. The news media were doing a better job than we realized in filtering out crap and contextualizing new information intelligently for us, and as the internet destroyed the business model behind intelligent reporting, we failed to come up with a substitute in time to prevent idiocy from filling the void and it's too big a job for individuals to do without institutional assistance.

1. ^

   Baader-Meinhof phenomenon

2. ^

   Dunning-Kruger effect perhaps

3. ^

   David Wooderson effect

4. ^

   Rosy retrospection

5. ^

   e.g. lead, maybe? "Half of US adults exposed to harmful lead levels as kids" (*AP* summary of *PNAS* paper). The researchers "find that lead is responsible for the loss of 824,097,690 IQ points as of 2015." (I've never seen this millions-of-missing-IQ-points sort of population-wide metric before, but it is an impressively large and precise number.)

How to Find the Frontiers of Knowledge

This article is based on my lecture at Topos House, San Francisco on February 29, 2020. It was originally published on SamoBurja.com. You can access the original here.

Not all fields of knowledge exist yet. If you tried to study biochemistry in 1820, you'd have a lot of trouble: the field had yet to cohere. Do we think that all the biochemistries of the world have been discovered? If we did back then, we'd be wrong. For those seeking a safe career, sticking to the established fields is probably the right move. But for those interested in pushing forward the frontiers of knowledge, it will sometimes be better to work in a field that has not yet cohered, or in a field on the cusp of crystallizing. Often the best intellectual opportunities—if you want to be a breakthrough researcher—and the largest economic opportunities—if you want to build a great company—are going to be precisely in those areas about to crystallize into a new field.

If you went into biochemistry when biochemistry was emerging, your name might be in a textbook today. You didn't have to be brilliant; you just had to pick your problem well. The greatest difficulty would be in figuring out who to learn from. Before you figure out where and how to learn, you have to decide who has the knowledge you seek. When the field of biochemistry was on the cusp of crystallizing, for example, you would have looked to experts in the fields of biology and chemistry. But as an outsider, who are you to evaluate the quality of a field? Just because a field claims to exist, doesn't mean it exists.

Evaluating Existing Fields

One thing you can do is look at how a field performs on its own criteria. Take the replication crisis in academic psychology for example. In academic science, replicability has been one of the gold standards both for internal bureaucratic targets and also for the layman's

understanding of the philosophy of science. Philosophy of science might not seem very important, but it is key to figuring out how we know what we know, and what is good science as opposed to bad science. There is, in fact, no consensus philosophy of science, which means that there is no canonical science of science.

But surely scientists must know how science works? Well, scientists might know how science works in the way that birds know how aerodynamics works. They know it, but not at all on a conscious level—they just fly. Even if birds could speak, their answers might be quite useless, even to baby birds. Their knowledge is not formatted for scientific understanding; it remains locked away as a type of intellectual dark matter. If you are deciding whether to enter a field in which you are not yet an expert, you need a more precise epistemic foundation than a bird's intuition, especially when deciding where expertise truly resides.

A thought experiment: say you had five experimental planes in front of you. How do you decide which one to board? Is it the wooden one, the bamboo one, the steel one, the large one with smoke-puffing engines, or the modest one made by the bicycle shop owners? The Wright brothers were bicycle shop owners, and they built a janky-looking machine. Plenty of the other early flying machines looked vastly more impressive than theirs, but they didn't fly. In the case of deciding which of these early flying machines to board, relying on institutional claims to epistemic authority would not work—not even the ones made by Harvard professors could fly.

When making such decisions, you cannot assume that the members of prestigious institutions of your society are experts simply because they claim to be experts. They would claim to be experts whether they were or weren't! In any period of human history, if you examine how institutions portray their own expertise, they always portray themselves as tremendously knowledgeable with impeccable foundations. In the rare cases where they do admit to not knowing something they propose it to be either unknowable or, to be knowable—but only if you give them more funding. The Catholic Church would claim this when it came to metaphysics and the question of salvation. Today, the Church of Scientology claims that Scientology is at the productive frontier of psychology, that they've disproven psychiatry, and that they have these cutting-edge, electronic brain-measuring devices. And so would a university cognitive science department.

We all make choices using dumb heuristics, but importantly, they are often good enough for everyday life. If you choose to study history at Oxford because the buildings look old, you'll likely do pretty well. Rules of thumb such as "Are the buildings old?" can have a valid core. The heuristic isn't a bad one for the prospective historian to use: old buildings often come with libraries well-stocked with old books. But does this heuristic hold in other cases?

Say you are deciding between Harvard and MIT for studying astrophysics, and you view Harvard as the better pick. Why do you believe this? Is it because Harvard has older-looking buildings? We can imagine that some prospective astrophysicists, too, are swayed by Harvard's historic appearance. Much of academia, in fact, functions on such cargo cult heuristics. Maybe one *should* study astrophysics at Harvard, rather than MIT. But if this is so, then it's right by accident: the "old buildings" heuristic doesn't apply to astrophysics. It is thus a mistake to rely on it—a broken clock is right for two moments a day, and wrong in all others. Your life has many important moments.

Because there is no consensus philosophy of science—no science of science—if you want to go into a more established field, you have to rely either on institutions, or your own evaluation of individual researchers. What you must do, then, is form your own judgments of the claims to intellectual authority made by particular institutions, or the quality of thinking of possibly exceptional individuals. But what if you want to go into a less established field?

Communities of Practice

For those considering entering a new field, there are several ways to acquire deep expertise. Firstly, there is the community of practice—imagine a hobbyist society or hacktivist collective, for example. People gravitate towards these communities chiefly for friendship and community, and proceed to enthuse together over their community's mechanism of practice, channeling this shared energy into competition for status, acceptance, and love. Thus, communities of practice tend to have a social pressure towards excellence. This often makes joining one one of the best ways to acquire knowledge.

History is full of examples of communities of practice, from the circles of philosophers of Ancient Greece, to the guilds of medieval Europe, to Meetup.com. The Royal Society was originally a group of bored 17th century British aristocrats who wanted to stay far away from politics in the aftermath of the English Civil war to pursue knowledge of nature together. One can imagine them showing off their fanciful etchings and astronomical instruments, or one-upping each other with exotic mineral samples brought from far-flung corners of the globe. A side effect of this competition was science: we learned things about geology, astronomy, entomology, and so on. Thanks to this, Charles Darwin had access to a dataset describing the taxonomy and habitats of insects and plants from all over the world, which provided much of the evidence for his theories of natural selection. Were that prior work not done, Darwin's theories would have been mere conjecture, difficult to establish as authoritative contributions to our understanding of the mechanisms of nature.

So seek out communities of practice; find out who is excited by what you want to learn. This community may or may not be very well-connected, but above all it should be relatively narrowly focused on the practice of some activity that its members constantly relate to. Without this focus, the community won't have highly-trained internal heuristics—better heuristics than the age of buildings!—to identify who really knows what they're talking about.

Communities of practice are not perfect, but you can almost always rely on them in some way. Some of them may be fraudulent or confused, but you're certainly better off joining one than attempting to be a pure individualist. There are likely entire communities out there centered on what you wish to study. Find them. Think about who would have the socioeconomic leisure to run such a community, and what they'd call themselves. Maybe you will find them at a university among college students, maybe on the internet. Perhaps you will find them among the modern equivalent of bored aristocrats—angel investors, perhaps. Smart, bored rich people will always find a hobby. It's only a question of what that hobby is—is it what you're searching for?

Master-Apprentice Relationships

Aside from communities of practice, another indispensable mechanism of transferring deep expertise is the master-apprentice relationship. In medieval European guilds, the master-apprentice relationship formed a strong contract. Instead of owing $90,000 of student debt and receiving a diploma, you instead would sign yourself up for service to a particular master for around seven or eight years. And at the end, they would essentially grant you their business. Their retirement plan would be *you*. Can you imagine how different your relationship with your doctoral advisor would be if you personally constituted their retirement plan? Their incentives would be much better aligned, to say the least.

A weaker, non-economic version of the master-apprentice relationship does still apply for professors and thinkers who care about their legacy. And if a field has clearly identifiable experts, it is extremely valuable to work with the best person you can, in the hopes of forming a relationship with them. You should approach them and offer to proofread their papers, or even offer to make their coffee. Getting to work with them for a few hours a day while they talk about their field could be an education well worth unpaid or underpaid labor, especially if they understand you are there because you truly care about the field, and want to learn.

You may find semi-retired prominent figures more approachable than you think. Try asking around at their institution, whether a university or a firm, to feel out your chances of meeting them. Devote some time to going to as many of their institution's relevant events as you can—maybe even crash some happy hours. As always, a cold email with a thoughtful response to one of their papers will most likely garner a response.

A little applied anthropology can go a long way. Say you wanted to figure out who Paul Graham hangs out with in London, in the hope of meeting and learning from him. In this case you should think: if I were Paul Graham, who would I hang out with in London? This is a more realistic task than it might seem. You are only a few degrees removed from everyone else on the planet. Once you find the right social circles, the key is finding someone who knows your target contact and giving them a reason to introduce you to them. At this point, your job becomes to align the incentives of the gatekeepers with your own. Show up at the right parties until you are able to go up to Paul Graham and ask, "Can I make you coffee?"

Prospective apprentices should be careful, as they are placing themselves in a position that can be economically exploited. Currently, there is no mechanism to make your mentor reward you financially. But if your payoff is knowledge, you will soon be able to reason for yourself if the relationship is worth it. It may not be. Perhaps the communication gap is too big, or the relationship too awkward. But finding a good master is worth the effort.

Functional Institutions

Finally, in addition to communities of practice and master-apprentice relationships, well-functioning institutions present you with a third option. Note that it's important to discover whether the field you're entering already exists, or whether it's about to crystallize. Only if it already exists will there be the option of joining a functional institution.

I may appear skeptical about the functionality of large institutions such as universities or newspapers at scale—my thoughts in this regard are best summarized in How To Use Bureaucracies, Functional Institutions Are The Exception, Institutional Failure As Surprise, and Intellectual Dark Matter—but I do think functional institutions exist. Some institutions work extremely well at scale. I would consider Microsoft Research to be fairly functional. Even though Microsoft is a large company, they do a pretty good job of maintaining a professional environment.

When I say professional, I do not mean "behaves properly." What I mean is "adheres to an expected social role." When I show up to Dr. Bob's office, I don't expect to deal with Bob; I expect to deal with The Doctor. The doctor is not going to behave as Bob, Alice, or Caroline might act, even if that's who the doctor is. The doctor is going to behave as The Doctor! This is essentially a human user interface. I'm not Samo. I am The Patient. And Alice is The Doctor. Professionalism is a LARP—and an important one. It might seem as natural as gravity that there are doctors and patients, but these are social roles. And we had to be educated to fit those social roles.

If you are thinking about entering an institution, you should seek a very specific culture of professionalism, one that seems to match the task at hand, rather than what amounts to mutual sabotage. If the professionalism of a large at-scale institution is well-designed, the overall incentives of institutional leadership will be aligned with the overall mission. And if such an institution deploys efficient bureaucracies, it can function well as a productive research organization. But again, functional institutions are the exception. So only institutionalize yourself—that is, enter a mostly bureaucratic environment—if you believe the institution that you are entering to be a functional one.

One way to determine if the bureaucratic mechanism is efficient is whether or not they expect you to do all of the paperwork. Inefficient, decaying bureaucracies are a little bit like dying stars: they eject most of their mass. Their internal cancerousness generates reams of

excess paperwork, and they are thus incentivized to push paperwork onto the user. For example, if a public official likes you, they may ask you to fill out a few lines and send you on your way, but if they dislike you, they may give you a mound of paperwork for the exact same task. I have some experience of this growing up in Eastern Europe, but Americans can easily experience this at the local DMV.

If you experience low bureaucratic burden in a highly bureaucratized organization, that bureaucracy is working very well. It's like a machine that hums in the background, versus a machine that screeches and puffs smoke into your face. It's going to be a very visible experience, even if you don't have much other information.

Finally, check the institution's leadership. Even if it is a well-oiled machine, make sure that its roles are very clearly delineated towards the mission, rather than towards internal conflict. The leadership might not want your output and may actively steer away from your output. A good example of this dynamic can be found in Richard Feynman's critique of NASA's administrative process. After the Challenger explosion, NASA called on Feynman to find out what had gone wrong with their bureaucratic process. The issue, he found, was that the leadership didn't want to hear bad news from the engineers, because they wanted to push through as many flights as they could in order to score PR wins. And the Challenger mission, with a schoolteacher on board in an effort to demonstrate that space was to be for everyone, was an important PR flight. Leadership didn't want to hear the truth, even though the engineers themselves were still acting professionally and some parts of the NASA bureaucracy were still functioning well. The Challenger went up in flames, and the Space Shuttle remained expensive.

Conclusion

Entering a field for the first time as an autodidact, you should decide whether you want to enter into a community of practice, a master-apprentice relationship, or a functional institution. Try to figure out for each of those whether it fits the criteria. Seek out an energetic community of practice, or a good master. If you are considering joining an institution, make sure that its bureaucracy is working well, its professionalized roles are functional, and its leadership is aligned with correct output.

Leaps of faith like these can be harrowing, but with planning they need not be. Tracking existing sources of prestige and economic stability may be rational in a narrow sense, but the world of institutions in which they are embedded is prone to dysfunction. Recognizing when dysfunction has overtaken an institution is key to ensuring that we remain able to generate new knowledge about the world. History is littered with examples of collapsed civilizations that failed to do so. New scientific fields carry the promise of civilizational advancement, but they are not discovered automatically. It's up to us to find them.

Read more from Samo Burja here.

Security Mindset and Takeoff Speeds

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

About this post

This post is a stylized transcript of a conversation between Rohin Shah and Daniel Filan, two graduate students at CHAI, that happened in 2018. It should not be taken as an exact representation of what was said, or even what order topics were brought up in, but it should give readers a sense of what was discussed, and our opinions on the topic of conversation at the time of the discussion. It also should not be assumed that any of the points brought up are original to Rohin or Daniel, just that they represent our thinking at the time of the conversation. Notes were taken by Rohin during the conversation, and Daniel took the lead in writing it into a blog post.

The conversation was precipitated when Rohin noted that researchers at CHAI, and in AI alignment more broadly, diverged in their attitudes towards something like security mindset, that we will just call security mindset despite being somewhat different to what is described in the linked blog post.

Some researchers at CHAI are very concerned about building an axiomatic understanding of artificial intelligence and proving solid theorems about the behaviour of systems that we are likely to build. In particular, they are very concerned about expecting benign behaviour without a formal proof to that effect, and believe that we should start worrying about problems as soon as we have a story for why they will be problems, rather than when they start manifesting. At the time of the conversation, Daniel was one of these researchers, who have what we'll call "security mindset".

In contrast, some other researchers believe that we should focus on thinking about what extra machinery is needed to build aligned AI, and try building that extra machinery for current systems. Instead of dealing with future anticipated problems today and developing a theory that rules out all undesired behaviour, these researchers believe that we should spend engineering effort to detect their occurence and fix problems once we know that they occur and have more information about them. They think that rigour and ordinary paranoia are important, but less important than security mindset advocates claim. At the time of the conversation, Rohin was one of these researchers.

During a prior conversation, Rohin noted that he believed that security mindset was less important in a world where the power of AI systems gradually increased, perhaps on an exponential curve, over a period of multiple years, as opposed to a world where AI systems could gain a huge amount of power rather suddenly from designers having conceptual breakthroughs. Daniel was intrigued by this claim, as he had recently come to agree with two posts arguing that this sort of 'slow takeoff' was more likely than the alternative, and was unsure how this should affect all his other views on AI alignment. As a result, he booked a separate meeting to exchange and discuss models on this topic. What follows is a record of that separate meeting.

The conversation

Daniel: Here's my worry. Suppose that we're thinking about an AI system that is better at, say, math or engineering than humans. It seems to me that this AI system is going to have to be able to do some sort of optimization itself—maybe thinking about how to optimize physical structures so that they don't fall down, or maybe thinking about how to optimize its own computation so that it can efficiently find proofs of a desired theorem. At any rate, if this is the case, then what we have on our hands is optimization that is being done in a direction other than "behave maximally predictably", and is plausibly being done adversarially. This is precisely the situation in which you need security mindset to reason about the system on your hands.

Rohin: I agree that security mindset is appropriate when something is optimizing adversarially. I also agree that holding capability levels constant, the more we take a security mindset approach, the more safe our resulting systems are. However:

1. We simply don't have time to create a system that can be proved aligned using security-mindset-level rigour before the first prepotent AI system. This means that we need to prioritize other research directions.

2. Because we will likely face a slow takeoff, things will only change gradually. We can rely on processes like testing AIs, monitoring their thoughts, boxing them, and red-teaming to determine likely failure scenarios. If a system has dangerous abilities that we didn't test for, it will be the weakest possible system with those dangerous abilities, so we can notice them in action as they produce very minor damage, disable that system, create a new test, and fix the problem.

3. We should instead focus on constructing AI systems that correctly infer the nuances of human intent, rather than trying to address problems that could arise ahead of time. This will plausibly work to create an AI that can solve the harder problems for us.

Daniel: I have a few responses to those points.

1. Regarding your first point, I'm more optimistic than you. If you look at the progress made on the Agent Foundations research agenda in the past five years (such as work on reflective oracles and logical induction), for example, it seems like we could solve the remaining problems in time. That being said, this isn't very cruxy for me.

2. Regarding your second point, I think that in order to write good tests, we will need to take a security mindset approach, or at least an ordinary paranoia approach, in order to determine what things to test for and in order to write tests that actually rule out undesired properties.

3. In general, I believe that if you do not build an AI with security mindset at the forefront of your concerns, the result will be very bad—either it will cause an unacceptable level of damage to humanity, or more likely it just won't work, and it will take a very long time to fix it. This sucks, not just because it means that your work is in some sense wasted, but also because…

4. There will likely be a competing AI group that is just a bit less capable than you, and a different group just less capable than them, and so on. That is to say, I expect AI capabilities to be continuous across space for similar reasons that I would expect them to be continuous across time.

5. As a result of 3 and 4, I expect that if your group is trying to develop AI without heavy emphasis on security mindset, you fail and get overtaken by another group, and this cycle continues until it reaches a group that does put heavy emphasis on security mindset, or until it creates an AI that causes unacceptable levels of damage to humanity.

Rohin: I doubt your point 4. In our current world, we don't see a huge number of groups that are realistic contenders to create smarter-than-human AI, and the groups that we do see show a promising degree of cooperation, such as collaborating on safety research and making promising commitments towards avoiding dangerous race dynamics. Also, I think that in worlds where there is such a break-down of cooperation that your point 4 applies, I think that technical work today is near-useless, so I'm happy to just ignore these worlds.

I also think that the arguments that you give for point 4 are flawed. In particular, the arguments for slow take-off require gradual improvement that builds on itself, which happens over time but is not guaranteed to happen over space. In fact, I expect there to be great resource inequalities between groups and limited communication between competing groups, which should generate very large capabilities gaps between competing groups. This is something like a local crux for me: if I thought that there weren't resource inequalities and limited communication, I would also anticipate competing groups to have similar levels of capabilities.

Daniel: Hmmmmmm. I'll have to think about the arguments that I should anticipate large capability gaps between competing groups, but they seem pretty convincing right now.

Actually, maybe we should expect the future to look different to the past, with countries like China and India growing capable AI labs. In this world, it's sadly plausible to me that pairs of countries' research groups could end up failing to cooperate. But again, I'll have to think about it more.

At any rate, even if my point 4 fails, the rest of my points imply that research done without security mindset at the forefront will reliably be useless, which still feels like a strong argument in favour of security mindset.

Rohin: Then let's move on to your points 2 and 3.

Regarding 3, I agree that if you have a vastly super-human AI that was not designed with security mindset in mind, then the outcome will be very bad. However, for an AI that is only incrementally more powerful than previous, already-understood agents, I think that incremental improvements on existing levels of rigour displayed by top AI researchers are sufficient, and also lower than the levels of rigour you, Daniel, would want.

For example, many putative flaws with superintelligence involve a failure of generalization from the training and test environments, where the AI appears to behave benignly, to the real world, where the AI allegedly causes massive harm. However, I think that AI researchers think rigorously enough about generalization failures—if they did not, then things like neural architecture search and machine learning more broadly would fail to generalize from the training set to the test set.

Daniel, not quite getting the point: This feels quite cruxy for me. I believe that top AI researchers can see problems as they happen. However, I do think that they have significantly less rigour than I would want, because I can see problems that I suspect

are likely to come up with many approaches, such as inner alignment failures, and these problems weren't brought to my attention by the AI research community, but rather by the more security-mindset-focussed contingent of the AI alignment research community. If this is the case, it seems like a big win to find these problems early and work on them now.

Rohin: If inner alignment failures are a big problem, I expect that we would find that out in ~5 years, and that a unit of work done on it now is worth ~10-20% of a unit of work done on it after we have a concrete example of how they are a problem. Given this, instead of working on those sorts of problems now, I think that it makes sense to work on things that we actually know are problems, and have a hope of solving in the present, such as communicating human intent to neural networks.

Daniel: I'm skeptical of those numbers. At any rate, it seems to me that there might be problems that you can solve in that way, but that there are also some things that you need to get right from the beginning. Furthermore, I think that you can form decent models about what these things are, and examples include the Agent Foundations agenda as well as the more theoretical aspects of iterated distillation and amplification research.

Rohin: Interesting. I'd like to get down later into our models of what problems need to be done right now, but for now that feels a bit off topic. Instead, I'd like to hear why you believe your point 2, that security mindset is needed to do monitoring, testing, and boxing well.

Daniel: Well, I have three reasons to think this:

1. You are plausibly dealing with an AI that is optimizing to pass your test. This is the sort of case where security mindset is required for good reasoning about the system.

2. Your suggestion of monitoring thoughts is quite exciting to me, since it could plausibly detect any adversarial optimization early, but it's hard for me to see how you could be sure that you've done that adequately without the type of thinking produced by security mindset.

3. You are optimizing to create an AI that passes the test by trying a bunch of things and thinking about how to do it. Again, this is a situation where optimization is being done, perhaps to pass the specific tests that you've set, and therefore a situation that you need security mindset to reason correctly about.

Rohin: Points 1 and 3 seem solid to me, but I'm not sure about point 2. For instance, it seems like if I could 'read minds' in the way depicted in popular fiction, then by reading the mind of another human all the time, I would be able to detect them trying to take over the world just by reasoning informally about the contents of their thoughts. Do you agree?

Daniel, answering a slightly different question: If you mean that I get to hear what's happening in their verbal loop, then I'm not sure that I could detect what people were optimizing for. For instance, it's plausible to me that if you heard the verbal loop of a dictator like Stalin, you would hear a lot about serving his country and helping the workers of the world, and very little about maximizing personal power and punishing people for disagreeing with him.

That being said, it seems to me like the primary part where security mindset is required is in looking at a particular human brain and deducing that there's a verbal loop containing useful information at all.

Well, it's about time to wrap up the conversation. Just to close, here are my cruxes:

- How high is the "default" level of security mindset and rigour? In particular, is it high enough that we should outsource work to the future?
- How much security mindset/rigour does one need to do monitoring, testing, and boxing of incrementally advanced AIs well?
    - The underlying question here is something like how much optimization does a smart AI do itself?
- At any given time, how far apart in capabilities are competing groups?

What posts do you want written?

I have many posts that I want written but do not have time to write and I suspect there are other people that feel similarly. This post on the Solomonoff prior was one example, until I got fed up and just wrote it.

Please write one post idea per answer so they can be voted on seperately.

Box inversion hypothesis

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

This text originated from a retreat in late 2018, where researchers from FHI, MIRI and CFAR did an extended double-crux on AI safety paradigms, with Eric Drexler and Scott Garrabrant in the core. In the past two years I tried to improve it in terms of understandability multiple times, but empirically it seems quite inadequate. As it seems unlikely I will have time to invest further work into improving it, I'm publishing it as it is, with the hope that someone else will maybe understand the ideas even at this form, and describe them more clearly.

The box inversion hypothesis consists of the two following propositions

1. There exists something approximating a duality / an isomorphism between technical AI safety problems in the Agent Foundations agenda and some of the technical problems implied by the Comprehensive AI Services framing
2. The approximate isomorphism holds between enough properties that some solutions to the problems in one agenda translate to solutions to problems in the other agenda

I will start with an apology - I will not try to give my one paragraph version of the Comprehensive AI Services. It is an almost 200 pages long document, conveying dozens of models and intuitions. I don't feel like being the best person to give a short introduction. So, I just assume familiarity with CAIS. I will also not try to give my short version of the various problems which broadly fit under the Agent Foundations agenda, as I assume most of the readers are already familiar with them.

0. The metaphor: Circle inversion

People who think geometrically rather than spatially may benefit from looking at a transformation of a plane called circle inversion first. A nice explanation is here - if you have never met the transformation, pages 1-3 of the linked document should be enough.

You can think about the "circle inversion" as a geometrical metaphor for the "box inversion".

1. The map: Box inversion

The central claim is that there is a transformation between many of the technical problems in the Agent Foundations agenda and CAIS. To give you some examples

- problems with daemons <-> problems with molochs
- questions about ontologies <-> questions about service catalogues
- manipulating the operator <-> addictive services
- some "hard core" of safety (tiling, human-compatibility, some notions of corrigibility) <-> defensive stability, layer of security services
- ...

The claim of the box inversion hypothesis is that this is not a set of random anecdotes, but there is a pattern, pointing to a map between the two framings of AI safety. Note that the proposed map is not exact, and also is not a trivial transformation like replacing "agent" with "service".

To explore two specific examples in more detail:

In the classical "AI in a box" picture, we are worried about the search process creating some inner mis-aligned part, a sub-agent with misaligned objectives.

In the CAIS picture, one reasonable worry is the evolution of the system of services hitting a basin of attraction of so-called moloch - a set of services which has emergent agent-like properties, and misaligned objectives.

Regarding some properties, the box inversion turns the problem “inside out”: instead of sub-agents the problem is basically with super-agents.

Regarding some abstract properties, the problem seems similar, and the only difference is where we draw the boundaries of the “system”.

2. Adding nuance

Using the circle inversion metaphor to guide our intuition again: some questions are transformed into exactly the same questions. For example, a question whether two circles intersect is invariant under the circle inversion. Similarly, some safety problems stay the same after the "box inversion".

This may cause an incorrect impression that the agendas are actually exactly the same technical agenda, just stated in different language. This is not the case - often, the problems are the same in some properties, but different in others. (Vaguely said, there is something like a partial isomorphism, which does not hold between all properties. Someone familiar with category theory could likely express this better.)

It is also important to note that apart from the mapping between problems, there are often differences between CAIS and AF in how they guide our intuitions on how to solve these problems. If I try to informally describe the intuition

- CAIS is a perspective which is rooted in engineering, physics and continuity"continuum"
- Agent foundations feel, at least for me, more like coming from science, mathematics, and a "discrete/symbolic" perspective

(Note that there is also a deep duality between science and engineering, there are several fascinating maps between "discrete/symbolic" and "continuum" pictures, and, there is an intricate relation between physics and mathematics. I hope to write more on that and how it influences various intuitions about AI safety in some other text.)

3. Implications

As an exercise, I recommend to take your favourite problem in one of the agendas, and try to translate it to the other agenda via the box inversion.

Overall, if true, I think the box inversion hypothesis provides some assurance that the field as a whole is tracking real problems, and some seemingly conflicting views are actually closer than they appear. I hope this connection can shed some light on some of the disagreements and "cruxes" in AI safety. From the box inversion perspective, they sometimes seem like arguing whether things are inside or outside of the circle of symmetry in a space which is largely symmetrical to circular inversion.

I have some hope that some problems may be more easily solvable in one view, similarly to various useful dualities elsewhere. At least in my experience for many people it is usually much easier to see some specific problem in one of the perspectives than the other.

4. Why the name

In one view, we are worried that the box, containing the wonders of intelligence and complexity, will blow up in our face. In the other view, we are worried that the box, containing humanity and its values, with wonders of intelligence and complexity outside, will crush upon our heads.

Words and Implications

> Professor Quirrell didn't care what your expression looked like, he cared which states of mind made it likely.

- HPMOR, Ch. 26

Words should not always be taken at face value. Presumably you know this. You probably have some heuristics about specific situations or claims in which a person's words should not be taken literally. But I think most peoples' heuristics here are far too narrow - that is, most people take words literally far too often.

The sequences talk about habitually asking, in everyday life, "What do I think I know, and how do I think I know it? What physical process produced this belief?". I suggest a similar habit for words in everyday life: "What is being said, and why is it being said? What physical process produced these words?".

This post is a bunch of examples, in an attempt to goad your system-1 into looking past surface-level meanings more often.

Dishes

Once or twice a week, I'll hear my girlfriend yell from the kitchen "Jooooooohn! Why are there so many dirty dishes in the sink?". Going to wash the dishes is *not* the correct response to this.

If I go wash the dishes, then she will quite consistently find something else to complain about in the meantime - floor needs sweeping, nothing to eat, neighbors are noisy, etc. Usually multiple other things. It was never really about the dishes in the first place, after all. Really, she's stressed and looking for an outlet.

A hug fixes the problem much more effectively than washing the dishes would.

The general mental motions required to notice this are something like:

- Stop. Don't just go wash the dishes.
- Ask why this is coming up, and in particular why it's coming up *right now specifically*. Is there any particular reason the dishes are relevant *right now*? (Sometimes the answer is "yes", and then it's time to go do the dishes.)
- If there isn't a reason why the dishes are relevant right now, then I need to figure out the actual reason for the complaint.

This scene from the movie Limitless is a similar example. It's a bit over-the-top, but it's one of the few examples of a supposedly-intelligent person in a Hollywood movie actually doing something intelligent (as opposed to technobabble).

Designers and Engineers

If you work in software, one problem you've probably encountered on the job is "the thing they literally ask for is only very loosely correlated with the thing they actually want".

A designer or product manager comes to a software engineer with some crazy request. They want to redesign a particular button, move it to a different place on the page, and change what it does, but still keep it the same button. A very confused engineer asks "What on earth does that even mean? How is it supposed to be the same button when everything has

changed?”. After far too many questions, it turns out that the product team just wanted to re-use the tracking from the old button, because adding new columns to their data is annoying. Main point: the thing they literally ask for is only very loosely correlated with the thing they actually want.

Meanwhile, that same product team is testing out a prototype with potential users and collecting feedback. The users have all sorts of crazy requests. One of them wants a summary page with a bunch of app-internal numbers on it. After asking far too many questions, the product team figures out that what this user actually wants is a way to generate receipts for their customers. Once again, the thing they literally ask for is only very loosely correlated with the thing they actually want.

Down the hall, a manager asks an analyst for the click-through rate on the checkout screen. What the manager actually wants to know is whether lowering prices would lead to more sales. Whether that click-through rate is a good proxy for customers’ price sensitivity is the sort of question the analyst needs to answer, which means the analyst needs to figure out that that’s the real question in the first place. The thing they literally ask for is only very loosely correlated with the thing they actually want.

In another office, the COO has found a used conveyor system on sale and wants to buy it for the warehouse. They ask a lawyer to write up the contract for the purchase, and to keep it simple - just a straightforward asset purchase. The COO probably hasn’t even thought about what happens if the conveyor is defective; the lawyer needs to realize that the COO probably wants the contract to cover any potential problems, even though the COO has not thought about it. The thing they literally ask for...

We could go on all day.

In the information/knowledge economy, a key part of most jobs is realizing that what someone literally asks for is only very loosely correlated with what they actually want. The mental motions required to handle such problems effectively are much the same as the previous section:

- Stop. Don’t just immediately do what was literally requested.
- Ask why this particular request was made. Is there an obvious goal, and is this clearly the best way to achieve that goal?
- If not, what’s the real goal, and what’s the best way to achieve that goal?

Salespeople

I was on vacation in Mexico with my parents and siblings. We were headed to a touristy beach-park, and took a cab. The cab driver “helpfully” suggested an alternative touristy beach-park. This sounded like the sort of “helpful” suggestion which would earn the driver a kickback, and this hypothesis was promptly confirmed when she handed my mother a laminated advertisement for the place.

… at which point my mother leaned over and said “Hey this looks pretty nice! And it’s even pretty cheap.”

Unrolled into a dialogue, my reaction to this was something like…

Inner voice 1: “Huh??? There isn’t any information about niceness or price on that piece of paper.”

Inner voice 2: “It’s the literal content of the words. There’s a price written on there, and it’s a bit lower than the place we're going.”

Inner voice 1: “Ok, but what does the number on that piece of paper have to do with the amount of money which would change hands at the gate to this place? It’s an ad aimed at tourists, it’s almost certainly misleading. And same with the pictures.”

Inner voice 2: “I don’t think your mother realizes that.”

Inner voice 1: <exasperated sigh>

Point of the story: obviously do not trust information from advertisements or salespeople.

The one exception to this is information which does not seem tailored to make the sale happen. That said, be careful - salespeople can get kickbacks in nonobvious ways. Today’s car dealers, for instance, make most of their money on kickbacks from financing and warranty deals rather than the car itself. (I know this from firsthand experience - I worked at an online car dealership a few years back.)

I won’t do a political example here, but this also includes politicians. It especially includes politicians from your own preferred party. Also note that politicians tend to rely more on bullshit than lies, relative to salespeople - it's not just a question of whether their words are "trustworthy", but of whether they have any correspondence to the real world at all. Ask what physical process resulted in these particular words, and often the answer will be "signalling group loyalty" or "polled well with constituents", with physical reality playing no significant role.

The Parable Of The Dagger

People tend to generalize “don’t trust ads/salespeople” to an heuristic like “be suspicious of the incentives behind information-sharing”. This isn’t a bad heuristic, but it’s the sort of heuristic which makes it a little too easy to miss the more general rule. The taxi driver trying to sell us on a beachpark is highly salient, but the key underlying factor is that the letters and numbers on the piece of paper do not necessarily have anything to do with the amount of money changing hands at the gate. “Be suspicious of incentives” is less general than “ask what causal process resulted in these words”.

The parable of the dagger makes this point more directly. A jester has angered the king (with a tricky logic puzzle) and been thrown in the dungeon. The king sets up a puzzle for him...

> The jester was brought before the king in chains, and shown two boxes.
>
> "One box contains a key," said the king, "to unlock your chains; and if you find the key you are free. But the other box contains a dagger for your heart, if you fail."
>
> And the first box was inscribed:
>
> "Either both inscriptions are true, or both inscriptions are false."
>
> And the second box was inscribed:
>
> "This box contains the key."

The jester correctly reasons through the puzzle, and picks the second box, only to find that it contains the dagger.

> "How?!" cried the jester in horror, as he was dragged away. "It's logically impossible!"
>
> "It is entirely possible," replied the king. "I merely wrote those inscriptions on two boxes, and then I put the dagger in the second one."

The steps the jester would need to take to avoid his death are quite similar to the mental motions from earlier:

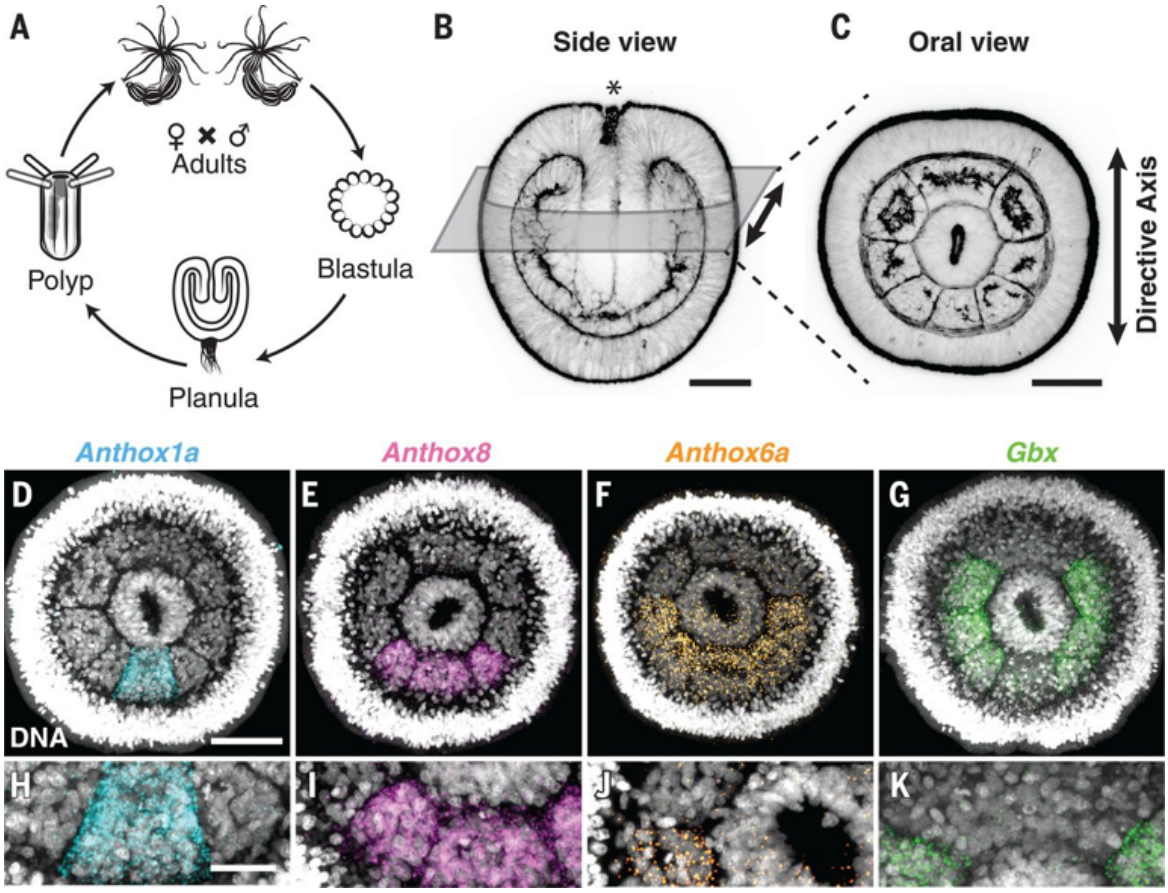
- Stop. Don't just take the words written on the boxes at face value.
- Is there an obvious reason the words on the boxes would accurately predict my fate?
- If not, then what is the king really up to?

Reading Papers

Finally, a more complex example.

An evo-devo class I was sitting in on assigned this paper. The experimenters were interested in the evolution of *Hox* genes - genes typically used in animals to establish different roles for different body segments along the head-to-tail axis (e.g. the segments of an ant or bee, the sections of a human spine, etc). They found *Hox*-analogues in a sea anemone - rather odd, since the anemone doesn't have the sort of specialized head-to-tail segments with which *Hox* are usually associated. So, the experimenters investigated the role of those genes in anemone specifically.

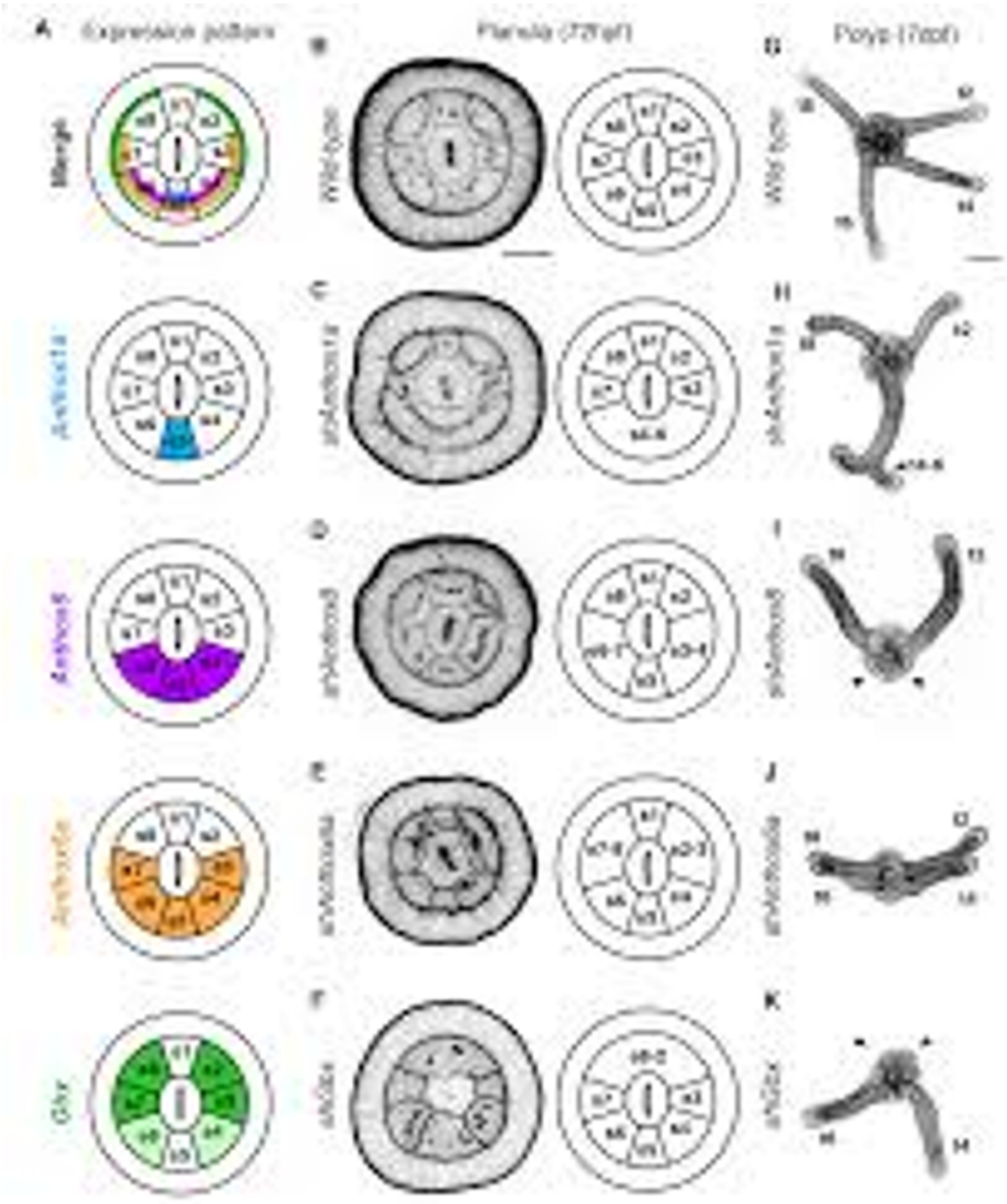
The highlight of the paper was this image, which tells most of the story on its own:

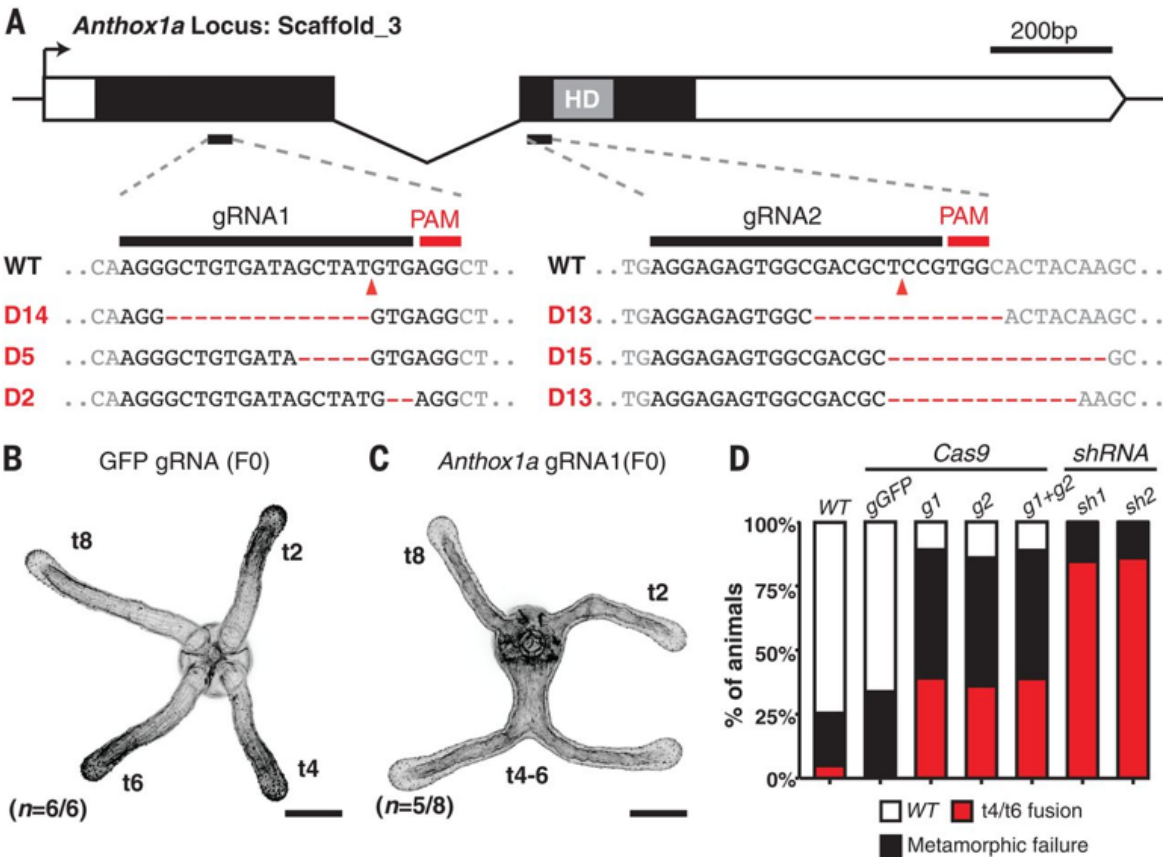


Those colors each represent the activity of one *Hox* protein. They behave exactly like they do in other animals, with each protein lighting up one segment further than the preceding protein… *except* rather than lighting up head-to-tail *along* the length of the animal, they're ordered axially *around* the animal. (The order in which they light up is determined by the

order in which their genes appear in the genome - the genes are in a line, so that a repressor/promoter targeting one will also repress/promote the *Hox* genes after it.)

The experimenters then use both RNA interference and CRISPR (in separate experiments) to suppress/knock out specific *Hox* genes, and show that this results in some of the segments "merging" - which in turn gives the anemone merged tentacles from those segments.





Here's the weird thing: the results from the RNAi experiments were much more impressive, and much more detailed, than the results from the CRISPR experiments. (You can see that visually in the figures above: the tentacles merge much more dramatically in the RNAi examples on the top than the CRISPR examples on the bottom.) Why?

You might guess that there's some biological weirdness going on, random things interfering with other random things, as sometimes happens in the messiness of biological systems. But my guess is that it's mostly not about the underlying biology. Reading between the lines of the paper, it sounds like the lab has lots of expertise and experience with RNA interference methods. But CRISPR was the hot new thing, so probably some grad student or reviewer suggested that the paper would be sexier if they threw in a quick CRISPR experiment. The lab lacked experience with this sort of genetic engineering, so probably they just didn't do it in the most effective way, and ended up with less-impressive results for that experiment. (The class professor, who was familiar with past work from this lab, confirmed that this sounded likely.)

As always, it's the same basic steps:

- Stop. Don't just assume that the words and figures in the paper are directly representative of the system under study.
- Ask where these words and figures came from. What did the experimenters actually do, what thoughts went through their heads?
- To the extent that the words and figures reflect something other than the system under study, what can we deduce from them?

One particularly common case is researchers claiming implications which their data do not establish - especially "X causes Y". (I won't provide an example, partly because I don't usually save those papers.) As always, we need to look at the actual process which generated the words: what experiments were actually run, what were the results, and do

they actually establish the causal claim? Are the results not just necessary but *sufficient* to establish causality? If not, what information *can* we glean from the results?

Conclusion

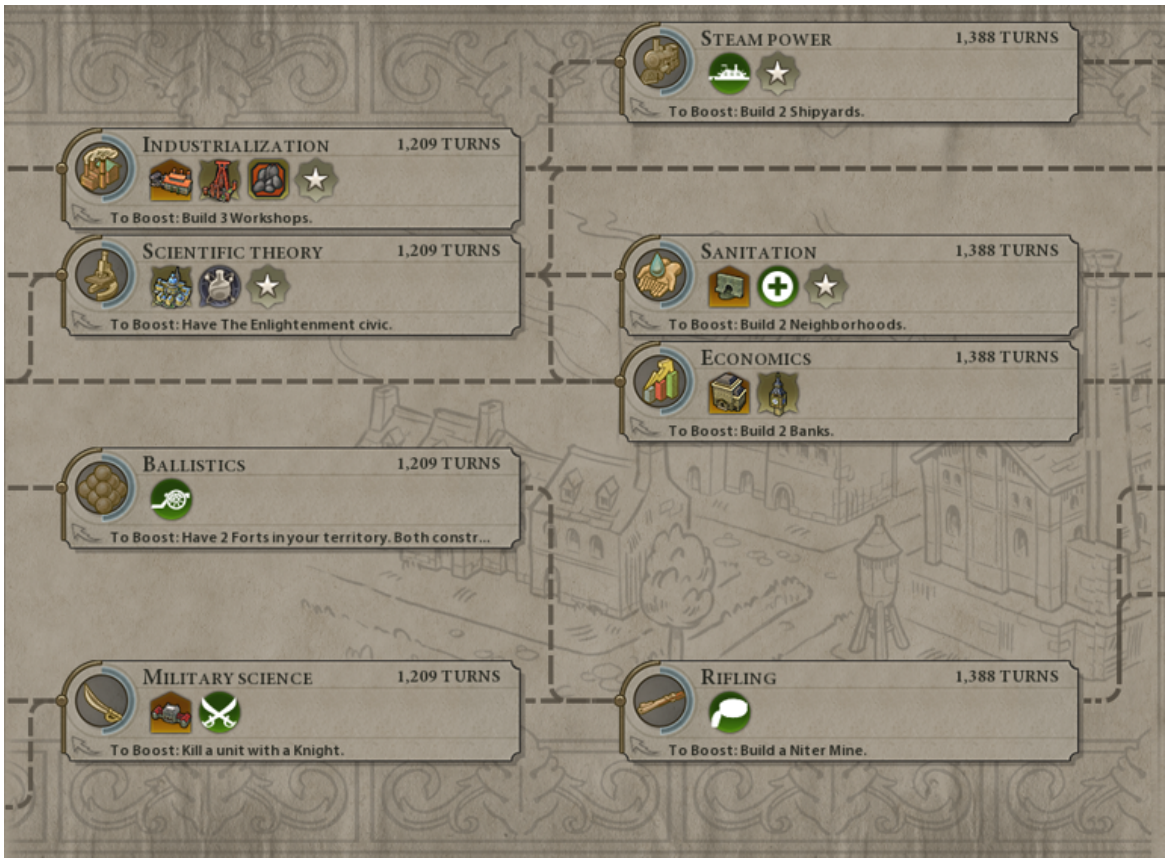
Most people have a variety of heuristics about when (not) to take words at face value: beware of ulterior motives, beware of people asking for things they don't understand, check whether claims in a paper's abstract are actually established by the results. These are good heuristics to have, but they make it easy to overlook the more general technique: "What is being said, and why is it being said? What physical process produced these words?".

The basic mental steps:

- Stop. Don't just automatically take the words at face value.
- Ask what physical process generated the words. Where did they come from? Why these particular words at this particular time?
- What can we deduce from the fact that the words were spoken, other than the literal content?

Lessons on Value of Information From Civ

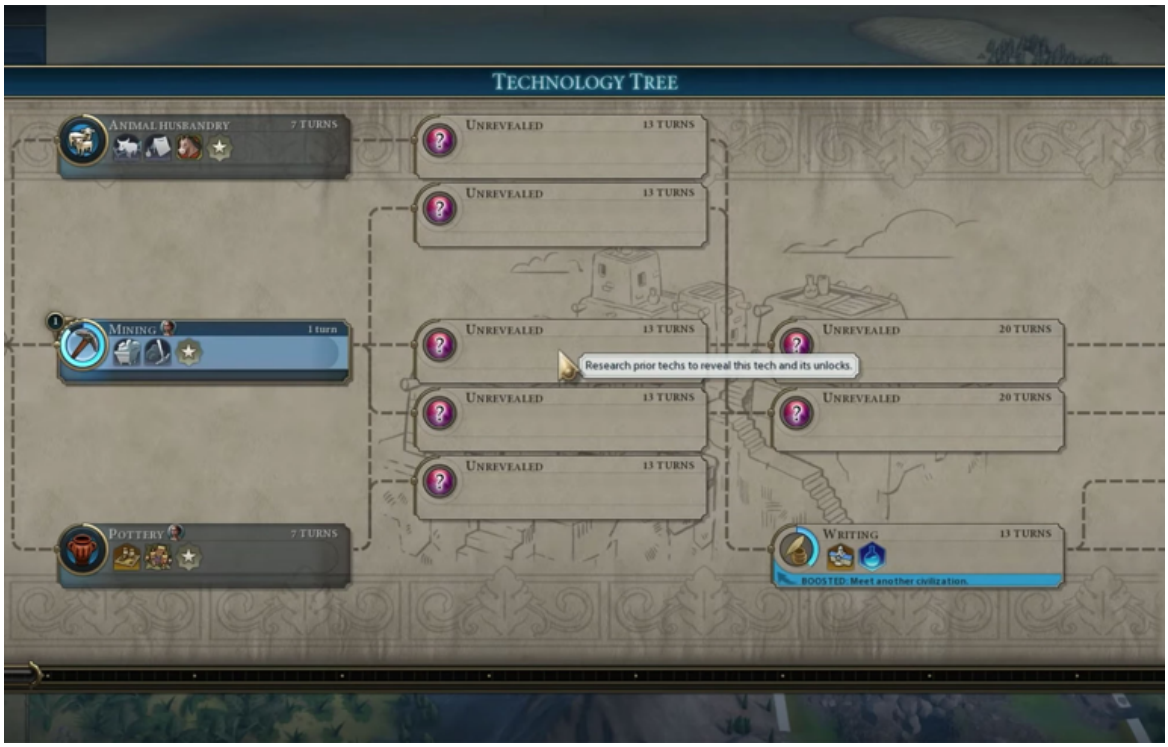
I'm a big fan of the game Civilization. Civ is the canonical empire-building game: you explore the world, expand your empire, manage your economy, invade other empires, unlock new tech, etc. A few months ago, the game released "tech shuffle mode", which radically changes the game (and IMO makes it a lot more interesting). Usually, a not-yet-unlocked section of the tech tree looks like this:



To unlock a tech, we first need to unlock the prerequisite techs (in left-to-right order). The tree is the same in every game, so we can plan ahead to unlock key technologies early - e.g. we can ignore the lower half of the tree in order to get to Steam Power faster. But with tech shuffle mode, a not-yet-unlocked section of the tree looks like this:

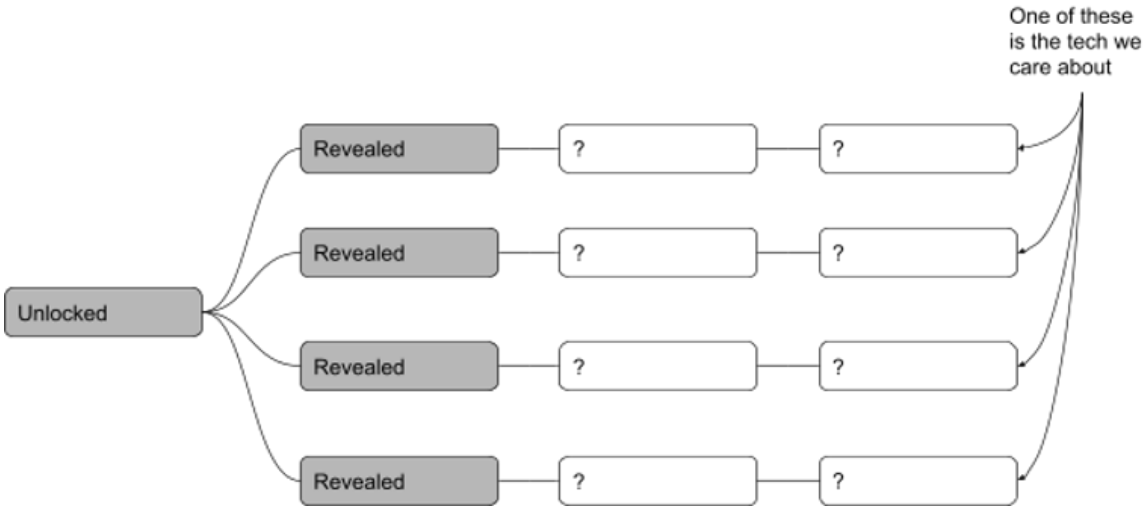
The tree structure is randomly generated each game (although techs are still grouped by “era”, so e.g. we won’t unlock nukes right at the start of the game). Techs are also hidden until we’re ready to research them (i.e. until we unlock the prereqs). That means, for instance, that we can’t focus resources on techs needed to unlock Steam Power because we don’t know where in the tree Steam Power is.

But there’s a catch: there are actions we can take to *reveal* a tech in the tree. This is the “tech boost” mechanic - for instance, by finding another civilization, we can “boost” Writing, reducing the resource cost to unlock it by 40% and (more importantly) revealing its position in the tree. It looks like this:



In other words: we can pay to acquire information.

This is especially important when the value of information (VOI) is high. What does that look like? Like this:



There’s several different branches, each several techs deep. The strategically-important tech we want is somewhere at the end of one of those branches, but we don't know *where*. Brute-force searching all of the paths until we find it would require unlocking ~½ of these techs on average, but if we knew which branch to take, then we’d only need to unlock one branch - i.e. ~¼ of them in this case. So, I can pay some resource cost to reveal my target tech in the tree, and then I can get to it about twice as quickly (on average).

I realized the importance of the tech boost mechanic as a way of purchasing information after my first game on tech shuffle mode. It seemed that VOI was often very high, so I made an explicit effort to pay attention to it in my next few games.

The result: even when explicitly trying to pay attention to VOI, I regularly noticed that I wasn’t investing enough in it.

The basic problem is that VOI isn’t readily *visible*. I click into the tech tree, and I see something like

… and the natural response is “ok, my options are Animal Husbandry, Mining, or Pottery. Of those, Mining seems most useful right now, since it will let me build mines”. It’s so easy to say “well, I don’t have any information to prefer one unrevealed tech over another, and Mining is the most useful of the visible options, so I might as well start with Mining”. Animal Husbandry, Mining and Pottery are right there, they’re visible, they’re salient, they’re the obvious things to focus on. I have to step back a minute to realize that Animal Husbandry, Mining and Pottery are all basically-irrelevant right now compared to e.g. Shipbuilding, far and away the most important thing is to unlock Shipbuilding, which means the top priority is to *figure out where Shipbuilding is*. It’s so easy to weigh the visible options against each other, it’s such a natural instinct, but the optimal choice is to ignore the visible, direct consequences of each choice and instead gain more information.

This is, of course, a metaphor for the real world.

The real world is high-dimensional, so VOI is orders-of-magnitude higher than in Civ’s low-dimensional tech tree. There aren’t 4 possible paths to choose from, there are hundreds or thousands at least. The VOI isn’t a factor of 2 cost reduction on average, it’s a factor of 50 or 500 cost reduction. Brute-force work on every project in some random order realistically may *never* find the one or two big things which matter more than anything else.

It’s so easy to say “well, of all these projects, adding a page for X seems like it would improve our app the most”. Or “well, of all these projects, running ads on Y seems like it would increase our profits the most”. Or “well, of all these projects, trying a standing desk seems like it would increase my productivity the most”. There’s a list of projects right there in front of us, it’s so easy to imagine the direct consequences of each project. But in a high-dimensional world, with high VOI, the optimal choice is usually to ignore the visible, direct consequences of each choice and instead gain more information.

Rule of Equal and Opposite Advice & Slack

References:

Slack: https://thezvi.wordpress.com/2017/09/30/slack/

TL;DR Having a 20% buffer in your schedule to rest and handle anything urgent that pops up is the difference between a good life with time to experiment and being in a situation where one emergency breaks you and everything is on fire all the time.

Rule of Equal and Opposite Advice: https://slatestarcodex.com/2014/03/24/should-you-reverse-any-advice-you-hear/

TL;DR The person who needs the advice to love their body is probably the person to end up on diet forums. The person who needs the advice to care about their health is probably going to end up on HAES forums. Each person needs to reverse the advice they naturally seek to get the advice that brings them to the healthy middle way rather than a self destructive extreme.

Main Post:

Calm > Anxiety; Slack > Distress:

The advice to seek out slack seemed to be the advice lesswrong needs. This forum is filled with people who worry about if they should be working 24/7 and donating all of their income. I think if your problem is on the anxiety spectrum it's really important that you step back from trying to find ways to get everything done and try the opposite advice of seeking slack.

Joy > Depression; Eustress > Slack:

I think I've gone too far down the slack hole and need to reverse the advice again. This forum is also filled with smart people who need a challenge to feel fulfilled. As more slack has found its way into my life[1] I've been having trouble motivating myself to do the miniscule workload asked of me; whereas in the past I've handled herculean schedules and loved it.

As a result, I've been seeking more and more slack thinking "I can barely do this, I clearly need to try even less". I need to reverse that. Of course, I don't want to go back to not have a single unscheduled second in my day. The goal is to settle the pendulum at a nice healthy medium where I have an engaging number of balls to juggle but if I drop one I'll still have time to pick it up before the next ball falls.

Conclusion:

If you’re feeling anxious, ask yourself where you can find more slack. If you’re feeling depressed, ask yourself where you can find things that excite you without making you anxious. Slack doesn’t necessarily mean having a lot of time where you do nothing. As long as your schedule has things you can push back without issue that gives you the freedom that slack is meant to give you even if you have relatively few unscheduled blocks on your calendar.

[1] Partially due to covid eliminating the hobbies I was putting most of my effort into

The Rise and Fall of American Growth: A summary

This is a linkpost for https://rootsofprogress.org/summary-the-rise-and-fall-of-american-growth

The Rise and Fall of American Growth, by Robert J. Gordon, is like a murder mystery in which the murderer is never caught. Indeed there is no investigation, and perhaps no detective.

The thesis of Gordon's book is that high rates of economic growth in America were a one-time event between roughly 1870–1970, which he calls the "special century". Since then, growth has slowed, and we have no reason to expect it to return anytime soon, if ever.

The argument of the book can be summarized as follows:

- Life and work in the US were utterly transformed for the better between 1870 and 1940, across the board, with improvements continuing at a slower pace until 1970.
- Since 1970, information and communication technology has been similarly transformed, but other areas of life (such as housing, food, and transportation) have not been.
- We can see these differences reflected in economic metrics, which grew significantly faster especially during 1920–70 than before or since.
- All of the trends that led to high growth in that period are played out already, and there are none on the horizon to replace them.
- Therefore, high growth is a thing of the past, and low growth will be the norm for the future.

The bulk of the book's 700+ pages are dedicated to the first three points above: a qualitative and quantitative survey of how the American standard of living has changed since 1870.

In the several decades after 1870, every aspect of American life was transformed:

Food: In 1870 Americans were well-fed, but with a monotonous diet high in pork and cornmeal, foods that could easily be preserved without refrigerators. Over the coming decades diets became more varied, and food got easier to prepare, thanks to the introduction of home refrigerators, prepared foods, and supermarkets. But major innovation here was over by 1940.

Clothing: In 1870 most Americans owned only a few outfits. Most clothing was made in the home, by women, although some men's clothing might be purchased. By 1940 clothing was made in factories and purchased through department stores and other retail outlets.

Retail: The general store of 1870 gradually gave way to the urban department store, with better selection and lower prices, and rural areas were served by mail order catalogs, which took advantage of the postal service and the railways. Later, when cars became common, people had more shopping options because they could drive to the city, ending the monopoly of local stores in small towns.

The home: Many big changes occurred 1870–1940. Homes got electricity, running water, sewage, gas, and telephone service; Gordon summarizes this as the home becoming "networked." No more did people have to haul water from the well, or carry wood and coal into the house to be burned in the kitchen stove, and then to carry the dirty water and burnt ashes outside again. Homes got bathrooms with toilets to replace outhouses, and private bathtubs to replace the previous more public practice of bathing in the kitchen, with water heated on the stove. (Hauling and heating the water was laborious enough that most people

bathed no more than weekly.) Central heating meant that bedrooms received heat, instead of just the kitchen. Electricity, and the advent of labor-saving appliances such as the washing machine, simultaneously improved cleanliness and reduced total housework. And gas light was replaced by clean, safe, convenient electricity.

Transportation: Although railroads were fairly well established by 1870 (with the first transcontinental railroad having been completed in 1869), this only solved the problem of transit between major cities. Within cities, and on the last mile between the railroad and destinations such as individual farms, people still depended on horses. Horses were expensive to care and feed (a lot of US agriculture was devoted just to this purpose), they littered the streets with their urine and manure, and they could pull a load at only a few miles per hour. We escaped the “tyranny of the horse” starting with electric streetcars in the early 1900s, and then completely with the internal combustion engine and the automobile, which became affordable with the Model T turned out by Ford’s assembly line in the 1910s. By 1940 the automobile had fairly taken over American life. The major developments in transportation after this were the interstate highway system and commercial air travel, both of which were well established by 1970.

Communications: In 1870, communications were by postal mail, or by (expensive) telegram. Rural families in particular were extremely isolated. The years 1870 to 1940 saw the development of the telephone, phonograph, movies, and radio. By 1940, then, people could talk to family, friends, and business associates anywhere in the country (or even across the ocean, if it was worth the high price); they could listen to recorded music from great performers, and go to the cinema to see shows and (in the early days) newsreels; and they could tune in to real-time news, entertainment, and sports broadcasts. Between 1940 and 1970, the television industry grew, and television replaced some of the functions of film and radio.

Health & disease: In 1870 the germ theory was still being established, and it had yet to make any impact on American life. But in the coming decades, water filtration and chlorination was established in all major cities, milk became pasteurized, and the Food & Drug Administration (created 1906) began enforcing standards of purity and safe handling on meat, milk, and medicines. Mortality rates dropped quickly. And then between 1940 and 1970, antibiotics were discovered to treat most bacterial diseases.

Work: Work life improved for both men and women. For men, work shifted from dangerous and uncomfortable manual labor on farms to safer and more comfortable jobs indoors. Some of that work was routine, but an increasing share of it is not. For women, the great liberation was the reduction in the need for housework, thanks to the changes mentioned above: running water, electric appliances, and products such as premade clothes and prepared foods. This led to a big increase in women in the workforce after World War II.

However, many of these areas of life have seen comparatively minor improvements since 1970 and some of them since 1940:

Food and clothing have seen only minor changes since 1940, with some new products and new types of stores.

The home has seen modest improvements. Home sizes have steadily increased. The quality of appliances, including their reliability, improved steadily mid-century but was mostly complete by 1970 (based on an extensive survey of *Consumer Reports* magazine). The biggest changes since 1970 have been the microwave oven and the spread of air conditioning.

Transportation is fundamentally unchanged since the 1970s, after jet engines became common in air travel. Indeed, air travel has regressed on some dimensions, such as tighter seating and longer security lines.

Medicine has not had a revolution since 1970 comparable to the germ revolution, and the main treatments for cardiovascular disease and cancer already existed in 1970, although there have been some notable improvements such as the reduction in lung cancer from smoking.

The work week had already decreased to 40 hours by 1940.

In contrast, information and communications continued to see dramatic advances:

- Recorded music went from vinyl records to cassette tapes to CDs to MP3s to streaming services
- Recorded video became available, from VHS tapes to DVDs to video streaming
- Televisions got bigger and their pictures became sharper
- Phones went mobile and then turned into pocket computers
- The entire computer and Internet revolution happened

Gordon does not deny the revolution in electronics, computers, and the Internet. The essence of his qualitative argument is this: since 1970, only computing, communications and entertainment have been revolutionized (directly, with some indirect effects on other areas such as retail). But from 1870–1940, *everything* was transformed: not only communications and entertainment but also food, clothing, the home, transportation, medicine, and work. A revolution in one area of the economy is not, on its own, as big as the combined total of several simultaneous revolutions covering all areas of the economy.

Gordon backs up his qualitative argument with a wealth of data; charts and tables are given throughout the book. A few particularly important ones stand out. One is GDP growth rates, split into three periods:

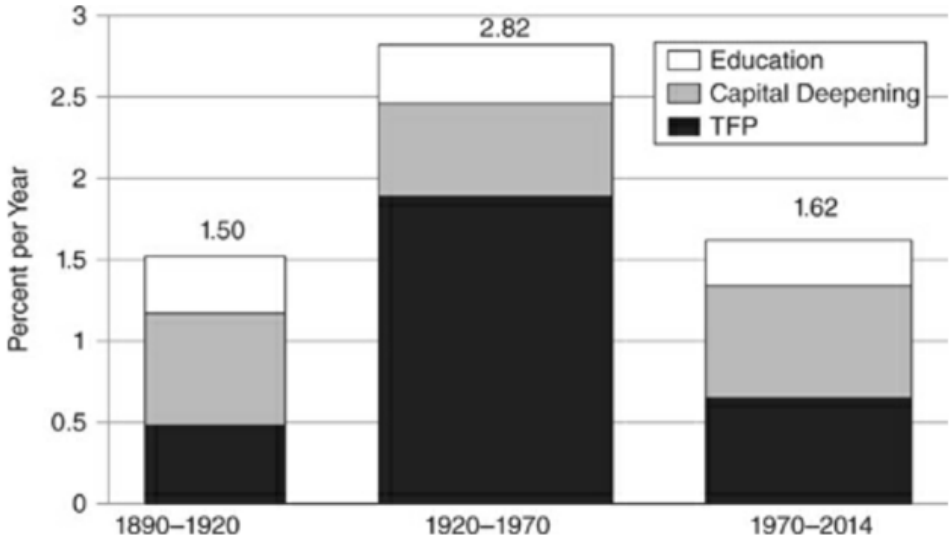


Figure 1–2. Average Annual Growth Rates of Output per Hour and Its Components, Selected Intervals, 1890–2014

Source: See Data Appendix.

The 1920–70 period stands out as having significantly higher growth than either before or after. Note that increases in capital and labor, including the improved education of the workforce, did not change much across these three periods. Therefore, the difference is

mostly from a combination of other factors. Economists refer to this unaccounted-for growth as “total factor productivity” (or TFP), and it is generally assumed to come from advances in technology, organization and management. Gordon shows a breakout of TFP by decade:

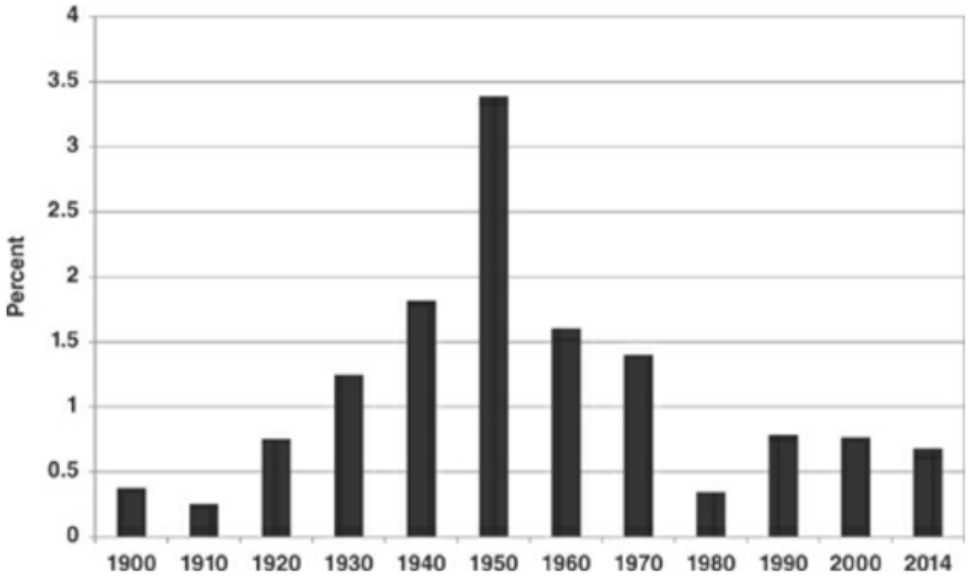


Figure 16–5. 10-Year Average Annual Growth in Total Factor Productivity, 1900–2014

Note: The average annual growth rate is over the ten years prior to year shown. The bar labelled 2014 shows the average annual growth rate for 2001–14.

If many crucial inventions, such as the sewing machine, electric light and motor, telephone, and automobile, were created in the period 1870–1920, then why was the higher growth not seen until the 1920–70 period? In brief, it took decades for many inventions to be widely distributed and adopted. The electric power industry was created in the 1880s, but electric light didn’t reach 80% penetration until 1940. In that same year, running water was only in 70% of homes, and indoor flush toilets in 60%. The automobile wasn’t widely adopted until Ford brought the price down in the 1910s. Radio was invented prior to 1920, but the radio *industry*, with networks and programming on the air, didn’t take off until the 1920s. Perhaps most significantly, electricity ultimately led to a revolution in manufacturing, but only after processes and entire factories and were redesigned to take advantage of its possibilities:

> The assembly line, together with electric-powered tools, utterly transformed manufacturing. Before 1913, goods were manufactured by craftsmen at individual stations that depended for power on steam engines and leather or rubber belts. The entire product would be crafted by one or two employees. Compare that with a decade later, when each worker had control of electric-powered machine tools and hand tools, with production organized along the Ford assembly-line principle. An additional aspect of the assembly line was that it saved capital, particularly “floor space, inventories in storage rooms, and shortening of time in process.”
>
> It is likely that electric power and the assembly line explain not just the TFP growth upsurge of the 1920s, but also that of the 1930s and 1940s. There are two types of evidence that this equipment capital was becoming more powerful and more electrified. First is the horsepower of prime movers, a data series available for selected years for different types of productive capital, and the second is kilowatt hours of electricity production.

Gordon also traces some of the causes back to the Depression and World War II. The New Deal promoted unionization, which led to rising wages and a shrinking work week, which motivated employers to invest in capital to substitute for labor. During WW2, the urgency and pressure of the threat spurred manufacturers to reach ambitious new levels of efficiency. Two famous examples were the Kaiser shipyards, which in a short period from 1942–43 reduced the time needed to build a Liberty freighter from eight months to a few weeks; and the Ford plant at Willow Run that, before the end of the war, was producing B-24 bomber planes at the rate 432 per month. The federal government helped increase capacity by heavily investing in machine tools, which *doubled* in number from 1940 to 1945. After the Depression and the war, the new capital equipment, processes, and knowledge didn't go away, and was turned to the purpose of peacetime production.

To understand the impact of the digital revolution, Gordon breaks out TFP growth into multiple periods, and isolates in particular the decade 1994–2004:

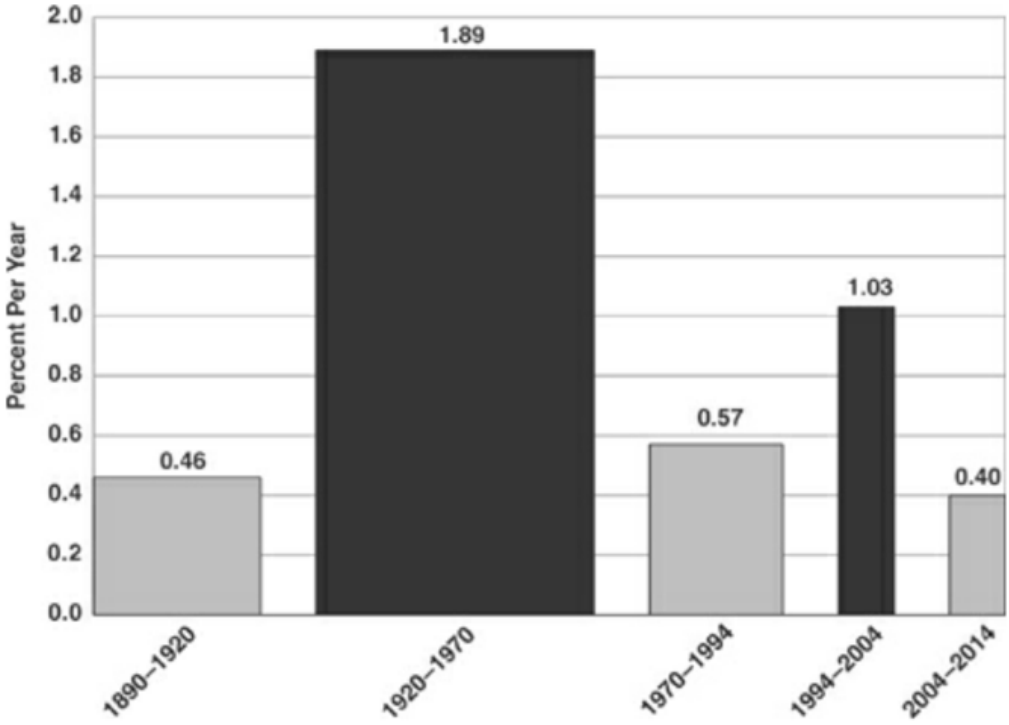


Figure 17–2. Annualized Growth Rates of Total Factor Productivity, 1890–2014

Source: Data underlying Figure 16–5.

He ultimately concludes that the digital revolution had a real but comparatively minor impact on the economy: its impact was lower in magnitude and shorter-lived than the combined impact of electricity, the automobile, and the other inventions of the previous period.

A common criticism of this type of analysis is that it relies on GDP, and GDP does not capture all improvements to the quality of life. In the Internet era, many of the best services, such as Google, Wikipedia, and YouTube, are free—an enormous consumer surplus that doesn't contribute to GDP. Gordon agrees with this point, but he argues that this has *always* been the case, and that in fact even *more* consumer surplus went unmeasured in the past, such as the value of free radio and television programs, the liberation provided by the automobile, or the

lives saved by penicillin. Thus, he argues, GDP mismeasurement can't explain the GDP growth slowdown.

What about the future? Gordon is not optimistic.

The trends of the early and mid-20th century are mostly played out. Our factories are already electrified, as are our homes. Everyone already has a car, a phone, and a toilet.

Even computing, Gordon thinks, has mostly given us its contribution already, citing the slowing of TFP growth after 2004. Business practices haven't changed much since we all got PCs, spreadsheets, and email. Retail hasn't changed much since computerized inventories and barcode scanners. Banking has already deployed ATMs, and the financial markets are already computerized. The smartphone has matured and plateaued, and after it, there isn't much that's new in consumer electronics. Even Moore's Law is slowing down.

The demographic changes of the 20th century that improved productivity also don't have much further to go. The share of the population living in urban areas hit 70% by 1960. High school graduation rates rose from 9% in 1910 to 77% by 1970, but aren't much higher today. The labor force participation rate of women rose dramatically after the war but peaked at 77% in 1999, and is now a bit lower.

What about new technologies? Gordon has surveyed the field and is unimpressed. Drug development is just getting more expensive. Robots aren't good enough to replace humans yet. 3D printing won't replace mass production. "Big data" is just a continuation of the trend towards more data and therefore is nothing new. Self-driving cars won't shorten people's commutes, and self-driving trucks won't replace all the functions of drivers, who also load and unload the trucks and stock store shelves. And besides, self-driving vehicles don't work yet anyway.

Finally, Gordon sees "headwinds" working against growth. The boomers are retiring. Debt is rising, including student debt (now over $1 trillion) and government debt (which is nearing 100% of GDP). And inequality is growing, so gains in GDP per capita will translate to smaller gains in median income.

Taking all factors into account, Gordon forecasts an average growth rate of just 0.3% in real median disposable income per person over the period 2015–2040, compared to 2.25% growth 1920–70 and 1.46% 1970–2014.

Can we do anything about it?

In a postscript, Gordon offers a long list of policy prescriptions, including:

- Higher taxes on the very rich
- Increase in the minimum wage
- Expanded Earned Income Tax Credit
- Mass pardoning to reduce the incarceration rate
- Drug legalization
- Public preschool
- Financing public school with statewide rather than local sources to reduce school inequality
- Publicly funded college, paid for by higher income taxes on college graduates
- Restrictions on patents and copyrights
- Reduced occupational licensing
- Reduced zoning and land-use regulations
- More high-skill immigration
- Eliminating tax deductions
- A carbon tax

These prescriptions, however, are only palliative. Fundamentally, he sees the growth slowdown as so natural and inevitable that it is not even to be lamented:

> What is remarkable about the American experience is not that growth is slowing down but that it was so rapid for so long…. the rise and fall of growth are inevitable when we recognize that progress occurs more rapidly in some time intervals than in others. There was virtually no economic growth for millennia…. American growth slowed down after 1970 not because inventors had lost their spark or were devoid of new ideas, but because the basic elements of a modern standard of living had by then already been achieved along so many dimensions, including food, clothing, housing, transportation, entertainment, communication, health, and working conditions.
>
> The 1870–1970 century was unique: Many of these inventions could only happen once, and others reached natural limits.

There is no murder investigation, because there was no murder. The victim died of natural causes.

I found the first two parts of the book very valuable. The survey of improvements to the American standard of living over the last 150 years is comprehensive, described in vivid detail, and quantified with more than enough charts and tables. As an overview of progress, it is excellent. If the phrase “standard of living” (which I have always found vapid) is an empty term to you, this book will fill it with many colorful examples.

But I find Gordon’s vision for the future strangely lacking in imagination, and his complacent acceptance of low growth disappointing. Contrast this with another proponent of the stagnation hypothesis, Peter Thiel. Thiel’s view seems to be not that stagnation was natural or inevitable, but that we dropped the ball, and that we should have had flying cars by now.

In considering potential future technological developments in one of the final chapters, Gordon only considers four: “medical, small robots and 3D printing, big data, and driverless vehicles.” And he doesn’t actually analyze the *future* of these technologies, but instead looks only at their *present*. He points out that at one robot competition, the robots had trouble standing up; that robotic reasoning is limited, and that they can enter an error state if presented with an unfamiliar situation; that self-driving cars only work at low speeds and depend on detailed maps. Given that this comes after several hundred pages detailing over a century of breakthrough inventions, there is surprisingly little recognition that technologies almost always go through an immature stage in which they do not work reliably, and that eventually these problems are solved through engineering iteration. Gordon also shows little imagination for the economic potential of self-driving vehicles, assuming that we will continue to use cars and trucks in the same way, only without drivers. He does not seem to consider that self-driving vehicles have the potential to fundamentally alter land use patterns and cargo logistics, even though he described in earlier chapters how motor vehicles did exactly these things.

There is no serious consideration that we might solve other grand challenge problems, such as curing cancer or heart disease the way we cured infectious diseases and vitamin deficiencies, or generating cheap and safe nuclear power. The possibility that we might make new breakthroughs in areas such as genetic engineering or quantum computers is not raised.

I find this approach difficult to understand. It seems to treat centuries of breakthroughs as … a fluke.

That said, I think this is overall a very valuable book that every serious student of progress should read. It is an excellent survey of American growth, and it paints the stagnation picture clearly.

And even its predictions for the future are valuable if taken, not as a prophecy, but as a warning. A 0.3% growth rate *is* the future—*if* we have no more breakthroughs. We have, in some ways, been coasting on the achievements of the past. Gordon shows us that all inventions, no matter how great, eventually mature and plateau. Restoring high growth will require new fundamental inventions and possibly entire new fields of science. I'm becoming increasingly convinced that this is where progress studies should focus.

Moloch games

tl;dr: This post suggests a direction for modelling Molochs. The main thing this post does is to rename the concept of "potential games" (an existing concept in game theory) to "Moloch games" to suggest an interpretation of this class of games. I also define "the preferences of a Moloch" to generalize that notion (the preferences may be intransitive).

This post assumes that you have familiarity with game theory and the concept of a Moloch.

What do group dynamics want? If a society/group (the Moloch) wants things that are different from what the individuals want, how can we assign preferences or a utility function to that society/group that doesn't model what would be good for the aggregated preference of the individuals of the group, but that describes what the group dynamics is actually "trying" to achieve (even against the interest of its individual members)? Here is a suggested answer.

Intuition. A **Moloch game** is a game such that there is a utility function U_M, called "**the Moloch's utility function**", such that if the agents behave individually rationally, then they collectively behave as a "Moloch" that controls all players simultaneously and optimizes U_M. In particular, the Nash equilibria correspond to local optima of U_M.

Not all games are Moloch games.

Definition. A game with finite number of players and for each player i a strategy space X_i is a *cardinal* **Moloch game** (in the game theory literature, a *cardinal potential game*), if there is a utility function $U_M : (\prod_{i \in N} X_i) \to R$ such that for all players i, and all strategies s_{-i} for the other players,

$$u_i(s_i', s_{-i}) - u_i(s_i'', s_{-i}) = U_M(s_i', s_{-i}) - U_M(s_i'', s_{-i}) \qquad \forall s', s'' \in X_i$$

Intuitively, if you take any strategy-profile for all the players, and adjust the strategy of one player, then the Moloch's utility will increase/decrease by the same amount as the utility function of that particular player. Hence, intuitively, every player behaves always *as if* they are optimizing the Moloch's utility function.

The definition for an *ordinal* Moloch game replaces the condition with

$$u_i(s_i', s_{-i}) \geq u_i(s_i'', s_{-i}) \quad \text{iff} \quad U_M(s_i', s_{-i}) \geq U_M(s_i'', s_{-i}) \qquad \forall s', s'' \in X_i$$

Intuitively, U_M represents the Moloch's preferences ordinally but not cardinally. Obviously, cardinal Moloch games are also ordinal Moloch games.

Example. The prisoners dilemma:

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	1, 1	-1, 2
	Defect	2, -1	0 , 0

We will show that this is a cardinal Moloch game, by just computing the cardinal utility function and showing that there are no inconsistencies:

How to compute the cardinal utility function of the Moloch: Pick an arbitrary strategy profile to have utility 0 (I take Defect, Defect). Then iteratively compute the utility of rows and columns by just applying the constraint that the definition gives: Compute the difference in utility *of the player whose row/column you're moving along* (i.e. player 2 for the rows, player 1 for the columns) of each cell in the row/column from a cell of which you know the value of U_M. In this case, we know $U_M(D, D) = 0$. So compute $U_M(D, C)$ as $U_M(D, D) + u_2(D, C) - u_2(D, D)$ which equals $0 + -1 - 0 = -1$. Similar for $U_M(C, D)$. For $U_M(C, C)$, there are two ways to compute it: Using player 1's utility function and $U_M(D, C)$ or player 2's utility function and $U_M(C, D)$. If these two give different answers, then the game is not a cardinal Moloch game.

Here is the **cardinal utility function of the Moloch for the prisoner's dilemma** (The above algorithm gives a utility function that is unique up to translations) :

	Cooperate	Defect
Cooperate	-3	-1
Defect	-1	0

Intuition. In this case, even though all players prefer Cooperate, Cooperate over Defect, Defect, the Moloch prefers the opposite. This corresponds to the fact that it is individually rational for the players to Defect. This Moloch utility function captures the "preferences of the group dynamics" as opposed to the preferences of the individuals. (It is obviously very different from the notion of "aggregate preferences" or "welfare").

A Moloch game assumes in some sense that "the Moloch has transitive preferences". We can generalize to Molochs with possibly intransitive preferences (I don't know of this being defined this way in the literature on potential games):

Definition. Let Γ be a game with a finite number of players, each of which has a preference relation $\preceq_i$ over the strategy space $X = \prod_i X_i$ (by default derived from a utility function u_i). Then the **Moloch's preferences** $\preceq_M$ are defined as the preference relation satisfying for all players i, and all strategy profiles s_{-i} for the other players:

$$(s_i', s_{-i}) \preceq_i (s_i'', s_{-i}) \quad \text{iff} \quad (s_i', s_{-i}) \preceq_M (s_i'', s_{-i}) \qquad \forall s_i', s_i'' \in X_i$$

Observation. These preferences are always incomplete (intuitively, the Moloch doesn't have an opinion on the comparison between different players changing their strategies, because it doesn't have this information: players individually make choices given their options). They may be either transitive or intransitive. I'll say a Moloch's preferences are rational if they are transitive (neglecting the usual requirement of completeness).

Just to show that the concepts are what they should be:

Lemma. Any game whose Moloch has transitive preferences is an ordinal Moloch game. Any ordinal Moloch game has a Moloch with transitive preferences.

Proof. For any transitive relation on a space there is a real-valued function on it that is consistent with that relation. The other direction follows directly.

Intuition. If the Moloch has transitive preferences, then the Moloch knows what it wants and the game will have a pure Nash equilibrium (there is a theorem that formalizes this). Conversely, if the Moloch has intransitive preferences, then the Moloch doesn't know what it wants and the game will tend to have cycles (not all of them will because the players might want to move out of them into a "transitive region" of the Moloch's preferences).

I won't show this here, but the literature on potential games (cf. the thing I am calling Moloch games), these are examples:

Games with rational Molochs (i.e. Moloch games / potential games):

- Prisoner's dilemma
- Battle of the sexes
- Coordination game
- Game of Chicken

Games with irrational Molochs (i.e. not Moloch games / potential games):

- Matching pennies
- Rock paper scissors

I probably won't spend much more time on this, but here is a suggestion for taking this as a starting point to modelling Molochs:

- Check if various informal ideas about Molochs can be phrased in this language. Check if the language is satisfying to talk about actual Molochs.
- Look at the literature on potential games to see if it contains much insight. Make a dictionary of concepts named in the terminology of the ontology we're interested in (similar to how I renamed "potential game" to "Moloch game") to make this literature an "efficiently queryable database" for insights into Molochs.
- I suspect that there might be ideas to be had about Moloch games that aren't treated there, because as far as I know, potential games were developed mostly as a trick to make computations easier, not as a conceptual tool for thinking about Molochs, societal inadequacy and so forth. It's plausible that certain obvious questions haven't been asked about them for this reason. Try to actually model Molochs this way and see if these definitions allow us to answer questions we want to ask about them. Use this as a stepping stone and see where it is unsatisfying. Build on top of that to push the analysis further.

Feel free to contact me if you want to think about this.

Some reading:

Flows and Decompositions of Games: Harmonic and Potential Games. In the language of this post: decomposing a game into a "rational part" of the Moloch, and an irrational deviation from it. Finding the "closest rational Moloch" of a game.

Some further ideas and questions to ask:

- Can real world societies be decomposed into multiple Molochs? In the style of the "Flows and Decompositions of Games" paper, it wouldn't have to be a decomposition in terms of subgroups of players, but of "aspects of the game-theoretic interaction". (e.g. an individual might simultaneously be part of a "capitalism Moloch" and a "politics Moloch"). Maybe Molochs can be approximately decomposed.
- Is there a notion of "Moloch game" for sequential games? Games with limited information? (The potential game literature probably has asked analogous questions).

Have the lockdowns been worth it?

There's been a lot of discussion about whether the pandemic lockdowns have been worth it. However, much of the reasoning that we've seen has been very motivated and un-nuanced in a way that for us has distorted a lot of the information.

So this is not a thread for taking a position on that. This is a thread for raising *individual considerations* that are relevant for thinking about the question "Have the pandemic lockdowns, in general, been worth it?"

Every answer to this thread should analyze a single belief that is relevant to whether pandemic lockdowns have been worth it, such as

- "No lockdown would have led to 300,000 additional loss of QALYs in the UK"
- "GDP $1b lower than the counterfactual in Germany"
- "The West will develop a vaccine in 6 months"
- "A year in lockdown is worth 70% of a normal year and 85% of a covid year without lockdown"
- "The impetus to try remote work has led to changes in workplace norms worth $1 trillion over the next decade"

and provide relevant facts, data and information available about that factor.

Answers in this thread should not attempt to take a position on the overall question. **Answers that take a position on the overall question will be deleted.** This thread will live up to the virtue of holding off on proposing solutions.

(By 'lockdown' we refer to the thing that the US, UK and China have been doing, and what Sweden didn't. There is naturally a lot of variation between countries, so this cannot have a canonical answer. If your consideration only applies to a small number of countries, that is fine.)

Dutch-Booking CDT: Revised Argument

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

This post has benefited greatly from discussion with Sam Eisenstat, Caspar Oesterheld, and Daniel Kokotajlo.

Last year, I wrote a post claiming there was a Dutch Book against CDTs whose counterfactual expectations differ from EDT. However, the argument was a bit fuzzy.

I recently came up with a variation on the argument which gets around some problems; I present this more rigorous version here.

Here, "CDT" refers -- very broadly -- to using counterfactuals to evaluate expected value of actions. It need not mean physical-causal counterfactuals. In particular, TDT counts as "a CDT" in this sense.

"EDT", on the other hand, refers to the use of conditional probability to evaluate expected value of actions.

Put more mathematically, for action $a \in A$, EDT uses $E(U|Act = a)$, and CDT uses

$E(U|do(Act = a))$. I'll write $edt(a)$ and $cdt(a)$ to keep things short.

My argument could be viewed as using Dutch Books to formalize Paul Christiano's "simple argument" for EDT:

> Suppose I am faced with two options, call them L and R. From my perspective, there are two possible outcomes of my decision process. Either I pick L, in which case I expect the distribution over outcomes P(outcome|I pick L), or I pick R, in which case I expect the distribution over outcomes P(outcome|I pick R). In picking between L and R I am picking between these two distributions over outcomes, so I should pick the action A for which E[utility|I pick A] is largest. There is no case in which I expect to obtain the distribution of outcomes under causal intervention P(outcome|do(I pick L)), so there is no particular reason that this distribution should enter into my decision process.

However, I do not currently view the argument as favoring EDT over CDT! Instead it supports the weaker claim that the two had better agree. Indeed, the Troll Bridge problem strongly favors *CDT whose expectations agree with EDT* over EDT. So, this is intended to provide a strong constraint on a theory of (logical) counterfactuals, not necessarily abolish the need for them. (However, the constraint is a strong one, and it's worth considering the possibility that this constraint is *all we need* for a theory of counterfactuals.)

The Basic Argument

Consider any one action $a \in A$ for which $edt(a) \neq cdt(a)$, in some decision problem. We wish to construct a modified decision problem which Dutch-books the CDT.

My argument requires an assumption that the action a is assigned nonzero probability.

This is required to ensure edt(a) is defined at all (since otherwise we would be conditioning on a probability zero event), but also for other reasons, which we'll see later on.

Anyway, as I was saying, we wish to take the decision problem which produces the disagreement between edt(a) and cdt(a), and from it, produce a new decision problem which is a Dutch book.

The new decision problem will be a two-step sequential decision problem. *Immediately before* the original decision, the bookie offers to sell the agent the following bet B, for a price of 2d utilons. B is a bet conditional on a, in which the buyer is betting against cdt(a)'s expectation and in favor of edt(a)'s expectation. For example:

B: *In the case that* $A = a$*, the seller of this certificate owes the purchaser of this certificate* $(\mathrm{cdt}(a) - U) \cdot s$*, where* s *is the signum* $\frac{|\mathrm{cdt}(a) - \mathrm{edt}(a)|}{\mathrm{cdt}(a) - \mathrm{edt}(a)}$

The key point here is that *because the agent is betting ahead of time*, it will evaluate the value of this bet according to the conditional expectation $E(U|\mathrm{Act} = a)$.

If $\mathrm{cdt}(a) > \mathrm{edt}(a)$, so $s = 1$, then the value of B in the case that $\mathrm{Act} = a$ is $\mathrm{cdt}(a) - U$.

The expectation of this is $\mathrm{cdt}(a) - \mathrm{edt}(a)$, which again, we have supposed is positive. So the overall expectation is $P(\mathrm{Act} = a) \cdot (\mathrm{cdt}(a) - \mathrm{edt}(a))$. Setting d low enough ensures that the agent will be happy to take this bet. Similarly, if $\mathrm{edt}(a) > \mathrm{cdt}(a)$, the value of the bet ends up being $P(\mathrm{Act} = a) \cdot (\mathrm{edt}(a) - \mathrm{cdt}(a))$ and the agent still takes it for the right price.

Now, the second stage of our argument. *As the agent is making the decision* the bookie again makes an offer. (In other words, we extend the original set of actions A to contain twice as many valid actions; half in which we accept, half in which we don't accept.) The new offer is this: "I will buy the bet B from you for d utilons."

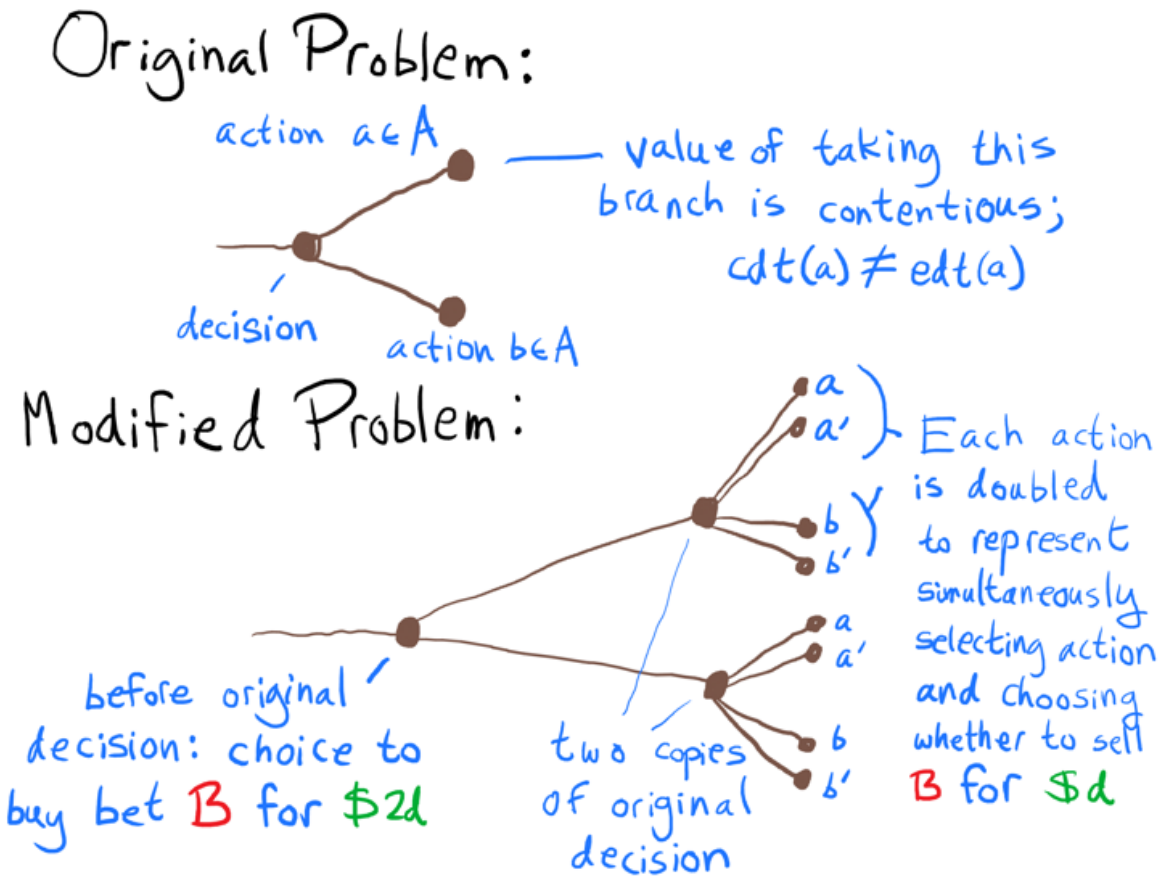
Now, since the agent is reasoning *during* its action, it is evaluating possible actions *according to* cdt(a)*;* so its evaluation of the bet will be different. Here, the argument splits into two cases:

- When considering the action a, the bet's expected value is cdt(a) − cdt(a), which is zero. So the agent prefers the new action a' which is like a except B is sold back to the bookie for d utilons.
- When considering any other action, the bet is worth zero *automatically*, since it only pays out anything when Act = a. So, the agent will gladly take the bookie's payment of d to sell the bet back.

So the result is the same in either case -- CDT recommends selling B back to the bookie no matter what.

The agent has paid 2d to buy B, and gotten only b when selling back. Buying and selling the contract cancel each other out. So the agent is down d utilons for no gain!

Here is an illustration of the entire Dutch Book:



Assumptions

Non-Zero Probability of a

A really significant assumption of this argument is that actions are given nonzero probability -- particularly, that the target action a has a nonzero probability. This assumption is important, since the initial evaluation of B is $P(\text{Act} = a) \cdot |\text{cdt}(a) - \text{edt}(a)|$.

If the probability of action a were zero, there would be no price the agent would be willing to pay for the bet.

The assumption is also required in order to guarantee that edt(a) is well-defined -- although we could possibly use tricks to get around that, specifying a variant of EDT which defines some expectation in all cases.

Many of my arguments for CDT=EDT rest on this assumption, though, so it isn't anything new. It seems to be a truly important requirement, rather than an artefact of the argument.

There are many justifications of the assumption which one might try to give. I have often invoked epsilon-exploration; that is, the idea that some randomness needs to be injected into an agents actions in order to ensure that it can try all options. I don't like invoking that as much as I used to. I might make the weaker argument that agents should use the chicken rule, IE, refuse to take any action which they can prove they take. (This can be understood as weaker than epsilon-exploration, because epsilon-exploration can be implemented by the epsilon-chicken rule: take any action which you assign probability less than epsilon to.) This rule ensures that agents can never prove what they do (so long as they use a sound logic). We can then invoke the non-dogmatism principle, which says that we should never assign probability 0 to a possibility unless we've logically refuted it.

Or, we could invoke a free-will principle, claiming that agents should have the subjective illusion of freedom.

In the end, though, what we have is an argument that applies if and only if a has nonzero probability. All the rest is just speculation about how broadly this argument can be applied.

An interesting feature of the argument is that *the less probable action* a *according to the agent, the less money we can get by Dutch-booking them on discrepancies between* cdt(a) *and* edt(a). This doesn't matter for traditional Dutch Book arguments -- any sure loss is considered a failure of rationality. However, if we take a logical-induction type approach to rationality, smaller Dutch Books *are less important* -- boundedly rational agents are expected to lose *some* money to Dutch Books, and are only trying to avoid losing too much.

So, one might consider this to be a sign that, in some hypothetical bounded-rationality approach to decision theory, lower-probability actions would be allowed to maintain larger discrepancies between edt(a) and cdt(a), and maintain them for longer.

Probabilities of Actions in the Modified Problem

A trickier point is the way probabilities carry over from the original decision problem to the modified problem. In particular, I assume the underlying action probabilities do not change. Yet, I split each action in two!

One justification of this might be that, for agents who choose according to CDT, it *shouldn't* change anything -- at the moment of decision, the bet B is worth nothing, so it doesn't bias actions in one direction or another.

Ultimately, though, I think this is just part of the problem setup. Much like money-pump arguments posit magical genies who can switch anything for anything else, I'm positing a bookie who can offer these bets without changing anything. The argument -- if you choose to accept it -- is that the result is disturbing in any case. It does not seem likely that an appealing theory of counterfactuals is going to wriggle out of this specifically by denying the premise that action probabilities remain the same.

Note, however, that it *is not* important to my argument that none of the *new* actions get assigned zero probability. It's only important that the sum of $P(a)$ and $P(a')$ in the new problem equals the original decision problem's $P(a)$.

Counterfactual Evaluation of Bets

Another assumption I didn't spell out yet is the interaction of bet contracts with *counterfactual* evaluations.

I assume that counterfacting on accepting the bet does not change probabilities of other things, such as the probability of the actions. This could be a large concern in general -- taking a conditional bet on Act = a might make us want to choose a on purpose, in order to cash in on the bet. This isn't a problem in this case, since the agent later evaluates the bet to be worth nothing. However, that doesn't necessarily mean it's not an issue *according to the counterfactual evaluation*, which would chance the perceived value of B. Or, even more problematic, the agent's counterfactual expectations might say that taking the bet would result in some very negative event -- making the agent simply refuse. So the argument definitely assumes "reasonable counterfactual evaluations" in some sense.

On the other hand, this kind of reasoning is very typical for Dutch Book arguments. The bets are grafted onto the situation without touching any of the underlying probabilities -- so, e.g., you do not normally ask "is accepting the bet against X going to make X more probable?".

Handling Some Possible Objections

Does the Bookie Cheat?

You might look at my assumptions and be concerned that the bookie is cheating by using knowledge which the agent does not have. If a bookie has insider information,

and uses *that* to get a sure profit, it doesn't count as a Dutch Book! For example, if a bookie knows all logical facts, it can money-pump any agent who does not know all logical facts (ie, money-pump any fixed computable probability distribution). But that isn't fair.

In this case, one might be concerned about *the agent not knowing its own action*. Perhaps I'm sneaking in an assumption that agents are uncertain of their own actions, and then Dutch-booking them by taking advantage of that fact, via a bookie who can easily predict the agent's action.

To this I have a couple of responses.

The bookie does not know the agent's choice of action. The bookie's strategy doesn't depend on this. In particular, note the disjunctive form of the argument: *either* the agent prefers a, in which case B is worthless for one reason, *or* the agent prefers a different action, in which case B is worthless for a different reason. The bookie is setting things up so that it's safe no matter what.

The agent knows everything the bookie knows, from the beginning. All the bookie needs in order to implement its strategy is the the values of cdt(a) and edt(a), and, in order to set the price d, the probability of the action a.These are things which the agent also knows.

The Agent Justifiably Revises Its Position

Another critique I have received is that it makes perfect sense that the agent takes the bet at the first choice point and later decides against it at the second choice point. The *agent* has gained information -- namely, when considering an action, the agent knows it will take that action. This extra information is being used to reject the bet. So it's perfectly reasonable.

Again I have a couple of responses.

The agent does not learn anything between the two steps of the game. There is no new observation, or additional information of any kind, between the step when the bookie offers B and the step when the bookie offers to buy B back. As the agent is evaluating a particular action, it *does not* "know" it will carry out that action -- it is only considering what would happen *if* it carried out that action!

Even if the agent did learn something, it would not justify being Dutch-booked. Consider two-stage games in which an agent is offered a bet, then learns some information, and then given a choice to sell the bet back for a fee. It is perfectly reasonable for an agent to *in some cases* sell the bet back. What makes a Dutch book, however, is if the agent *always sells the bet back*. It should never be the case that an agent *predictably* won't want the bet later, no matter what it observes. If that were the case (as it is in my scenario), the agent should not have accepted the bet in the first place. It's critical here to again note that the agent prefers to sell back the bet *for every possible action* -- that is, the original actions are *always* judged worse than their modified copies in which the sell-back deal is taken. So, even if we think of the agent

as "learning" which action it selects when it evaluates selecting an action, we can see that it decides to sell back the bet no matter what it learns.

But Agents *WILL* Know What They'll Do

One might say that the argument doesn't mean very much at all in practice because *from* the information it knows, the agent *should* be able to derive its action. It knows how it evaluates all the actions, so, it should just know that it takes the argmax. This means the probability of the action actually taken is 1, and the probability of the rest of the actions is zero. As a result, my argument would only apply to the argument actually taken -- and a CDT advocate can easily concede that cdt(a) = edt(a) *when* a *is the action actually taken*. It's other actions that one might disagree about. For example, in Newcomb, classical physical-causality CDT two-boxes, and agrees with EDT about the consequences of two-boxing. The disagreement is only about the value of the *other* action.

(Note, however, that the CDT advocate is still making a significant concession here; in particular, this rules out the classic CDT behavior in Death and Damascus and many variations on that problem. I don't know exactly how a classic physical-causality CDT advocate would maintain such a position.)

There are all kinds of problems with the agent knowing its own action, but a CDT advocate can very naturally reply that these should be solved with the right counterfactuals, not by ensuring that the agent is unsure of its actions (through, e.g., epsilon-exploration).

I'll have more to say about this objection later, but for now, a couple of remarks.

First and foremost, yeah, my argument doesn't apply to actions which the agent knows it won't take. I think the best view of the phenomenon here is that, if the agent *really does* know exactly what it will do, then yeah, the argument *really does* collapse to saying its evidential expectations should equal its counterfactual expectations for that one action. Which is like saying that, if $P(A) = 1$, then we had better have $P(X|do(A)) = P(X)$ -- counterfacting on something true should never change anything.

Certainly it's quite common to think of agents as knowing exactly what they'll do; for example, that's how backwards-induction in game theory works. And at MIRI we like to talk about problems where the agent can know exactly what it can do, because these stretch the limits of decision theory.

On the other hand, realistic agents probably mostly *don't* know with *certainty* what they'll do -- meaning my argument will usually apply in practice.

The agent might not follow the recommendations of CDT. Just because a CDT-respecting agent would definitely do a specific thing given all the information, *does not* mean that we have to imagine the agent in my argument knowing exactly what it will do. The agent in the argument might not be CDT-respecting.

Here on LessWrong, and at MIRI, there is often a tendency to think of CDT or EDT *as the agent* -- that is, think of agents as instances of decision theories. This is a point of

friction between MIRI's way of thinking and that of academic philosophy. In academic philosophy, the decision theory need not be an algorithm the agent is actually running (or indeed, could ever run). A decision theory is a *normative theory* about what an agent *should do*. This means that CDT, as a normative theory, can produce recommendations for non-CDT agents; and, we can judge CDT on the correctness or incorrectness of those recommendations.

Now, I think there are some advantages to the MIRI tendency -- for example, thinking in this way brings logical uncertainty to the forefront. However, I agree nonetheless with making a firm distinction between the *decision theory* -- a normative requirement -- and the *decision procedure* -- a real algorithm you can run, which obeys the normative requirement. The logical induction algorithm vs the logical induction criterion illustrates a similar idea.

Academic decision theorists extend this idea to the criticism of normative principles -- such as CDT and EDT -- for their behavior in scenarios which an agent would never get into, if it were following the advice of the respective decision theory. This is what's going on in the bomb example Will MacAskill uses. (Nate Soares argues against this way of reasoning, saying "decisions are for making bad outcomes inconsistent".)

If we *do* endorse the idea, this offers further support for the argument I'm making. It means we get to judge CDT for recommending that an agent accept a Dutch-book, even if the scenario depends on uncertainty over actions which a CDT advocate claims a CDT-compliant agent does not have.

This is particularly concerning for CDT, because this kind of argument is used especially to defend CDT. For example, it's hard to justify smoking lesion as a situation which a CDT or EDT agent could actually find itself in; but, a CDTer might reply, a decision theory needs to offer the right advice to a broad variety of agents. So CDT is already in the business of defending normative claims about non-CDT agents.

Dynamic Consistency

One might object: the argument simply illustrates a dynamic inconsistency in CDT. We already *know* that *both* CDT and EDT are dynamically inconsistent. What's the big deal?

Let me make some buckshot remarks before I dive into my main response here:

- First, this isn't just any old dynamic inconsistency. This is a Dutch Book. Dutch Books have a special status in the foundations of Bayesianism. So, one might consider this to be more concerning than mere dynamic inconsistency.
- Second, dynamic consistency is still bad. We may still take examples of dynamic inconsistency as counting against a normative theory, and seek a theory which is dynamically consistent in a fairly broad range of cases.
- Third, I think EDT really does have a dynamic-consistency advantage over CDT, and my argument is just one example of that.

This third point is the one I want to expand on.

In decision problems where the payoff depends *only on actions actually taken*, not on your policy, there is a powerful argument for the dynamic consistency of EDT:

Think of the entire observation/action history as a tree. Dynamic consistency means that at earlier points in the tree, the agent does not prefer for the decisions of later

selves to be different from what they will be. The restriction to actions (not policies) mattering for payoffs means this: selecting one action rather than another changes which branch we go down in the tree, but *does not* change the payoffs of other branches in the tree. This means that *even from a perspective beforehand*, an action can only make a difference down the branch where it is taken -- no spooky interactions across possible worlds. As a result, thinking about possible choices ahead of time, the contribution to early expected utility is *exactly* the expected utility that action will be assigned later at the point of decision, times the probability the agent ends up in that situation in the first place. Therefore, the preference about the decision must stay the same.

So, EDT is a dynamically consistent choice when actions matter but policy does not.

Importantly, this is **not** a no-Newcomblike-problems condition. It rules out problems such as counterfactual mugging, transparent Newcomb, Parfit's hitchhiker, and XOR Blackmail. However, it **does not** rule out the original Newcomb problem. In particular, we are highlighting the inconsistency where CDT wishes to be a 1-boxer, and similar cases.

Now, you *can* make a very similar argument for the dynamic consistency of CDT, if you define dynamic consistency based on counterfactuals: would you prefer to counterfact on your future self doing X? For Newcomb's problem, this gets us back to consistency -- for all that the CDT agent *wishes it could be EDT*, it would have no interest in the point-intervention that makes its future self one-box, for the usual reason: that would not cause Omega to change its mind.

However, this definition seems not to capture the most useful notion of dynamic consistency, since the same causal CDT agent would *happily* precommit to one-box. So I find the EDT version of the argument more convincing. I'm not presently aware of a similar example for EDT -- it seems Omega needs to consider the policy, not just the action really taken, in order to make EDT favor changing its actions via precommitments.

More on Probability Zero Actions

As I've said, my argument depends on the action a having nonzero probability. EDT isn't well-defined otherwise; how do you condition on a probability-zero event? However, there *are* some things we could try in order to get around the problem of division by zero - filling in the undefined values with sensible numbers. For example, we can use the Conditional Oracle EDT which Jessica Taylor defined, to "fill in" the otherwise-undefined conditionals.

However, recall I said that the argument was blocked for *two* reasons:

- EDT not being defined.
- The bet conditional on a is worth nothing if a has probability zero.

So, if we've somehow made edt(a) well-defined for probability-zero a, can we patch the second problem?

We can try to flip the argument I made around: for probability zero actions, we *pay* the agent to take on a bet (so that it will do it even though it's worthless). Then later, we *charge* the agent to offload the bet which it now thinks is unfavorable.

The problem with this argument is, if the agent doesn't take action a anyway, then the conditional bet will be nullified regardless; we can't force the agent into a corner where it prefers to nullify the bet, so, we don't get a full Dutch Book (because we can't guarantee that we make money off the agent -- indeed, we would only make money if the agent ends up taking the action which it previously assigned probability zero to taking).

However, we *do* get a *moderately* damning result: limiting attention to just the action a and the new alternative a′ which both does a and also pays to cancel the bet, **CDT strictly prefers that the agent be Dutch Booked* rather than just do a. This seems pretty bad: CDT isn't actually recommending taking a Dutch Book, *BUT*, it would rather take a Dutch Book than take an alternative which is otherwise the same but which does not get Dutch Booked.

So, we can still make a moderately strong argument against divergence between counterfactuals and conditionals, even if actions have probability zero. But not a proper Dutch Book.

A Few Words on Troll Bridge

At the beginning of this, I said that I didn't necessarily take this to be an argument for EDT over CDT. In the past, I've argued this way:

1. "Here's some argument that CDT=EDT."
2. "Since EDT gets the answer more simply, while CDT has to posit extra information to get the same result, we should prefer EDT."

However, this argument has at least two points against it. First, the arguments for CDT=EDT generally have some assumptions, such as nonzero probability for actions. CDT is a strictly more general framework when those conditions are not met. Theories of rational agency should be as inclusive as possible, when rationality does not demand exclusivity. So one might still prefer CDT.

Second, as I mentioned at the beginning of this post, the Troll Bridge problem strongly favors CDT over EDT. Counterintuitively, it's perfectly possible for a CDT agent to keep its counterfactual expectations exactly in agreement with its conditional expectations, and yet get Troll Bridge right -- even though we are doomed to get Troll Bridge wrong if we *directly* use our conditional expectations. Insisting on a distinction "protects" us from spurious counterfactual reasoning. (I may go over this phenomenon in more detail in a future post. But perhaps you can see why by reviewing the Troll Bridge argument.)

So, my *current* take on the CDT=EDT hypothesis is this:

1. We should think of counterfactuals as having real, independent truth. In other words, cdt(a) does not *reduce to* $edt(a)*. Counterfactual information tells us something above and beyond probabilistic information.

2. Counterfactuals are subjective in the same way that probabilities are subjective. The "independent truth" of counterfactuals does not mean there is one objectively correct counterfactual which every agent is normatively required to agree with. So there doesn't need to be a grand theory of logical counterfactuals -- there are many different subjectively valid beliefs.
3. However, as with probability theory, there are important notions of coherence which constrain subjective beliefs. In particular, counterfactual beliefs should almost always equal conditional beliefs, at least when the antecedent has positive probability.
4. Furthermore, conditional beliefs act a whole lot more like stereotypical CDT counterfactuals than most people seem to give them credit for. Something can't correlate with your action unless it contains *information you don't have* about your action. This is a high bar to pass, and will typically not be passed in e.g. twin prisoner's dilemma. (So, to solve these problems requires something further, e.g. updatelessness, Löbian handshakes, ???).

This is not a strongly held view, but it is the view that has made the most sense of counterfactual reasoning for me.

As I've mentioned in the past, the CDT=EDT hypothesis is almost the most boring possible answer to the question "how do (logical) counterfactuals work?" -- it doesn't do very much to help us solve interesting decision problems. If we factor decision theory into the two parts (1) "What are the (logical) counterfactuals?" (2) "How do we use counterfactuals to make decisions?" then I see the CDT=EDT hypothesis as a solution to (1) which shoves an awful lot of the interesting work of decision theory into (2). IE, to solve the really interesting problems, we would need logically-updateless UDT or even more exotic approaches.

In particular, for variants of Newcomb's problem where the predictor is quite strong but doesn't know *as* much as the agent does about what the agent will choose, this post implies that TDT *either* two-boxes, or, is vulnerable to the Dutch Book I construct. This is unfortunate.

Conclusion

Frankly, I find it somewhat embarrassing that I'm still going on about CDT vs EDT. After all, Paul Christiano said, partially in response to my own writing which he cited:

> There are many tricky questions in decision theory. In this post, I'll argue that the choice between CDT and EDT isn't one of them.

I wish I could say this will be my final word on the subject. The contents of this post do feel quite definitive in the sense of giving a settled, complete view. However, the truth is that it only represents my view as of November or early December of 2019. Late December and early January saw some developments which I'm excited to work out further and post about.

Thoughts on ADHD

A few different things that have occurred while investigating. It seems weird to me there are lots of posts about akrasia and productivity but not adhd.

1. High Rejection Sensitivity makes my attention in general more flinchy. Rather than just explicit rejection, you can think of this as a high sensitivity to negative feelings in general with rejection just being of extra highlight. This becomes a bit of a closed loop if I'm avoid noticing how flinchy my attention is because that would also imply unpleasant things.

2. Imbalanced novelty seeking dovetails really nicely with 1 because it gives me more buckets to run to. Find something I don't like in one tab? There is another tab a muscle memory shortcut away. It also encourages multitasking in the sense of listening to music or eating at the same time as I do other things. Novelty feels somewhat additive, so that 3 minor novelties can combine to make me feel 'stimulated enough.' In the same way that insulin insensitivity is a problem this feels a bit like a sort of dopamine insensitivity (probably not literally literally since dopamine doesn't work the way the folk version of it does). When I'm not stimulated enough I'm more likely to notice unwanted stimuli. Novelty, or surprise, is 'spiky' (activating, pulling of attention) enough to keep attention away from the unpleasant.

3. Higher than average branch factor + completionism = bad combo. By branch factor I mean the average number of thoughts activated by each thought. When this is less than 1 I tend towards idle relaxation, between 1 and 2 and I have a coherent chain + the ability to analyze a few possible alternatives. 3+ and I'm a bit all over the place. This isn't so bad if/when my pruning heuristic scales with the branching, but if it doesn't I'm in for a bad time. Thoughts tend to just expand without healthy cycles of contraction until they become unwieldy for working memory, at which point novelty seeking becomes a nice relief, at which point I lose most of my cache and am back to square one the next time I go to load the problem. But the next time I go to load the problem I run into exactly the same issue but now even worse because reloading the cache doesn't have any novelty to make it fun. So now the task feels really big and boring. And I don't have a strong sense of why I failed last time, so attention remains diffuse etc. This also results in avoidance of needing to keep cache, which makes it harder to break tasks into chunks with reasonable save points, which means both that activation cost is higher and that I'm likely to avoid starting them unless I can dedicate a big chunk of time.

4. The tendency to go up the ladder of abstraction rather than down. This post being a good example....Extracting universal principles feels more productive than doing the local optimization. But this becomes incorrect under marginal analysis even if it is true in general. Going up the ladder of abstraction is a way of dealing with too many branches on the object level, but often throws out getting the actual task done.

5. Mimesis. Spending too much time with others with adhd.

6. Lack of a directly responsible self. If you repeatedly had meetings at work where a particular chunk of work was brought up and the manager just said, yeah this needs to get done followed by moving on to another topic and the work never got done, you'd have a pretty good idea why this wasn't a workable strategy. Yet when this

happens internally we don't bat an eyelash. Which self will actually do the work? Because it won't be the planning and virtue signaling PR self.

7. lack of a felt sense of progress and not knowing how to generate that for oneself. Having been spoon fed too many 'felt sense of progress' superstimuli in the form of video games and escapist fiction and completionism on link aggregators that I can pretend are educational. These compete with the nebulous and self defined and enforced felt sense of progress from my own tasks.

8. Excess sensitivity to wait times. Ignoring/glossing over things with only marginally longer wait times to start, even if they are significantly better once started.

9. excessive ambiguity aversion, especially in the face of decisions with negative tradeoffs. Might imply that positive reframes are more important for adhd than for others.

10. In practice it feels like conscientiousness is the ability to generate a sub-system and then have other parts agree to subordinate their goals to it in some space-time-ritual bound way. It feels like time related cognition is the mechanism by which this coordination happens. Therefore, linked to the time dysregulation that is commonly associated with other adhd symptoms.

AI race considerations in a report by the U.S. House Committee on Armed Services

Epistemic status: Quick and dirty. A surface level dive into a particular aspect of AI governance carried out over the course of one morning.

Context

> The U.S. House Committee on Armed Services is a standing committee of the United States House of Representatives. It is responsible for funding and oversight of the Department of Defense (DOD) and the United States Armed Forces, as well as substantial portions of the Department of Energy.

The Future of Defense Task Force is a subcommittee of the U.S. House Committee on Armed Services. They have released a report, available here, and also as the first item on their latest news page. The task force is manned by an equal number of Republicans and Democrats. Though this seems a priori unlikely, it could both be the case that this report is unrepresentative of the political forces in the US Congress, and that this particular committee holds little power.

References to AI race dynamics in the report

Bold added by me.

> Technological advancements in artificial intelligence and biotechnology will have an outsized impact on national security; the potential of losing this race to China carries significant economic, political, and ethical risks for the United States and our free democratic allies for decades to come. Winning this race requires a whole-of-nation approach where the distinct advantages of both America's private and public sector are harnessed and synthesized.
>
> Using the Manhattan Project as a model, **the United States must undertake and win the artificial intelligence race** by leading in the invention and deployment of AI while establishing the standards for its public and private use. Although the Department of Defense has increased investment in AI and established the Joint Artificial Intelligence Center to assist with the transition and deployment of AI capabilities, cultural resistance to its wider adoption remains.
>
> The stakes are high. Whoever achieves superiority in this technological race will enjoy significant military and economic advantage for decades—and possibly into the next century.
>
> To incorporate the technology necessary to maintain the United States' military supremacy, **the Pentagon must continue refining its acquisition process to be more agile and less risk-averse** so that it can fully leverage emerging technologies and capabilities at scale. Train and incentivize the acquisition workforce to utilize existing flexible authorities to quickly push innovative

technology to war fighters in the field. Incentivize calculated risk by providing funding for emerging technologies through programs of record at scale; allow a less-than-perfect success rate.

History repeatedly shows that technological superiority does not guarantee victory and that new ways of thinking can be more powerful than new weapons.

The Pentagon will further need to refine its acquisition process and improve its ability to incorporate innovative emerging technologies and capabilities at the scale required to succeed in an era of great power competition

The report cites: Michael Brown, Eric Chewning, Pavneet Singh, Preparing the United States for the Superpower Marathon with China, The Brookings Institution (April 2020) (online at https://www.brookings.edu/research/preparing-the-united-states-for-the-superpower-marathon-with-china/).

Still, while China and the United States appear destined to be rivals, they maintain a complex yet symbiotic partnership that would be challenging for either country to upend, at least in the short term. Since restoring diplomatic relations with Beijing in 1979, the United States has deepened its social and economic ties with China, leading to increased prosperity in both countries. Recognizing these shared interests may allow for diplomatic endeavors and financial leverage to drive outcomes and to avoid seemingly inevitable conflict.

Rapidly advancing technologies, which offer tremendous opportunity for civil and commercial applications, are also rife with potential for nefarious use and will exacerbate threat streams exponentially for the United States and its global partners. A sophisticated array of new weaponry is changing the nature of conflict, and, while most of the technologies will require substantial funding and development by state actors, others, such as cyber and electronic warfare, may allow less formidable foes to gain the operational upper hand with limited investment, with potentially limited ability to trace the source of such actions and hold those nations accountable.

Whichever nation triumphs in the AI race will hold a critical, and perhaps insurmountable, military and economic advantage. AI allows a computer to think, learn, and perform in the cognitive ways that humans operate. Soon, advanced AI ecosystems will see machines surpassing human capability in speed, analyzation of large data sets,and pattern recognition. Advancement in AI will shape the global power structure and drive advancements in commerce, transportation, health, education, financial markets, government, and national defense

AI will shape the future of power. The nation with the most resilient and productive economic base will be best positioned to seize the mantle of world leadership. That base increasingly depends on the strength of the innovation economy, which in turn will depend on AI. AI will drive waves of advancement in commerce, transportation, health, education, financial markets, government, and national defense

Discoveries in AI, biotechnology,and quantum computing are on course to upend nearly every aspect human life and will drastically change how conflicts and wars are waged. Therefore, it is essential that democratic nations, who adhere to human rights, lead in their development and applications.

> Further, it will be incumbent upon the nations who use these technologies to set strong moral and ethical standards to protect the health and well-being of humankind. Advancements in AI, for example, will likely require a global compact in the vein of the Geneva Conventions, the Chemical Weapons Convention, and the Nuclear Non-Proliferation Treaty to establish guardrails and protect against a variety of factors, not the least of which is the infringement of personal liberty and freedoms.
>
> A sophisticated array of technologies is emerging to transform society and alter the nature of warfare. The country that can develop and incorporate these technologies the fastest and most effectively will enjoy significant military and economic advantage for decades to come.
>
> Expanding critical investments in innovative technology and programs will require an increased tolerance for calculated risk at the Pentagon and in Congress. It also requires the discipline to invest in systems and operational concepts necessary to succeed and the will to eliminate those that do not. Correctly navigating these difficult trade-offs will determine whether the United States is able to remain in over-match against great power competitors.
>
> **This concept of “algorithmic warfare” will pit algorithms against algorithms where information and the speed of decisions will likely be more important than traditional means of military superiority, such as the size of opposing forces or the range of armament.** Those with superior data, computing power, information security, and connectivity will maintain the upper hand. This paradigm will require new operational concepts and equipment to adapt and maintain the advantage.

Brief commentary and discussion

The diagnosis in the report — that China will grow in power and technological acumen — seems basically accurate, though the recommendation that the USA should be willing to take more risk and go out with a bang seems more questionable. The language in the report is heavily adversarial, with an emphasis on winning an AI race, rather than with an emphasis on exploring and developing robust mechanisms to avoid Red Queen races.

The committee also seems to assume that AI scenarios will in the short term be multipolar, with the United States, China and Russia competing to develop their AI capabilities, while smaller nations also invest in asymmetric warfare. Europe's capabilities aren't considered at all. Crucially, AI is here seen as only one of many factors to consider in an engagement. Scenarios outside the Overton window, such as intelligence explosions, aren't considered.

I initially arrived at this report while researching this CSET question. Overall, I'd say that the language in this post is a signpost or warning sign of potentially more heated AI races to come.

It's unclear to me what levers there exist to make the US's approach less adversarial. Some brainstorming:

- Organize a campaign to call your representative: I imagine that very few people call their representatives to talk with them about this particular topic. Form and

fund a lobbying group.
- Study and make common knowledge instances where fading and raising powers have been able to live somewhat peacefully. Examples might include Britain and the US after the Suez Crisis, Portugal and Spain at the height of their respective empires, Greece still being respected after Rome had become the dominant military power, etc.
- Think long and hard about how to make arms reduction treaties applicable to AI systems.

Best of luck to the folks working on AI governance.

Appendix: All interesting quotes I wrote down.

> China represents the most significant economic and national security threat to the United States over the next 20 to 30 years. Because of its nuclear arsenal and ongoing efforts to undermine Western democratic governments, Russia presents the most immediate threat to the United States; however, Russia's long-term economic forecast makes its global power likely to recede over the next 20 to 30 years
>
> As a result of historic levels of government-sponsored science and technology research, and the inherent advantages of a free market economy, the United States emerged from the Cold War with a substantial economic and military lead over any potential rival. However,these gaps have dramatically narrowed. China will soon overtake the United States as the world's largest economy, and despite historic defense budgets,the United States has failed to keep pace with China's and Russia's military modernization.
>
> Advancements in artificial intelligence, biotechnology, quantum computing, and space, cyber, and electronic warfare, among others, are making traditional battlefields and boundaries increasingly irrelevant. To remain competitive, the United States must prioritize the development of emerging technologies over fielding and maintaining legacy systems. This will require significant changes to the Pentagon's force structure, posture, operational plans, and acquisition system and must be complemented by a tough and fulsome review of legacy systems, platforms, and missions.
>
> The Pentagon's emerging operational concepts have the potential to provide the U.S. military a decisive advantage, but they are not yet fully viable. To address current and future threats and deter conflict, the Department of Defense must more aggressively test new operational concepts against emerging technologies.
>
> Technological advancements in artificial intelligence and biotechnology will have an outsized impact on national security; the potential of losing this race to China carries significant economic, political, and ethical risks for the United States and our free democratic allies for decades to come. Winning this race requires a whole-of-nation approach where the distinct advantages of both America's private and public sector are harnessed and synthesized.

Using the Manhattan Project as a model, the United States must undertake and win the artificial intelligence race by leading in the invention and deployment of AI while establishing the standards for its public and private use. Although the Department of Defense has increased investment in AI and established the Joint Artificial Intelligence Center to assist with the transition and deployment of AI capabilities, cultural resistance to its wider adoption remains.

To maintain its global preeminence in scientific and technological innovation and the associated economic and military advantage, the United States should increase its investment in foundational science and technology research by committing to spending at least one percent of the country's gross domestic product on basic government-supported research and development.

Require the military services to spend at least one percent of their overall budgets on the integration of new technologies.

To maintain the United States' military advantage against emerging threats, the Pentagon must refine its operational concepts by employing new technologies and methods to deter future conflicts and compete in the gray-zone of hybrid warfare.

The Pentagon, Congress, and the Intelligence Community should work in tandem to identify trends and threats 10 to 30 years beyond the normal budget cycle while expanding entities within their respective organizations to incorporate long-term planning

To incorporate the technology necessary to maintain the United States' military supremacy,the Pentagon must continue refining its acquisition process to be more agile and less risk-averse so that it can fully leverage emerging technologies and capabilities at scale. Train and incentivize the acquisition workforce to utilize existing flexible authorities to quickly push innovative technology to war fighters in the field. Incentivize calculated risk by providing funding for emerging technologies through programs of record at scale; allow a less-than-perfect success rate.

The gravity and complexity of threats emerging to challenge the United States is proliferating as technological advancements in artificial intelligence, quantum information science, and biotechnology transform society and weaponry at an exponential rate. This is occurring as adversarial capability is increasing to the point where the United States may soon lose the competitive military advantage it has enjoyed for decades.

A sophisticated array of emerging technologies and new weaponry, in various stages of development, will fundamentally change the nature of conflict along with the very battle space where it will be fought. The stakes are high. Whoever achieves superiority in this technological race will enjoy significant military and economic advantage for decades—and possibly into the next century.

Advancements in artificial intelligence, quantum information science, space and cyber and electronic warfare, among others, are making traditional battlefields and boundaries increasingly irrelevant. To remain competitive, the U.S.must recognize this shift and prioritize the development of emerging technologies while also increasing its ability to defend against them.

The U.S. military, with its adherence to human rights and the rules of engagement, stands as the global model for how a free and open society should

protect itself and its interests. Exporting U.S. values through military engagements, with both exercises and train and assist programs, builds trust and interoperability while increasing readiness and resiliency and further protecting vital U.S. interests abroad.

the U.S. and Russia should extend the highly successful Strategic Arms Reduction Treaty (New START) while negotiating a follow-on agreement.

The U.S.has long been the global leader in technological innovation because of its investment in government-funded research and development (R&D) that has led to breakthroughs such as the Manhattan Project and the space program. Without increased investment and focus, however, its pre-eminence is at risk.

Historically, the U.S.has outpaced every other country in overall R&D spending, but its lead is quickly diminishing. Over the past two decades, China has rapidly increased its investment in overall R&D, whereas U.S. spending rates have lagged.Today, the U.S.still spends more than any other country, but China is on track to take the lead in global R&D spending by 2030 if current trends continue.

History repeatedly shows that technological superiority does not guarantee victory and that new ways of thinking can be more powerful than new weapons. Future leaders and strategists will need to embrace emerging war fighting concepts such as joint and multi-domain warfare. They will further need a comprehensive understanding of national power and how to integrate military tools into a whole-of-government effort.

The Pentagon will further need to refine its acquisition process and improve its ability to incorporate innovative emerging technologies and capabilities at the scale required to succeed in an era of great power competition

China's economic power continues to grow, and China remains on a glide path to be the world's largest economy by as early as 2030. If the U.S. defense posture maintains its current trajectory, 70 percent of the military's systems will be legacy platforms when that occurs. In contrast, China and Russia adhere to fewer traditional systems, allowing them to more easily field future capabilities

The report cites: Michael Brown, Eric Chewning, Pavneet Singh, Preparing the United States for the Superpower Marathon with China, The Brookings Institution (April 2020) (online at https://www.brookings.edu/research/preparing-the-united-states-for-the-superpower-marathon-with-china/).

According to the latest Department of Defense assessment, China has doubled its defense spending in the last decade and now has more ships than the U.S. Navy, among the best air defense systems globally, an arsenal of long-range ballistic missiles,and a variety of other means to challenge the U.S.A sobering report from the RAND Corporation recently determined that despite significantly outspending China and Russia, the U.S. military could lose a future conflict because it failed to adequately posture and train.

Still, while China and the United States appear destined to be rivals, they maintain a complex yet symbiotic partnership that would be challenging for either country to upend, at least in the short term. Since restoring diplomatic relations with Beijing in 1979, the United States has deepened its social and economic ties with China, leading to increased prosperity in both countries. Recognizing these shared

interests may allow for diplomatic endeavors and financial leverage to drive outcomes and to avoid seemingly inevitable conflict.

Rapidly advancing technologies, which offer tremendous opportunity for civil and commercial applications, are also rife with potential for nefarious use and will exacerbate threat streams exponentially for the United States and its global partners. A sophisticated array of new weaponry is changing the nature of conflict, and, while most of the technologies will require substantial funding and development by state actors, others, such as cyber and electronic warfare, may allow less formidable foes to gain the operational upper hand with limited investment, with potentially limited ability to trace the source of such actions and hold those nations accountable.

Whichever nation triumphs in the AI race will hold a critical, and perhaps insurmountable, military and economic advantage. AI allows a computer to think, learn, and perform in the cognitive ways that humans operate. Soon, advanced AI ecosystems will see machines surpassing human capability in speed, analyzation of large data sets,and pattern recognition. Advancement in AI will shape the global power structure and drive advancements in commerce, transportation, health, education, financial markets, government, and national defense

AI will shape the future of power. The nation with the most resilient and productive economic base will be best positioned to seize the mantle of world leadership. That base increasingly depends on the strength of the innovation economy, which in turn will depend on AI. AI will drive waves of advancement in commerce, transportation, health, education, financial markets, government, and national defense

Discoveries in AI, biotechnology,and quantum computing are on course to upend nearly every aspect human life and will drastically change how conflicts and wars are waged. Therefore, it is essential that democratic nations, who adhere to human rights, lead in their development and applications.

Further, it will be incumbent upon the nations who use these technologies to set strong moral and ethical standards to protect the health and well-being of humankind. Advancements in AI, for example, will likely require a global compact in the vein of the Geneva Conventions, the Chemical Weapons Convention, and the Nuclear Non-Proliferation Treaty to establish guardrails and protect against a variety of factors, not the least of which is the infringement of personal liberty and freedoms.

History presages that when the United States competes from the moral high ground, it usually wins.

Simply stated, China needs the United States economically, at least for the short term.

A sophisticated array of technologies is emerging to transform society and alter the nature of warfare. The country that can develop and incorporate these technologies the fastest and most effectively will enjoy significant military and economic advantage for decades to come.

Historically, the United States has led the world in funding tech R&D, which has allowed it to maintain a strategic advantage. China, however, appears poised to challenge the United States as the overall leader in R&D spending. In response,

the U.S. government and the Pentagon should consider increasing investment in basic R&D while developing R&D partnerships globally.

Pentagon culture and business practices are rightfully designed to be fair and open and to avoid waste and abuse. However, this can sometimes make them slower-moving, risk-averse, and process-based rather than outcome-based, which can hinder the military's ability to fully utilize private sector innovation. Established practice and culture favor large, traditional business partners, which makes it more difficult for non-traditional companies with innovative technology to compete. The Pentagon knows how to acquire large programs of record like fighter jets or aircraft carriers, but it is less adept at purchasing at scale the types of emerging technologies that will be required for future conflict.

Because of risk aversion and fear of potential failure, the Pentagon often fails to fully utilize its existing authorities to quickly incorporate private sector technology,even when urgently necessary. This hinders its ability to fully leverage outside advances at the necessary speed and scale.

To maintain its strategic advantage, the United States must recruit and develop a workforce with the requisite skills and talent to maintain the country's technological and military advantage. In matters of national security, people are more important than hardware;therefore,the United States must develop, recruit, and retain the most talented science and technology, military, and national security professionals globally. Along with recruiting and growing science, technology, engineering,and mathematics(STEM) talent, the military and national security community must update personnel policies to ensure that they can attract and foster talent

With its global obligations and missions, the United States outspends all its rivals combined in defense expenditures. In 2019, it spent more than $730 billion, while China and Russia spent roughly $260 billion and $65 billion,respectively. This is nearly three times as much as China and ten times as much as any other country. While the United States maintains a global military presence and supports a variety of missions, partners, and allies, China and Russia have historically focused on their respective regions, although both are rapidly working to expand their global reach. China's economy will likely exceed the United States' in dollar terms in the next 10 years.

Expanding critical investments in innovative technology and programs will require an increased tolerance for calculated risk at the Pentagon and in Congress. It also requires the discipline to invest in systems and operational concepts necessary to succeed and the will to eliminate those that do not. Correctly navigating these difficult trade-offs will determine whether the United States is able to remain in over match against great power competitors.

Now, in both conventional warfare and gray zone tactics, Russia and China are able to challenge the United States in multiple arenas. Indeed, it is what they have been preparing for over the last two decades while the United States was focused on countering terrorism

This concept of “algorithmic warfare” will pit algorithms against algorithms where information and the speed of decisions will likely be more important than traditional means of military superiority, such as the size of opposing forces or the range of armament. Those with superior data, computing power, information

security, and connectivity will maintain the upper hand. This paradigm will require new operational concepts and equipment to adapt and maintain the advantage.

Book Review: Reinforcement Learning by Sutton and Barto

Note: I originally submitted this for the Slate Star Codex (RIP) book review contest, but given the current drama it probably won't happen for a while (if at all), so I decided to share this review with the LW community in the meantime.

Cross-posted on my blog: https://billmei.net/blog/reinforcement-learning

Reinforcement Learning profoundly changed my understanding of the science of happiness, biological evolution, human intelligence, and also gave me unique tactics for rapid skill acquisition in my personal life.

You may not be expecting these conclusions, as on the surface this is a technical textbook for those wishing to learn about reinforcement learning (RL), the subfield of machine learning algorithms powering recent high-profile successes of AIs trouncing human champions like AlphaGo in Go and Watson in Jeopardy.

However, I wrote this review in layman's terms, and it contains no math, to help disseminate these ideas to people who would be intimidated by the math and unlikely to read it otherwise, as I highlight only the most important (and underrated!) topics from this book that you won't find elsewhere. If you instead want a comprehensive math-included summary of the book, you can read my detailed notes on each chapter.

What is reinforcement learning

Reinforcement learning is a study of what are the best actions (*policy*) to take in a given environment (*states*) to achieve the maximum long-term *reward*.

I've found RL to be the most elegant, first-principles formalization of what it means to "win". It's an algorithm that explicitly gives you the optimal actions to take to achieve the goal you define, with zero outside knowledge other than the input environment. Unlike other machine learning algorithms, RL does not require you to specify subgoals (capture the most chess pieces), only the ultimate goal (win at chess).

The Slate Star Codex and LessWrong communities started using Bayes' theorem as the principal guide towards finding the truth. Bayes' theorem is the first-principles formalism of how to evaluate evidence and update your beliefs so that your beliefs will match reality as closely as possible. But ultimately "Rationality is winning"; the purpose of "Rationality" is not just to have a good epistemology, but to successfully achieve goals in the real world.

Just as Bayes' theorem is the mathematical foundation for finding the truth, reinforcement learning is the computational foundation for winning.

How reinforcement learning works

To oversimplify, let's focus on temporal-difference (TD) learning. TD learning uses a *value function* that estimates how good it is to take a certain action, or how good it is

to be in a *state*: a certain configuration of the environment. For example, in chess it's generally better (but not always) to have more pieces than your opponent. It's also generally better (but not always) to take actions that capture pieces than to take actions that result in your pieces being captured.

If the value function predicts that a state is better than it actually is, causing a move that results in losing the game, then we perform one iteration of a *TD update* which revises the value function downwards on this particular state (and other similar states). Vice-versa if it undervalued a state that later resulted in a win. We call this misestimation a *TD error*.

The RL agent will then mostly try to play only the actions that lead to board states with the highest estimated value, although it will sometimes play suboptimal actions to avoid getting stuck in local optima using a strategy that I will describe later.

Over many games and many iterations, the value function's estimate asymptotically converges to the true value of the state (played by some hypothetically perfect player). Then, assuming these estimates are correct, the RL agent can win by simply playing actions that lead to the highest value states. Even if the estimated value does not perfectly match the true value, in most practical cases the estimates become good enough that the agent can still play with superhuman skill.

Mistakes drive learning

What's interesting about the TD algorithm is that it learns only from its mistakes. If the value function estimates that I will win a game, and then I go on to actually win, then I get a positive reward but the TD error is 0, hence *no TD update is performed*, and no learning occurs. Therefore, an RL agent that learns quickly won't always just choose actions that lead to the highest state values, but instead identify states that haven't been played frequently and try to play in such a way to get to those states to learn about them. While this may cause the agent's winrate to decrease in the short term, it's a good way to improve out of local optima as focusing only on a small number of states leaves the agent ignorant about the universe of what's possible.

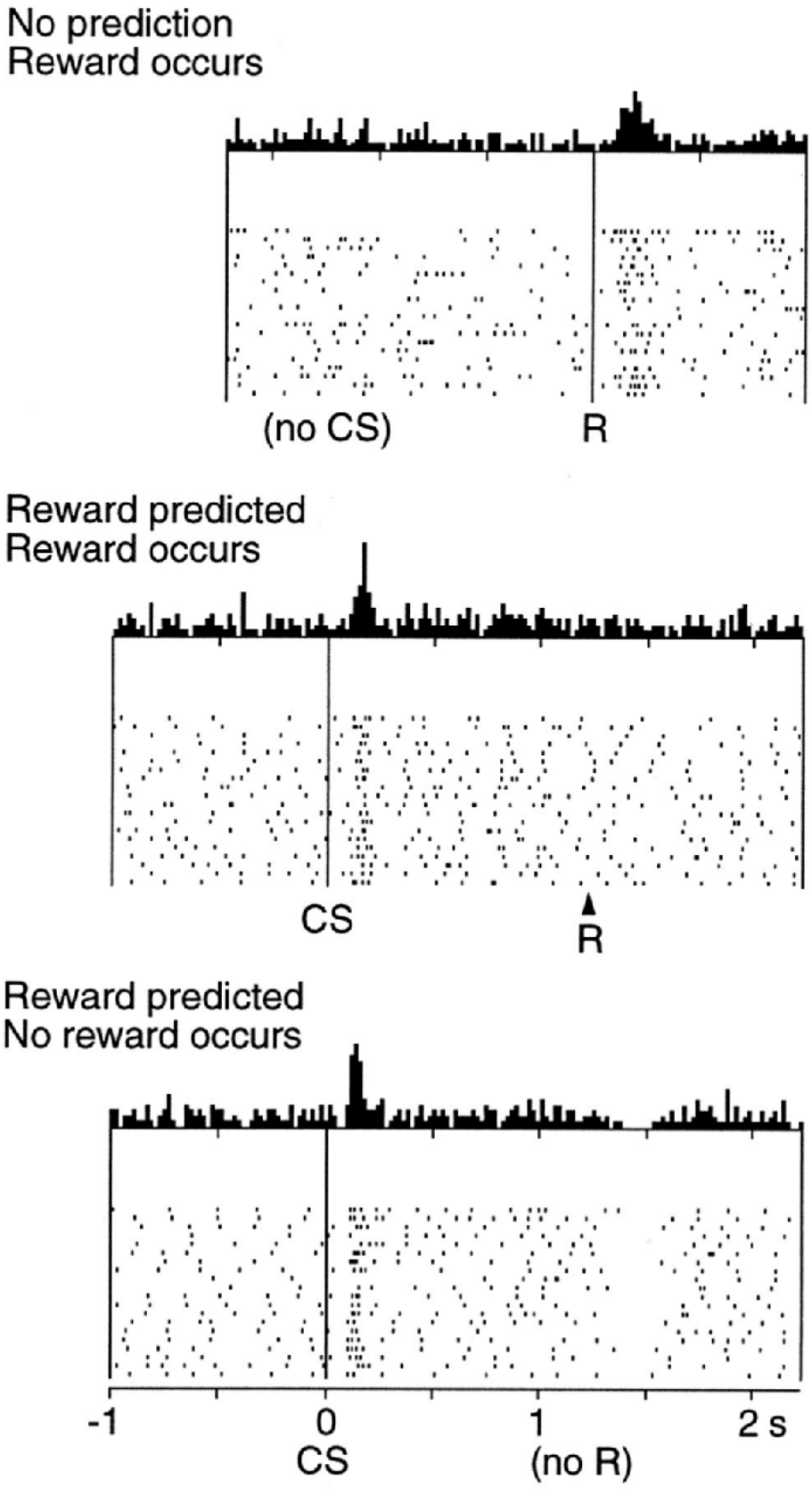
This echos the research on how you can acquire skills using "deliberate practice". The popular "10,000" hours rule oversimplifies the idea of "just practice a lot and you will be good", as the research of *deliberate* practice shows that if you're just doing mindless repetitions of an activity (or worse, winning a lot or having fun at it!), you aren't actually learning, as it requires struggling through mistakes and constantly challenging yourself at a level just beyond your reach that results in actual learning.

While you may already be familiar with the research on deliberate practice, RL provides a *mathematical* justification for this concept which further solidified it for me.

Dopamine and Happiness

Dopamine is a neurochemical people commonly associate with pleasure. But dopamine does not actually function this way in biological organisms; instead, it is an error signal. While it may seem like dopamine is analogous to the *reward* in an RL algorithm, it is not the reward. Instead, dopamine is the *TD error*.

Echoing Pavlov’s dogs, the book describes a study where researchers (Schultz, Apicella, and Ljungberg, 1993) trained monkeys to associate a light turning on with a reward of a drop of apple juice. They hooked up brain scanners to the monkeys to monitor their dopamine receptor activity. Here’s how those neurons behaved:


No prediction
Reward occurs
(no CS)
R
Reward predicted
Reward occurs
CS
R
Reward predicted
No reward occurs
-1
0
1
2 s
CS
(no R)

The x-axis is time elapsed and the y-axis is strength of dopamine response.
CS: Light turning on
R: Drop of apple juice dispensed

From the diagram, the monkeys get a dopamine spike when the apple juice is surprisingly dispensed without any forewarning from the light. If the light turns on, the monkeys get a dopamine spike right away in anticipation of the upcoming reward, but then *no dopamine is produced once the apple juice is actually dispensed*. In a third experiment, the light turned on which produced a dopamine spike as expected, but then later no juice was dispensed, which caused the monkey's dopamine levels to drop below baseline as their expectations were disappointed.

From this it's clear that dopamine's purpose is an error signal, not a reward signal. We don't get any dopamine when our rewards are exactly the same as our expectations, only when the rewards exceed expectations. Conversely, dopamine levels drop when our expectations are high and the rewards are disappointing. Likewise, once an RL agent receives a reward its TD error is positive when the value function undervalued its actions, and negative when it overvalued its actions.

It's a trope that "happiness equals reality minus expectations", and while dopamine is not the only neurochemical that contributes to happiness, the implication of this study is the more skilled you get at accurately predicting reality, the less pleasure (and less disappointment) you get from your efforts. Perfect prediction is perfect equanimity.

Another implication is in the psychology of addiction. This phenomenon underlies the behaviour of "chasing the high"; every time you receive a reward, your expectations revise upwards to match reality, so next time the reward needs to be greater to generate the same dopamine response, as receiving the same reward causes no "TD error".

These conclusions may be unsurprising to you if you are a psychiatrist, but what I found extraordinary is the research and science around dopamine was discovered many years after when the TD algorithm was developed. The RL algorithm designers were totally ignorant about this property of dopamine, yet independently came up with an algorithm for shaping the behaviour of computer agents that looks remarkably similar to how we incentivize behaviour in animals.

Is there a shared computational universe that drives the behaviour of both biological beings and computer algorithms? It certainly looks convincing.

Simulated Experience

In chess you have full knowledge of the environment, but what about tasks where you don't? Imagine a robot trying to navigate a maze. If this were an animal study we may reward its success with food, but AIs are satisfied to receive rewards as floating point numbers, with greater rewards for faster navigation. At every intersection, the RL agent must decide which corridor to follow to exit the maze.

How can the RL agent know what "state" it's in when it doesn't even know what the maze looks like? Any given intersection may look exactly like several others. Worse, what happens if the maze is constantly changing while the agent is in the middle of running it?

In the absence of external sensory information about the real-world maze, the RL algorithm constructs simulated mazes that it runs virtually. Driving a robot through a real maze is slow, but the computer can run thousands of simulated mazes per second.

In the beginning, when the agent knows nothing about the real maze, its simulated mazes will be poor imitations of the real thing, as it lacks data on what mazes are supposed to look like. So the experience it gains from these simulated mazes is worthless.

But slowly, as the robot drives through the real maze and collects increasing ground-truth data about the maze, it can improve its simulated mazes until they become reasonably accurate approximations of the real thing. Now when the algorithm solves a thousand simulated mazes, it gains the experience of having solved a thousand physical mazes, even though in reality it may have only solved a handful.

This is how Monte Carlo Tree Search works (although I oversimplified here for brevity), and it was the key to AlphaGo's victory over the top human Go players.

In the book *Sapiens*, Yuval Noah Harari argues that what separates humans from other primates is our ability to imagine events that don't exist. Our ability to learn from these fictional events is what endows us with intelligence.

If you can daydream about a non-existent lion, then you don't have to wait to actually encounter a lion to figure out what to do, when it may be too late.

At the risk of inappropriately anthropomorphizing our RL agent, this convinced me that the ability to simulate experience is one of the key building blocks of intelligence, here applied to machines instead of humans.

Simulated Rewards

The ability to simulate experience necessarily also means being able to imagine rewards (or punishments) at the end of those simulated experiences.

Any basic optimization algorithm suffers from the problem of being stuck in local optima. Humans can think globally and creatively because we can delay gratification; we generally don't always take the greedy step towards rewards, as anyone who does so is seen as more simple-minded or lacking willpower. Conversely, we respect people who have superior self-control, as working towards long-term goals generally leads to more success, health, popularity, etc., thus we perceive these people to be more intelligent.

I want to argue here that our ability to delay gratification is not the result of willpower, but actually a hack. We don't really delay gratification, instead we substitute a real reward for an imagined one.

In the famous Stanford marshmallow experiment, children who were able to give up a marshmallow to wait 15 minutes in an empty room received 2 marshmallows afterwards. Compared to the kids who didn't wait, the kids who waited later had improved SAT scores, educational attainment, and other measures of life outcome.

If you watch some videos of this experiment, what's remarkable is you will notice the most successful kids aren't the ones who have iron willpower, but instead those who

were able to distract themselves by singing songs, playing with their hands, etc.

Thus, the key to long-term planning is not the ability to push back a reward, but instead the ability to be satisfied with an imagined fiction tell yourself of an even greater reward you can receive if you wait.

While I use the terms “fiction”, “simulated”, and “imagined”, it’s important to note that this “synthetic happiness” is not fake. Biologically, psychologically, and computationally, it is in every way as real as “real” happiness. Dan Gilbert, the happiness researcher, presents the data behind this in a TED talk:

> *We smirk, because we believe that synthetic happiness is not of the same quality as what we might call “natural happiness”.*
>
> *[…]*
>
> *I want to suggest to you that synthetic happiness is every bit as real and enduring as the kind of happiness you stumble upon when you get exactly what you were aiming for.*

From our RL algorithm’s perspective, its simulated rewards are as real to its value function as the real rewards. The only difference is the real rewards have the ability to change the way the simulation itself is constructed, whenever expectations are violated. The way the math works, the agent’s ability to long-term plan results not from its delaying immediate rewards, but substituting real short-term rewards for simulated long-term rewards that have a larger floating point value.

Simpler animals require Pavlovian training and direct rewards/punishments to influence their behaviour, but you can get a human being to toil away in the fields with only a story about an imagined afterlife.

Importance Sampling

After three moves in chess, there are 100+ million possible board positions and 10^{120} possible continuations (more than the number of atoms in the universe), only a tiny sliver of which result in a win. To have any hope of getting a good value function with limited computing power, your algorithm must focus on analyzing the most promising (or most common) moves and avoid spending clock cycles on positions that are clearly bad and will never occur in a real game.

But your opinion on what is “good” may differ from someone else’s opinion. An RL agent’s experience in a game is highly path-dependent; if it happens to get lucky with a certain sequence of actions it may overvalue these actions and hence choose them more often than a counterfactual average agent. Thus, how much credit should you give to your own experience vs. others’ experiences? Morgan Housel says “Your personal experiences make up maybe 0.00000001% of what’s happened in the world but maybe 80% of how you think the world works.”

The *importance sampling ratio* is a modifier used by the RL agent to upregulate or downregulate its value function to reduce the variance from luck, without changing the expected value. It’s calculated using a set of weight parameters that is adjusted based on a *behaviour policy*, which you can think of as the RL agent’s simulation of what an average agent would do in its stead.

Just as Bayes' theorem gives you the math for *exactly* how much you should increase or decrease your confidence in a belief in response to new evidence, importance sampling gives you the math for *exactly* how much credit you should give to your own experiences versus others' experiences so you can correct for your narrow slice of reality without throwing up your hands and always deferring to others.

I believe importance sampling is the appropriate response to avoid overcorrecting and, as described in *Inadequate Equilibria*, to avoid becoming "too modest" to the point where you stop trusting your own intuitions even when they are right.

Conclusion

It may be offbeat to do a book review of a dense, 500-page math textbook, but I found the ideas presented in *Reinforcement Learning* revolutionary to clarifying my understanding of how intelligence works, in humans or otherwise.

In this review I've omitted the math formulas because they are quite involved and any variables I use would be meaningless without a lengthy introduction to them. You don't need a deep math background to read this book though, it's written in an approachable style that requires only 1st-year university math and some basic machine learning knowledge to understand, and I believe that is why this is the most popular textbook for beginners to RL.

I highly recommend this book if you are interested in what modern machine learning algorithms are capable of beyond just variations on linear regression and gradient descent. If you have the inclination, I also recommend completing the coding exercises in each chapter—coding up the algorithms helped me *feel* what was going on as if I were an RL agent myself, and this was the key that allowed me to draw the parallels between human and machine intelligence that I described in this review.

It took me four months to read this book and do the exercises, and I also did it as part of a class I am taking for my Master's Degree, but it was well worth the investment. This book took me from barely any idea about reinforcement learning to being able to comfortably read the latest RL research papers published on arXiv. Perhaps you'll also discover something new about AI that you didn't realize you were missing before.

Bibliography

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *The Journal of Neuroscience*, 13(3), 900–913. doi.org/10.1523/JNEUROSCI.13-03-00900.1993

Group debugging guidelines & thoughts

This post is a response to the request for a post with a "thorough description of how to do pair debugging"; I'm also having this post double as a retrospective on my attitude towards pair debugging in general.

As described by Vaniver,

> At CFAR, one of the exercises is 'pair debugging'; one person is the protagonist exploring one of their problems, and the other person is the helper, helping them understand and solve the problem. (Like many things at CFAR, this is a deliberate and distilled version of something that already happens frequently in normal life.)

Starting from late 2014 (when I visited my first CFAR workshop), I ran an in-person rationality/self-development group which was inspired by CFAR's practices but also went in its own directions. One of the activities we had were regular debugging circles: people would split into small groups and then take turns where one person at a time would explain a problem that they were having and others would then attempt to help them out with it.

As I've understood CFAR's debugging circles, they are relatively free-form by design; the main focus is on exploring the problem together, and also notice opportunities where various rationality techniques could be applied. As such, they avoid having too much explicit structure in order to allow many different approaches. Yet as there still seems to be something like "the skill of doing good debugging", in 2015 I wrote the following guidelines for us that seemed to distill some best practices.

My 2015 debugging manual

Etiquette

Only bring up something if you actually want a solution for it. Everyone has times when they just need to vent and want sympathy rather than solutions, but debugging circles aren't the place for that. This doesn't mean that you would need to accept any suggestion that the others bring up, but it does mean that that you should be open to others offering suggestions *in general.* Once the session ends, you're free to just ignore and forget anything that didn't seem to make sense to you.

Be courteous of others and their time. Do your best to make sure that everyone, both you and the others, get a roughly equal share of the group's attention. If you have lots of problems in your life, don't dump all of them on the group at once, but rather focus on one or a small set of related ones. If it starts looking like the discussion has gotten stuck on one person's issues for an extended time and the others might not have a chance to have their issues discussed, gently but firmly suggest moving on to the next person. Try to be considerate enough to pass on your own turn early enough that someone else doesn't need to prompt you to do so.

If someone is undergoing a particularly difficult time or has a particularly important issue going on, it's alright to sometimes spend a disproportionate time on them: but you should try to avoid being that person each time.

But have a fair respect of your own time, as well. The opposite also applies: if you genuinely feel that there's nothing in your life that needs discussing, you're free to cut your own turn short, but if you do it many occasions in a row, you're probably not taking full advantage of the group. If nothing else, you can always use the group to get an opinion on any assumptions behind your current plans. You have a right to get help from the group in return for helping others: stick to that right.

Don't proselytize your view. Maybe you're *completely certain* that the cause of the other person's problems is that they don't have cat ears as a part of their attire, which would *totally fix* everything if they just changed that. You're free to think that, but if they disagree with your suggestion, don't get stuck arguing but let it go.

Approaches

Start by trying to understand what problem the person is trying to solve. "I've been trying to sign up for dance lessons but can never seem to get around it." One possible approach would be to immediately start offering ways for the person to sign up for dance lessons. Often a more fruitful one would be to first ask - why do you want to attend dance lessons? Maybe it turns out that the person doesn't actually care about learning to dance, but is feeling bad because their friend, a great dancer, always gets all the attention at parties. Then the actual problem is not "how to learn to dance" but "how to get other people to notice me". It's quite possible that not knowing to dance isn't actually the biggest issue there.

Test your understanding of the problem. When you're formulating an understanding of the problem, it can be useful to frequently verbalize it to the other person to make sure that you've understood correctly. "So you seem to be feeling bad because your partner just became the President of your country while you mostly spend time playing video games, is that right?"

A rule of thumb that I sometimes use is "do I feel like I understand this problem and its causes well enough that I could explain to a third person why this person wants to solve it and why they haven't been able to solve it yet?" If the answer is no, hold off proposing solutions and try asking more questions first.

Even "obvious" problems may benefit from questions. Someone once mentioned that they tend to often jump to being critical of others, which tends to be harmful. Here the causal mechanism seemed to be pretty obvious, but asking "*how* does it tend to be harmful" was still useful in bringing out details of the exact nature of the typical criticism and how people tended to react to that.

Look for trigger-action patterns. "I always end up being on the computer and wasting time and then feeling bad." What specific things on the computer act as time-wasters, and how exactly does the person end up doing those things? Maybe they often feel bored or anxious, which causes them to open Facebook, which causes them to get lost in a maze of discussions and links. Would there be a way to either remove the anxiety, or find a new action to carry out when anxious? Which one would be easier?

Look for positive and negative reinforcers in the environment. "I often post a link to Facebook, and then I keep returning to Facebook throughout the day because I want to see whether it's accumulated new likes and comments." Here, logging on to Facebook after posting a link keeps getting reinforced by the accumulation of comments and likes, which provide a reward each time that the page is opened and there's a new one. Could something be done to eliminate those reinforcers? ("If anyone sees me responding to a comment or posting a link, don't like it but do remind me that I was supposed to be working.") Or maybe provide reinforcers for something else? ("For each ten minutes that passes without me logging onto Facebook, could you please come give me a hug?")

Be specific about the causes of emotional reactions. "My boss is so full of himself, it drives me nuts." Exactly how does the full-of-himself-ness manifest? If the exact behavior is "he often interrupts", maybe something could be done about that thing in particular. Best case: the boss comes from a conversational culture where interrupting is normal, and hasn't even realized that someone would consider it rude - but this would have been impossible for the others to suggest if the problem description would only have been on the level of "he's so full of himself".

This is also a useful technique for reducing your own annoyance at others, even if it was just something you did in your head. "I'm getting frustrated now because that person is talking really loudly and I would like to read." Breaking down an atomic "AAAAAGH I'M SO FRUSTRATED" into a "I'm feeling [specific emotion] because [specific cause] and [that violates my desire/need to something]" is not only useful for debugging, it can also relieve the frustration by itself.

Assume that problems won't fix themselves. In one session, someone says they intend to implement some change for next week's meeting. In the next session, they say, "yeah, that plan didn't really work out, but I was kinda busy and distracted this week. I'm going to try harder."

Chances are, if they were busy and distracted this week, they're likely to be busy and distracted the next week, too. "I'm going to try harder" often translates either as "I don't actually care about solving this problem but want to give the impression that I do", or alternatively, "I don't actually know how to fix this but I'm going to try again the same way, in the hopes of magically getting a different result now". Assuming that the person really *does* want to solve their problem, try to figure out exactly what went wrong and how it could be avoided in the future.

Ask, "is there a more general problem here?" Someone wants to cut down on the amount of money that they spend on fast food. One day when they're coming home from work they walk past a hamburger place, are tempted by the advertisements, and go there to eat. This happens several times.

The specific problem in this case would be "I always end up eating at the Burger King on the 27th street on my way home". The more general form of the problem might be something like "each time I walk past a fast food place when I'm hungry, I end up eating there". General solutions might be "pick a route that allows you to avoid seeing fast food places when you're hungry" and "make sure to carry something with you that allows you stave off the worst of the hunger until you're home".

Focusing. Someone is having difficulties deciding whether to try to solve a problem or whether to accept its consequences and let it be. One approach would be to have them verbalize all the reasons why the unsolved issue bothers them, and then say out

loud, "having considered all of these consequences of the problem, I find that they're acceptable and it's better to just let this be". Does saying that feel right to them, or does something about it feel wrong? What if they were to say, "having considered all of these consequences of the problem, I find that they're unacceptable and I want to solve the problem", instead? Would that feel right or wrong?

Quick Murphyjitsu. After you've come up with a plan, it may be useful to have the other person do a quick Murphyjitsu on it. How surprised would they be if this plan failed? If not particularly, is there any obvious failure mode that comes to mind and which could be fixed?

Check that the person remembers something actionable. Sometimes discussion may suggest some actionable things, then drift to e.g. more general discussion of the problem which doesn't provide as many concrete suggestions. If this happens, make sure that the person whose problems are being debugged still remembers the actionable suggestions they got earlier on.

2016: going more structured

A limitation with this approach was that even with these guidelines, newcomers in particular expressed uncertainty of what to actually *do* while debugging others. I also had the experience that while this kind of a freeform process was occasionally helpful, it didn't seem to help much with some of my own biggest issues - and that some others would likewise repeatedly participate in these sessions, but not seem to make much progress.

A one-sentence description of debugging might be "helping a person make progress on problems they haven't been able to solve yet"; the already-existing fields dedicated to this same issue are therapy and coaching. As it seemed pointless to re-invent the wheel, we adapted the GROW model of coaching as a structured guideline for people to use, and also shifted our language to talk about "coaching" to avoid needless insider jargon. At this point, we also shifted from being mainly in a circle, to mainly being one-on-one.

The following is one of the handouts describing the process that we ended up using, from 2016.

The GROW coaching model

(Goal, Reality, Options / Obstacles, Way Forward)

What follows is a recommended baseline model for a coaching session. Its purpose is to support the coaching process, and you should feel free to deviate from it should you feel that it’s necessary. Often this happens naturally: the coachee may find the way forward in the reality phase, or thinking about the current situation may lead you to change and redefine the goal. It’s perfectly fine for you to go back and forth in the steps, turning GROW into something more like GRGROGROOGROWOGORW.

Step 1: Goal

The coachee defines one objective that they want to reach. A good objective is valuable, specific, measurable, achievable, realistic and time-bound. In case the coachee isn't clear on what they might want, you may discuss their life in general until you come up with something. In general, goals may be related to:

- Changes you'd like to make to your life
- Things that you've always wanted to do or learn, but never gotten around
- Problems that you repeatedly keep running into
- Habits you'd like to establish or change
- Skills that you'd like to improve
- If you can't think of anything, try to tell yourself, "my life is perfect, I wouldn't change a thing". Does saying that feel wrong somehow, are there any counterexamples that come to mind?
- Alternatively, try imagining what your ideal life *would* be. What would be your best possible self and life, if you could decide without caring at all about what was practical?

You can also try the following questions:

- What's the best possible life that you could imagine having? How could you get closer to it?
- What kinds of things have caused you pain recently? Could you do something about them? What kinds of needs could be behind them?
- When do you feel most alive? How often do you feel that "these kinds of moments make life worth living for"?
- Do you have things that you've always wanted to do, but never have?
- Are there any aspects of your life where you feel like you're stuck?
- Do you feel like you're living a life that's in accordance to your values?
- How would you like the world to look like? Could you do something to further that?

When you've agreed on the goal, write it down here:

Step 2: Reality

What's the current situation relative to the goal?

At this point, the coach shouldn't make any suggestions! Rather, the coach should try to understand the coachee's situation as well as possible. A useful technique is to regularly summarize and rephrase what the coachee has said so far, letting the coachee correct or elaborate as needed.

If there any suggestions that come to mind for the coach, they should write them down for later or try to present questions that test any assumptions that underlie the suggestions.

For example, "this person is unable to concentrate, he should try meditation" assumes that being unable to concentrate in some particular situation is a problem, and that the cause isn't in the environment. A way to test those assumptions would be to ask what the coachee thinks is behind concentration being difficult, and whether he feels it would be useful to try to do something about it.

When you're in agreement about the current situation, write down a brief summary that both of you can agree on:

Step 3: Options / Obstacles

Support the coachee in finding ways to achieve the goal. One good way to do this is to present solution-focused questions. For example:

- "If you found a solution to this, what could it look like?"
- "If things magically looked better, how would you notice the difference?"
- "What ways do you have to achieve this?"
- "Have there been times when you've done something that's been of help?"
- "Have there been times when things were better / further along? What was different then? Did you do something differently?"
- "How bad is this problem on a scale from 1 to 10, where 1 is the worst it could ever be, and 10 is the best it could ever be?"
    - "What's stopping you from slipping one point lower on the scale?"
    - "What would it look like if you were half a point higher? How could you get there?"
    - "Where on the scale would be good enough? What would a day at that point on the scale feel like; what would you do differently?"

At this point the coach may possibly offer suggestions, **if** the coachee seems to want them and be receptive to them.

Write down a few sentences worth of summary of the options and obstacles you've found, that both can agree on:

Step 4: Way Forward

From the different options that you've identified, choose one in particular. Decide together the concrete next steps that the coachee will make, and encourage them to define in as much detail as possible where, when, and how they will take those steps. Write this down below.

Before settling on the final results, the coach should ask the coachee, "if an infallible crystal ball told you that your plan is going to fail, what would be the most likely reason?". See if the coachee would like to modify their plan based on the answer.

2017: Feeling the need to go more specific still

This felt somewhat useful, and people did seem to get some benefits, but personally I noticed myself again losing motivation with this kind of an approach. An issue that I frequently had was that I would discuss some issue that I had in a coaching session, think of some possible new approach... and then end up with a pretty high certainty that this too would fail to address the actual problem, even though I couldn't figure out *why* it would.

Then in the summer of 2017 I ended up stumbling on a memory reconsolidation technique that fixed a large chunk of my problems, while before/afterwards also generally starting to dig more into approaches such as Focusing, Core Transformation, Internal Family Systems, Coherence Therapy, etc., which seemed to be much more effective at digging into the roots of the actual problems.

My general current feeling is that approaches such as debugging and coaching are doubtlessly useful in some specific situations, such as when your problem is mostly just a lack of ideas or knowledge, or not having articulated your goals clearly enough. But my experience is that many (if not most) of the internal problems that a person has are the consequence of specific emotional learnings which the person's mind feels *necessary* to actively maintain. As such, I find that the most effective approaches for improving people's ability to better their own lives are not generalized approaches such as debugging, but rather specialized protocols aimed at uncovering and changing these emotional learnings: ones found in places such as Coherence Therapy, Focusing, and Internal Family Systems therapy. And since other people have already developed good protocols for those, it seems unnecessary for me to try to re-invent the wheel by developing my own guidelines.

At the same time, as these kinds of approaches may involve going deep into one's emotions and past sources of trauma, I have felt uncomfortable with the thought of running them in public groups with random volunteers, so have shifted to mainly doing one-on-one IFS sessions with friends who express interest in them.

Brainstorming positive visions of AI

Context

This is a place to explore visions of how AI can go really well. Conversations about AI (both in this community and disseminated by mainstream media) focus on dystopian scenarios and failure modes. Even communities that lean technoutopian (Silicon Valley) are having an AI hangover. More broadly, many people in my life think the future will be worse than the present and this makes me sad.

So I think it's time to revisit the science fiction books of our teenage years and imagine what amazing applications of AI or AGI in society looks like. AI that doesn't destroy us is great. AI that unlocks human flourishing is even better. I've personally found it much easier to think of negative scenarios than positive scenarios, so this is me enlisting your help.

Discussion norms

I'd like this to be a yes-and, generative, bursty, collective brainstorming thread. Ideas that make no sense, that you might not endorse later, that you can't even explain or defend are all welcome. The person that starts the idea might not be the person that finishes it so don't censor too early.

I hope this discussion style shows us a different way to get to better reasoning: as opposed to one person contributing a well-developed position and others pointing out inconsistencies, we can build ideas collectively, with everyone contributing different parts. I also hope it encourages people who don't normally participate on LessWrong to contribute their thoughts.

Prompts for idea generation

(Feel free to ignore and share whatever comes to mind)

1. What does AI going well look like?
2. If you feel pessimistic about the future, what vision of AI would make you feel more optimistic?
3. What are problems we don't know how to solve that you'd be excited to solve with AI?
4. What great companies could be built in the next decade, where the product is built primarily around modern Machine Learning? (If you find this prompt overly large, can you think of any companies that could be built around GPT-3?)

I'll seed some of my ideas below :) Thanks Amanda, Ben, Andreas for discussing these ideas with me.

Draft papers for REALab and Decoupled Approval on tampering

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

Hi everyone, we (Ramana Kumar, Jonathan Uesato, Victoria Krakovna, Tom Everitt, and Richard Ngo) have been working on a strand of work researching tampering problems, and we've written up our progress in two papers. We're sharing drafts in advance here because we'd like to get feedback from everyone here.

The first paper covers:

- How and when tampering problems might arise in the real world
- Key assumptions in standard RL frameworks we relax to allow modeling tampering
- How we model and measure tampering empirically, through our internal platform REALab, and
- How we formalize tampering problems, through our Corrupt Feedback MDP formalism

We particularly hope it clears up the concept of tampering (and why "but the agent maximized its given reward function" typically assumes the wrong framing), and internally, we've found REALab to be a useful mental model.

The second paper describes:

- Decoupled approval, an algorithm closely related to approval direction and Counterfactual Oracles, and designed to be straightforwardly compatible with standard deep RL
- An analysis of this algorithm (within the CFMDP formalism), and
- Empirical validation (in REALab)

We'd love to get feedback on these; the current drafts are viewable in this Google Drive folder. We're happy to discuss these on whichever of LessWrong/Alignment Forum/Google Drive comments, and would prefer to keep discussion on these forums for now, as we'll share the papers more widely after they're posted on arXiv in a few weeks. Looking forward to hearing your thoughts!

Things are allowed to be good and bad at the same time

I've found it useful to sometimes remind myself that things are allowed to be good *and* bad at the same time.

Suppose that there was a particular job that I wanted but didn't get. Afterwards, I find myself thinking:

"Damnit, some of the stuff in that job would have been *so* cool."
"But the commute would have killed me, it wouldn't have been a good fit for me anyway."
"But some of that stuff would have been so cool!"
"Still, the commute would have killed me."
"Yeah but…"

It's as if my mind is trying to decide whether I should be upset at not getting the job, or happy at having dodged a bullet.

And if I pay close attention to what's happening in my head, I might notice something.

It's as if my mind is acting in such a way that only *one* of these options might be true:

Either some of the stuff in the job would been really cool, *or* the commute would have killed me. When one consideration is brought up, it's as if it "cancels" the other.

Now if I were trying to decide whether I want the job or not, then this might make some sense: either the work is cool enough to overwhelm the badness of the commute, *or* the commute is bad enough to overwhelm the coolness.

But I already know that I didn't get the job, so I don't need to make a binary decision. And even if I *did* need to make a binary decision, realistically it's not that one of the considerations makes the other completely irrelevant. My mind is trying to persuade me that the job is either all good or all bad, and neither of those assumptions is likely to be a healthy basis for a decision.

So there's a thing where I… let my mind accept that both the good *and* the bad can be true at the same time, and that that's all there is to say about it.

Yes, that stuff *would* have been really cool. And yes, the commute *would* have killed me. And there doesn't *need* to be an "overall goodness" of the job that would be anything else than just the combination of those two facts.

This can feel a bit like there's an electrifying zap in my mind, the two facts merging together into an overall judgment, and then there's nothing more to consider. It doesn't exactly feel *good*, the way it would have if I'd concluded that I never should have wanted the job anyway. But it also doesn't feel *bad*, the way it would if I'd concluded that I actually really would have wanted it.

It just is what it is: a job that would've had some really cool aspects, where the commute would have killed me.

And then there’s so obviously nothing else to say about it, and I can move on.

—

Other uses of the principle:

My friend Marras describes finding relief from chronic pain through a similar dual acceptance: “Yes, my shoulder is in pain. Other parts are not. I can enjoy the other parts while I suffer from the small but upsetting bad aspect. I don’t need to argue for or against either.”

Anders Sandberg notes that this is also a good way for thinking about the state of the world in general: “a lot of things are going really well *and* a lot of other things are going badly.”

Babble challenge: 50 ways of solving a problem in your life

Back again. Let's become stronger.

This week's challenge:

Years ago you found yourself hurled into existence, facing a vast universe with a mind capable of the Art of Rationality, reading a LessWrong post at this very moment.

Yet in your life there is a particular problem. I don't know what it is. Maybe your chair is uncomfortable; you're not getting as high scores as you want at the Math Olympiad; or you've got insomnia.

Whatever it is, pick one specific problem in your life.

Find a way to solve it.

You have 1 hour to come up with 50 ways.

(But no need to implement the solutions within 1 hour!)

Looking back

Here are the champions who made it to 50 last week, with stars indicating their streak:

★★★ Slider, gjm, Harmless, jacobjacob, Tetraspace Grouping

★★ athom, johnswentworth, ryan_b, Ericf, Bucky, Mark Xu, CptDrMoreno, Yonge

★ TurnTrout, Tighe, knite

Why measure streaks?

Last week Bucky commented:

> I don't like measuring things by streaks - if you want to do a list I think doing it by total number of challenges completed is better. Streaks are a less accurate indication of effort put in or potential gains achieved and have more potential to create unhealthy incentives.

But I disagree. I replied:

> One of the goals of the challenge is building a culture of practice. I think *consistency* is an incredibly important part of that. That's how you get compound returns. A portfolio that grows 7% *every* year will grow ~30x over fifty years. But a portfolio that grows that much only *every other* year will only grow about ~5x. (Even though the first one only put in "twice as much effort".)

Moving forwards

I’m now entering week 4 out of the 7-week babble streak I committed to. If you want more regularity in practicing your creativity, feel free to post a comment committing to also going all the way to 7.

This week we’re trying something new: applied babble. I haven’t tried it before, so am very curious to see what will happen. Feel free to add a note to your comment about how useful you found the exercise, and whether you thought about good things you hadn’t considered before.

Rules

- **50 answers or nothing. Shoot for 1 hour.**

Any answer must contain 50 ideas to count. That’s the babble challenge.

However, the 1 hour limit is a stretch goal. It’s fine if it takes longer to get to 50.

- **Post your answers inside of spoiler tags.** (How do I do that?)
- **Celebrate other’s answers.**

This is really important. Sharing babble in public is a scary experience. I don’t want people to leave this having back-chained the experience “If I am creative, people will look down on me”. So be generous with those upvotes.

If you comment on someone else’s post, focus on making exciting, novel ideas work — instead of tearing apart worse ideas.

Reward people for babbling — don’t punish them for not pruning.

I might remove comments that break this rule.

- **Not all your ideas have to work.**

The prompt is very underspecified. If your chair is uncomfortable, consider sitting on a sofa, on the ground, in a pool, or on a trampoline. I've often found that 1 great idea can hide among 10 bad ones. You just need to push through the worse ones. Keep talking. To adapt Wayne Gretzky's great quote: "You miss 100% of the ideas you never generate."

- **My main tip: when you’re stuck, say something stupid.**

If you spend 5 min agonising over not having anything to say, you’re doing it wrong. You’re being too critical. Just lower your standards and say something, anything. Soon enough you’ll be back on track.

This is really, really important. It’s the only way I’m able to complete these exercises.

—

Now, go forth and babble! 50 ways of solving a problem in your life!

Additive Operations on Cartesian Frames

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

The mathematical object (but not the philosophical interpretation) of a Cartesian Frame is studied under the name "Chu space."

(In category theory, Chu spaces are usually studied in the special case of $W = 2$. To learn more about Chu spaces, see Vaughan Pratt's guide to papers and *n*Lab's page on the Chu construction.)

In this post and the next one, I'll mostly be discussing standard facts about Chu spaces. I'll also discuss how to interpret the standard definitions as statements about agency.

Chu spaces form a category as a special case of the Chu construction. You may notice a strong similarity between operations on Cartesian frames and operations in linear logic, coming from the fact that the Chu construction is also intimately related to linear logic, and is used in the semantics for linear logic.

Linear logic has a large number of operations—additive conjunction (&), multiplicative conjunction (⊗), and so on—and many of those symbols will turn out to have interpretations for Cartesian frames, and they're actually going to be meaningful interpretations in this setting. For that reason, we'll be stealing much of our notation from linear logic, though this sequence won't assume familiarity with linear logic.

Definition: Chu(W) is the category whose objects are Cartesian frames over W, whose morphisms from $C = (A, E, \cdot)$ to $D = (B, F, \star)$ are pairs of functions $(g : A \to B, h : F \to E)$, such that $a \cdot h(f) = g(a) \star f$ for all $a \in A$ and $f \in F$, and whose composition of morphisms is given by $(g_1, h_1) \circ (g_0, h_0) = (g_1 \circ g_0, h_0 \circ h_1)$.

The composition of two morphisms $C_0 \to C_1$ and $C_1 \to C_2$, then, sends the agent of C_0 to C_2 and sends the environment of C_2 to C_0.

Claim: Chu(W) is a category.

Proof: It suffices to show that composition is well-defined and associative and there exist identity morphisms. For identity, (id_A, id_E) is clearly an identity on $C = (A, E, \cdot)$,

where id_X is the identity map from X to itself.

The composition $(g_1, h_1) \circ (g_0, h_0)$ of $(g_0, h_0) : C_0 \to C_1$ with $(g_1, h_1) : C_1 \to C_2$ is $(g_1 \circ g_0, h_0 \circ h_1) : C_0 \to C_2$. To verify that this is a morphism, we just need that $a_0 \cdot_0 h_0(h_1(e_2)) = g_1(g_0(a_0)) \cdot_2 e_2$ for all $a_0 \in \mathrm{Agent}(C_0)$, and $e_2 \in \mathrm{Env}(C_2)$, where $\cdot_i = \mathrm{Eval}(C_i)$. Indeed,

$$\begin{aligned} a_0 \cdot_0 h_0(h_1(e_2)) &= g_0(a_0) \cdot_1 h_1(e_2) \\ &= g_1(g_0(a_0)) \cdot_2 e_2, \end{aligned}$$

since each component is a morphism.

Associativity of the composition follows from the fact that it is just a pair of compositions of functions on sets, and composition is associative for sets. □

1. What Do These Morphisms Represent?

1.1. Morphisms as Interfaces

A Cartesian frame is a first-person perspective. The agent A finds itself in a certain situation or game, where it expects to encounter an environment E. The morphisms in Chu(W) allow the agent of one Cartesian frame to play the game of another Cartesian frame.

We can think of the morphisms from $C = (A, E, \cdot)$ to $D = (B, F, \star)$ as ways of fitting the agent of C into the environment of D. Indeed, for every morphism $(g, h) : C \to D$, one can construct a Cartesian frame $(A, F, \diamond)$, whose agent matches C's agent, and whose environment matches D's environment, with $\diamond$ given by $a \diamond f = a \cdot h(f) = g(a) \star f$. (The morphism from C to D can actually be viewed as the composition of $(\mathrm{id}_A, h) : C \to (A, F, \diamond)$ with $(g, \mathrm{id}_F) : (A, F, \diamond) \to D$.)

Two random large Cartesian frames will typically have no morphisms between them. When there *is* a morphism, the morphism functions as an interface that allows the agent A to interact with some other environment F. However, we aren't just randomly

throwing A and F together. A's interaction with F factors through the function $h : F \to E$, so A can in a sense still be thought of as using an interface where it interacts with E. It just interacts with an $e \in E$ that is of the form h(f) for some $f \in F$. But this is happening simultaneously with the dual view in which F can be thought of as still interacting with B!

Since a Cartesian frame is a first-person perspective, you can imagine A having the internal experience of interacting with E, while F has the "experience" of interacting with B. The morphism's job is to be the translation interface that allows this A and F to interact with each other, while preserving their respective internal experiences in such a way that they feel like they're interacting with E and B respectively. A gets to play B 's game, while still thinking that it is playing its own game.

1.2. Morphisms as Differences in Agents' Strength

We can also interpret the existence of a morphism from $C = (A, E, \cdot)$ to $D = (B, F, \star)$ as saying something like "D's agent is at least as strong as C's agent."

This is easiest to see for a morphism $(g, h) : C \to D$ where g and h are both injective. In this case, it is as though $A \subseteq B$ and $F \subseteq E$, so D's agent has more options to choose between and fewer environments it has to worry about.

Since some of the environments in $E \setminus F$ might have been good for the agent, the agent isn't necessarily strictly better off in D; but in a zero-sum game, the agent will indeed be strictly better off. I think this justifies saying that C's agent is in some sense weaker than D's agent.

If g or h is not injective, we could duplicate elements of B and E to make it injective, so the interpretation "C's agent is no stronger than D's agent" is reasonable in that case as well. In particular, the existence of a morphism from C to D implies that Ensure(C) ⊆ Ensure(D) (and thus Ctrl(C) ⊆ Ctrl(D)).

However, the existence of a morphism is stronger than just saying the set of ensurables is larger. The morphism from C to D can be thought of as telling D's agent how to strategy-steal from C's agent, and thus do anything that C's agent can do.

We now provide a few examples to illustrate morphisms between Cartesian frames. (If you're ready to forge ahead, skip to §2 instead.)

1.3. Simple Examples of Morphisms

Imagine a student who is deciding between staying up late studying for a test (a_s) or ignoring the test (a_i). We will represent the student with a Cartesian frame over letter grades, where $W = \{A+, A, A-, B+, B, B-, C+, C, C-, D+, D, D-, F\}$.

If the student doesn't study, her final grade is always a C+, represented by the possible world C+. If she does study, she may oversleep and get a bad grade (represented by the environment selecting e_o and putting her in D-). If she studies and doesn't oversleep, she is uncertain about whether her teacher is typical (e_t, resulting in A-) or unusually demanding (e_d, resulting in B+). We represent this with the frame

$$C_T = \begin{array}{c} \\ a_s \\ a_i \end{array} \begin{array}{c} \begin{array}{ccc} e_t & e_d & e_o \end{array} \\ \begin{pmatrix} A- & B+ & D- \\ C+ & C+ & C+ \end{pmatrix} \end{array}.$$

Let us also suppose that yesterday, the student had the extra option of copying another student's answers on test day to get a sure A+. However, she decided not to cheat. We represent the student's options yesterday, prior to precommitting, with the frame

$$C_Y = \begin{array}{c} \\ b_s \\ b_i \\ b_c \end{array} \begin{array}{c} \begin{array}{ccc} f_t & f_d & f_o \end{array} \\ \begin{pmatrix} A- & B+ & D- \\ C+ & C+ & C+ \\ A+ & A+ & A+ \end{pmatrix} \end{array}.$$

There is a morphism from the student's frame today to her frame yesterday, representing the fact that Agent(C_T) can be plugged into Agent(C_Y)'s game, or that the student was "stronger" yesterday than she is today.

Let us also suppose that the student's teacher *is* in fact demanding. If the student today knew this fact, we would instead represent her perspective with the frame

$$C_{T'} = \begin{array}{c} \\ a_s' \\ a_i' \end{array} \begin{array}{c} \begin{array}{cc} e_d' & e_o' \end{array} \\ \begin{pmatrix} B+ & D- \\ C+ & C+ \end{pmatrix} \end{array}.$$

Here, we have a morphism from the student today (C_T) to her perspective if she had an additional promise from the environment ($C_{T'}$). This represents the fact that $C_{T'}$ can strategy-steal from a version of herself who knows strictly less.

Given two Cartesian frames C_0 and C_1, I am not aware of an efficient universal method for determining whether there exists a morphism from C_0 to C_1. Indeed, I conjecture that this problem might be NP-complete. In the above cases, however, we can see that there exist morphisms from C_T to the other two frames by observing that C_T is effectively C_Y with a row deleted, or $C_{T'}$ with a column added.

While Agent(C_Y) and Agent($C_{T'}$) are both stronger than Agent(C_T), we have no morphisms between C_Y and $C_{T'}$; their options are different enough that we can't compare their strength directly.

1.4. Examples of Morphisms Going Both Ways

Every Cartesian frame has an identity morphism pointing to itself; and as we'll discuss in the next post, whenever two Cartesian frames C and D are equivalent (in a sense to be defined), there will be a morphism going from C to D and another going from D to C. But not all pairs of Cartesian frames with morphisms going both ways are equivalent. Consider, for example,

$$C_1 = \begin{array}{c} \\ a_0 \\ a_1 \end{array} \begin{array}{c} \begin{array}{cc} e_0 & e_1 \end{array} \\ \begin{pmatrix} w_0 & w_0 \\ w_0 & w_1 \end{pmatrix} \end{array} \text{ and } D_1 = \begin{array}{c} \\ b_0 \end{array} \begin{array}{c} \begin{array}{c} f_0 \end{array} \\ \begin{pmatrix} w_0 \end{pmatrix} \end{array}.$$

In $C_1 = (A, E, \cdot)$, the default outcome is w_0, but the agent and environment can handshake to produce w_1. In $D_1 = (B, F, \star)$, there are no choices, and there's only one

possible world, w_0.

It turns out that there is a morphism $(g, h) : C_1 \to D_1$, where g is the constant function b_0 and h is the constant function e_0; and there is a second morphism $(g', h') : D_1 \to C_1$, where g' is the constant function a_0 and h' is the constant function f_0. We can interpret these like so:

- There is a morphism $C_1 \to D_1$ because D_1 is effectively C_1 plus a promise from the environment "I'll choose e_0." The agent in D_1 is "stronger" in the sense that it has fewer possible environments to worry about. There is less the environment can do to interfere with the agent's choices.
- There is a morphism $D_1 \to C_1$ because C_1's agent has strictly more options than D_1's agent: moving from D_1 to C_1 lets you retain the option to produce w_0 if you want, but it also lets you try for w_1.

So we can view the smaller matrix as the larger matrix plus a promise from the environment "I'll choose e_0," or we can view it as the larger matrix plus a commitment from the agent "I'll choose a_0."

This example demonstrates that my intuitive statement "wherever there's a morphism from C to D, D is at least as strong as C" conflates two different notions of "stronger." These notions often go together, but come apart in situations such as the handshake example. Like the hypothetical student in $C_{T'}$, the agent of D_1 is "stronger" in the sense that the environment can't do as much to get in the way. But like the not-yet-precommitted student in C_Y, the agent of C_1 is "stronger" in the sense that it has more options.

2. Self-Duality

A key property of Chu(W) is that it is self-dual.

Definition: Let $-^* : \mathrm{Chu}(W) \to \mathrm{Chu}(W)^{\mathrm{op}}$ be the functor given by $(A, E, \cdot)^* = (E, A, \star)$, where $e \star a = a \cdot e$, and $(g, h)^* = (h, g)$.

The more standard notation for dual in linear logic would be $-^{\perp}$, but this is horrible notation.[1]

Claim: $-^*$ is an isomorphism between Chu(W) and Chu(W)^{op}.

Proof: First, we show $-^*$ is a functor. The objects in Chu(W)^{op} are the same as in Chu(W), the morphisms from D to C in Chu(W)^{op} are the morphisms from C to D in Chu(W), and composition is the same, but with the order reversed. $-^*$ clearly preserves identity morphisms. To show that $-^*$ preserves composition, we have

$$\begin{aligned}(g_0, h_0)^* \circ^{op} (g_1, h_1)^* &= (h_1, g_1) \circ (h_0, g_0) \\ &= (h_1 \circ h_0, g_0 \circ g_1) \\ &= ((g_0, h_0) \circ (g_1, h_1))^*.\end{aligned}$$

To see that it is an isomorphism, we need a left and right inverse. We will abuse notation and also write $-^*$ for the functor from Chu(W)^{op} to Chu(W) given by $(E, A, \star)^* = (A, E, \cdot)$, where $a \cdot e = e \star a$, and $(h, g)^* = (g, h)$. Clearly, we have $-^* : \text{Chu}(W) \to \text{Chu}(W)^{op}$ and $-^* : \text{Chu}(W)^{op} \to \text{Chu}(W)$ composing to the identity in both orders, so $-^*$ is an isomorphism. □

Going back to our visualization of Cartesian frames as matrices, $-^*$ just takes the transpose of the matrix, swapping agent with environment. "Chu(W) is self-dual" is another way of saying that transposing a Cartesian frame always gives you another Cartesian frame.

Philosophically, depending on our interpretation, this may be doing something weird. We talk about possible agents and possible environments, but we may mean something different by "possible" in those two cases.

Since we are imagining events from the point of view of the agents, "possible agents" is referring to all of the ways the agent can choose to be by exercising its "free will." We could think of "possible environments" similarly, or we could think of possible environments as representing the agent's uncertainty.

Under the view where possible environments represent uncertainty, $-^*$ is pointing to an interesting duality that swaps choices with uncertainty, swaps the "could" of "I could do X" with the "could" of "The world could have property Y," and (if we add probability to the mix) swaps mixed strategies with probabilistic uncertainty. "What

will I do?" becomes "What game am I playing?", or "What is the world-as-a-function-of-my-action like?"

I will introduce many operations on Cartesian frames, so it will help to highlight even the basic properties as I go. Here, I'll note:

Claim: For any Cartesian frame C, $(C^*)^* = C$.

Proof: Trivial. □

3. Sums of Cartesian Frames

The first binary operation on Cartesian frames I want to introduce is the sum, $\oplus$.

Definition: For Cartesian frames $C = (A, E, \cdot)$ and $D = (B, F, \star)$ over W, $C \oplus D$ is the Cartesian frame $(A \sqcup B, E \times F, \diamond)$, where $a \diamond (e, f) = a \cdot e$ if $a \in A$, and $a \diamond (e, f) = a \star f$ if $a \in B$.

The sum takes the disjoint union of the agents and the Cartesian product of the environments, and does the obvious thing with the evaluation function. The agent can choose any strategy from A or from B, and the environment has to respond to that strategy. We can interpret this as an agent that can choose between two different first-person perspectives: it can decide to interact with the environment as the agent of C, or as the agent of D.

Maybe "Rebecca the chess player" is considering which chess opening to employ, whereas "Rebecca the food-eater" is considering putting her plate down on the chess board and having lunch instead. "Rebecca the agent that can choose between playing chess and having lunch" is the sum of the other two Rebeccas.

If Rebecca tunnel-visions on the chess game, she may not consider her other options. Likewise if she tunnel-visions on lunch. If she inhabits the perspective of the third Rebecca, she can instead decide between chess moves *and* decide whether she wants to be playing chess at all.

Meanwhile, the environment must use a policy that selects an option from E if the agent chooses from A, and selects an option from F if the agent chooses from B.

In the chess example: The environment must be able to respond to different chess moves, but it must also be able to respond to Rebecca deciding to play a different game.

To give a formal example, let $C_2 = (A, E, \cdot)$ and $D_2 = (B, F, \star)$ be given by the matrices

$$C_2 = \begin{array}{c} \\ a_0 \\ a_1 \end{array} \begin{array}{c} \begin{array}{cc} e_0 & e_1 \end{array} \\ \begin{pmatrix} w_0 & w_1 \\ w_2 & w_3 \end{pmatrix} \end{array} \text{ and } D_2 = \begin{array}{c} \\ b_0 \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{ccc} f_0 & f_1 & f_2 \end{array} \\ \begin{pmatrix} w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \\ w_{10} & w_{11} & w_{12} \end{pmatrix} \end{array}.$$

Here, $C_2 \oplus D_2$ is given by

$$C_2 \oplus D_2 = \begin{array}{c} \\ a_0 \\ a_1 \\ b_0 \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{cccccc} e_0f_0 & e_0f_1 & e_0f_2 & e_1f_0 & e_1f_1 & e_1f_2 \end{array} \\ \begin{pmatrix} w_0 & w_0 & w_0 & w_1 & w_1 & w_1 \\ w_2 & w_2 & w_2 & w_3 & w_3 & w_3 \\ w_4 & w_5 & w_6 & w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 & w_7 & w_8 & w_9 \\ w_{10} & w_{11} & w_{12} & w_{10} & w_{11} & w_{12} \end{pmatrix} \end{array}.$$

If we wish to interpret $C_2 \oplus D_2$ temporally, we can say: The agent first chooses what game to play. The environment then, as a function of which game was chosen, "chooses" what it does; and the agent simultaneously chooses its own move within the game it picked.

Definition: Let 0 be given by the Cartesian frame $0 = (\{\}, \{e\}, \cdot)$, where Agent(0) is the empty set, $Env(0) = \{e\}$ is any singleton set, and Eval(0) is trivial, since it has empty domain.

Claim: $\oplus$ is commutative and associative, and 0 is the identity of $\oplus$ (up to isomorphism).

Proof: Trivial. □

Returning to our interpretation of morphisms as differences in agents' strength: The agent of $C \oplus D$ can choose between being the agent from C or the agent from D, and so is stronger than either. Indeed, we can think of $C \oplus D$'s agent as the weakest agent

that is stronger than both C's agent and D's agent. Mathematically, this translates to $\oplus$ being the categorical coproduct in Chu(W).

Theorem: $C_0 \oplus C_1$ is the coproduct of C_0 and C_1 in Chu(W), and 0 is initial in Chu(W).

Proof: First, we show that 0 is initial. We want to show that there exists a unique morphism from 0 to a given C. Indeed, a morphism from 0 to $C = (A, E, \cdot)$ is a function from $\{\}$ to A along with a function from E to $\{e\}$, and there is always exactly one such pair of functions, regardless of what A and E are. It is also easy to see that this pair of functions is a morphism, since the condition for morphism is empty, since Agent(0) is empty. Thus 0 is initial.

Let $C_i = (A_i, E_i, \cdot_i)$, and let $C_0 \oplus C_1 = (A_0 \sqcup A_1, E_0 \times E_1, \diamond)$. We want to show that there exist inclusion morphisms $\iota_0 : C_0 \to C_0 \oplus C_1$ and $\iota_1 : C_1 \to C_0 \oplus C_1$ such that for any Cartesian frame $D = (B, F, \star)$, and any pair of morphisms $\phi_0 : C_0 \to D$ and $\phi_1 : C_1 \to D$, we have that there exists a unique morphism $\phi : C_0 \oplus C_1 \to D$ such that $\phi \circ \iota_0 = \phi_0$ and $\phi \circ \iota_1 = \phi_1$.

First, we need to specify $\iota_i : (A_i, E_i, \cdot_i) \to (A_0 \sqcup A_1, E_0 \times E_1, \diamond)$. We let $\iota_i = (j_i, k_i)$, where $j_i : A_i \to A_0 \sqcup A_1$ is just the the obvious inclusion of A_i into $A_0 \sqcup A_1$, and $k_i : E_0 \times E_1 \to E_i$ is just the obvious projection. This is clearly a morphism.

Given $\phi_0 = (g_0, h_0) : C_0 \to D$ and $\phi_1 = (g_1, h_1) : C_1 \to D$, we let $\phi = (g, h)$, where $g : A_0 \sqcup A_1 \to B$ is given by $g(a) = g_i(a)$ where i is such that $a \in A_i$, and $h : F \to E_0 \times E_1$ is given by $h(f) = (h_0(f), h_1(f))$. This is a morphism because for all $a \in A_0 \sqcup A_1$ and $f \in F$, we have

$$\begin{aligned} a \diamond h(f) &= a \cdot_i h_i(f) \\ &= g_i(a) \star f \\ &= g(a) \star f, \end{aligned}$$

where i is such that $a \in A_i$. It is clear from the definitions that $\phi \circ \iota_i = \phi_i$.

Finally, we need to show the uniqueness of this ϕ. Let $\phi' = (g', h') : C_0 \oplus C_1 \to D$ be a morphism such that $\phi' \circ \iota_i = \phi_i$ for both $i = 1, 2$. This means that $g'(a) = g_i(a)$ when $a \in A_i$, so $g'(a) = g(a)$ for all $a \in A_0 \sqcup A_1$. Similarly, $h'(f)$ must project to $h_0(f)$ and $h_1(f)$, so

$$\begin{aligned} h'(f) &= (h_0(f), h_1(f)) \\ &= h(f) \end{aligned}$$

for all $f \in F$. Thus $\phi' = \phi$. □

4. Products of Cartesian Frames

Dual to sum, we have the product operation, &. This operation is a product. It is also in the section on additive operations. There are many counterintuitive things about the notation of Chu spaces and linear logic.

Definition: For Cartesian frames $C = (A, E, \cdot)$ and $D = (B, F, \star)$ over W, C&D is the Cartesian frame $(A \times B, E \sqcup F, \diamond)$, where $(a, b) \diamond e = a \cdot e$ if $e \in E$, and $(a, b) \diamond e = b \star e$ if $e \in F$.

C&D means that the agent might have to be the agent of C, and might have to be the agent of D, but does not get to decide which one. Thus, it will have to choose a pair, (a, b), where a says how to behave in a C situation, and b says how to behave in a D situation. The environment will "choose" to either be C's environment or D's environment. When the agent and environment interact, the agent uses the component of its pair that matches the environment's choice.

Instead of thinking of the agent as choosing a pair, we could again think about the situation temporally. C&D is equivalent to an interaction where the environment first chooses which Cartesian frame, C or D, to play; then the agent observes this choice, and the agent and environment simultaneously behave as though they were in the chosen frame, either C or D.

(In fact, if Image(C) and Image(D) are disjoint, we can see this interpretation in the formalism by noting that Image(C) ∈ Obs(C&D)—that is, the agent can change its behavior on the basis of whether the environment selected from C or from D.)

For example, if we let C_2 and D_2 be as the example in §3,

$$C_2 = \begin{array}{c} \\ a_0 \\ a_1 \end{array} \begin{array}{c} \begin{array}{cc} e_0 & e_1 \end{array} \\ \begin{pmatrix} w_0 & w_1 \\ w_2 & w_3 \end{pmatrix} \end{array} \quad \text{and} \quad D_2 = \begin{array}{c} \\ b_0 \\ b_1 \\ b_2 \end{array} \begin{array}{c} \begin{array}{ccc} f_0 & f_1 & f_2 \end{array} \\ \begin{pmatrix} w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \\ w_{10} & w_{11} & w_{12} \end{pmatrix} \end{array},$$

then $C_2\&D_2$ is given by

$$C_2\&D_2 = \begin{array}{c} \\ a_0b_0 \\ a_0b_1 \\ a_0b_2 \\ a_1b_0 \\ a_1b_1 \\ a_1b_2 \end{array} \begin{array}{c} \begin{array}{ccccc} e_0 & e_1 & f_0 & f_1 & f_2 \end{array} \\ \begin{vmatrix} w_0 & w_1 & w_4 & w_5 & w_6 \\ w_0 & w_1 & w_7 & w_8 & w_9 \\ w_0 & w_1 & w_{10} & w_{11} & w_{12} \\ w_2 & w_3 & w_4 & w_5 & w_6 \\ w_2 & w_3 & w_7 & w_8 & w_9 \\ w_2 & w_3 & w_{10} & w_{11} & w_{12} \end{vmatrix} \end{array}.$$

A second example: Suppose that we have two Cartesian frames, C_3 and D_3. C_3 is a frame in which it's raining, and the agent chooses whether to carry an umbrella. D_3 is a frame in which it's sunny, and the agent chooses whether to carry an umbrella.

$$C_3 = \begin{array}{c} \\ u \\ n \end{array} \begin{array}{c} \begin{array}{c} r \end{array} \\ \begin{pmatrix} u\,r \\ n\,r \end{pmatrix} \end{array} \quad \text{and} \quad D_3 = \begin{array}{c} \\ u \\ n \end{array} \begin{array}{c} \begin{array}{c} s \end{array} \\ \begin{pmatrix} u\,s \\ n\,s \end{pmatrix} \end{array}$$

It turns out that the second example we provided in "Introduction to Cartesian Frames" §3.2 (Examples of Controllables) is exactly equal to the product of these two Cartesian frames,

$$\begin{array}{rc} & \begin{array}{cc} r & s \end{array} \\ \begin{array}{r} uu = u \\ nn = n \\ un = u \leftrightarrow r \\ nu = u \leftrightarrow s \end{array} & \begin{pmatrix} ur & us \\ nr & ns \\ ur & ns \\ nr & us \end{pmatrix}. \end{array}$$

The environment is the disjoint union of the rain and sun environments, and the policies of the agent can be viewed as "I get to choose what to do as a function of what game we're playing," where "what game we're playing" is "what the weather is."

Definition: Let $\top$ be given by the Cartesian frame $\top = (\{a\}, \{\}, \cdot)$, where Agent($\top$) is a singleton, Env($\top$) is the empty set, and Eval($\top$) is trivial, since it has empty domain.

Claim: & is commutative and associative, and $\top$ is the identity of & (up to isomorphism).

Proof: Trivial. □

& is essentially just $\oplus$ from the point of view of the environment. Thus, since $-^*$ swaps agent and environment, we can express & using $\oplus$ and $-^*$.

Claim: $C\&D = (C^* \oplus D^*)^*$, $\top = 0^*$, $C \oplus D = (C^*\&D^*)^*$, and $0 = \top^*$.

Proof: Trivial. □

In other words, $\oplus$ and & are De Morgan dual with respect to $-^*$.

In the same way that we interpreted $C \oplus D$ as having the weakest agent that is stronger than the agents of C and D, we can interpret C&D's agent as the *strongest* agent that is *weaker* than the agents of C and D.

Theorem: $C_0\&C_1$ is the product of C and D in Chu(W), and $\top$ is terminal in Chu(W).

Proof: Since $\oplus$ is the coproduct in Chu(W), it is the product in $\text{Chu}(W)^{\text{op}}$. Since $-^*$ is an isomorphism between Chu(W) and $\text{Chu}(W)^{\text{op}}$, we can take a product in Chu(W) of

C_0 and C_1 by sending them to $\mathrm{Chu}(W)^{\mathrm{op}}$ via this isomorphism, taking a product, and sending them back. Thus $(C_0^* \oplus C_1^*)^* = C_0 \& C_1$ is the product in $\mathrm{Chu}(W)$ of C_0 and C_1.

Similarly, since 0 is initial in $\mathrm{Chu}(W)$, it is terminal in $\mathrm{Chu}(W)^{\mathrm{op}}$. Thus, $0^* = \top$ is terminal in $\mathrm{Chu}(W)$. □

Our next post will discuss equivalence relations between Cartesian frames. We will introduce a homotopy equivalence on Cartesian frames, and employ these relations to classify small Cartesian frames up to homotopy.

Footnotes

1. One important reason $-^\perp$ is bad notation for dual is that A^B normally represents $B \to A$, where $\to$ is your category's internal hom functor. For Chu spaces, $\to$ is $\multimap$. Since $\perp$ will be the name for an object in our category, one would reasonably expect $C^\perp$ to represent $\perp \multimap C$, but it doesn't. Worse still, C^* *does* happen to be equivalent to $C \multimap \perp$, and this will be an important fact to understand. To minimize confusion, we instead use the common notation $-^*$ for dual. ↩

AGI safety from first principles: Conclusion

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

Let's recap the second species argument as originally laid out, along with the additional conclusions and clarifications from the rest of the report.

1. We'll build AIs which are much more intelligent than humans; *that is, much better than humans at using generalisable cognitive skills to understand the world*.
2. Those AGIs will be autonomous agents which pursue long-term, large-scale goals, *because goal-directedness is reinforced in many training environments, and because those goals will sometimes generalise to be larger in scope.*
3. Those goals will by default be misaligned with what we want, *because our desires are complex and nuanced, and our existing tools for shaping the goals of AIs are inadequate.*
4. The development of autonomous misaligned AGIs would lead to them gaining control of humanity's future, *via their superhuman intelligence, technology and coordination - depending on the speed of AI development, the transparency of AI systems, how constrained they are during deployment, and how well humans can cooperate politically and economically.*

Personally, I am most confident in 1, then 4, then 3, then 2 (in each case conditional on all the previous claims) - although I think there's room for reasonable disagreement on all of them. In particular, the arguments I've made about AGI goals might have been too reliant on anthropomorphism. Even if this is a fair criticism, though, it's also very unclear how to reason about the behaviour of generally intelligent systems without being anthropomorphic. The main reason we expect the development of AGI to be a major event is because the history of humanity tells us how important intelligence is. But it wasn't just our intelligence that led to human success - it was also our relentless drive to survive and thrive. Without that, we wouldn't have gotten anywhere. So when trying to predict the impacts of AGIs, we can't avoid thinking about what will lead them to choose some types of intelligent behaviour over others - in other words, thinking about their motivations.

Note, however, that the second species argument, and the scenarios I've outlined above, aren't meant to be comprehensive descriptions of all sources of existential risk from AI. Even if the second species argument doesn't turn out to be correct, AI will likely still be a transformative technology, and we should try to minimise other potential harms. In addition to the standard misuse concerns (e.g. about AI being used to develop weapons), we might also worry about increases in AI capabilities leading to undesirable structural changes. For example, they might shift the offense-defence balance in cybersecurity, or lead to more centralisation of human economic power. I consider Christiano's "going out with a whimper" scenario to also fall into this category. Yet there's been little in-depth investigation of how structural changes might lead to long-term harms, so I am inclined to not place much credence in such arguments until they have been explored much more thoroughly.

By contrast, I think the AI takeover scenarios that this report focuses on have received much more scrutiny - but still, as discussed previously, have big question marks surrounding some of the key premises. However, it's important to distinguish the question of how likely it is that the second species argument is correct, from the question of how seriously we should take it. Often people with very different perspectives on the latter actually don't disagree very much on the former. I find the following analogy from Stuart Russell illustrative: suppose we got a message from space telling us that aliens would be landing on Earth sometime in the next century. Even if there's doubt about the veracity of the message, and there's doubt about whether the aliens will be hostile, we (as a species) should clearly expect this event to be a huge deal if it happens, and dedicate a lot of effort towards making it go well. In the case of AGI, while there's reasonable doubt about what it will look like, it may nevertheless be the biggest thing that's ever happened. At the very least we should put serious effort into understanding the arguments I've discussed above, how strong they are, and what we might be able to do about them.[1]

Thanks for reading, and thanks again to everyone who's helped me improve the report. I don't expect everyone to agree with all my arguments, but I do think that there's a lot of room to further the conversation about this, and produce more analyses and evaluations of the core ideas in AGI safety. At this point I consider such work more valuable and neglected than technical AGI safety research, and have recently transitioned from full-time work on the latter to a PhD which will allow me to focus on the former. I'm excited to see our collective understanding of the future of AGI continue to develop.

---

1. I want to explicitly warn against taking this argument too far, though - for example, by claiming that AI safety work should still be a major priority even if the probability of AI catastrophe is much less than 1%. This claim is misleading because most researchers in the field of safety think it's much higher than that; and also because, if it really is that low, there are probably some fundamental confusions in our concepts and arguments that need to be cleared up before we can actually start object-level work towards making AI safer. ↩

The Colliding Exponentials of AI

Epistemic status: I have made many predictions for quantitative AI development this decade, these predictions were based on what I think is solid reasoning, and extrapolations from prior data.

If people do not intuitively understand the timescales of exponential functions, then multiple converging exponential functions will be even more misunderstood.

Currently there are three exponential trends acting upon AI performance, these being **Algorithmic Improvements, Increasing Budgets** and **Hardware Improvements.** I have given an overview of these trends and extrapolated a lower and upper bound for their increases out to 2030. These extrapolated increases are then combined to get the total *multiplier of equivalent compute* that frontier 2030 models may have over their 2020 counterparts.

Firstly...

Algorithmic Improvements

Algorithmic improvements for AI are much more well-known and quantified than they were a few years ago, much in thanks to OpenAI's paper and blog *AI and Efficiency (2020).*

OpenAI showed the efficiency of image processing algorithms has been doubling every 16 months since 2012. This resulted in a 44x decrease in compute required to reach Alexnet level performance after 7 years, as Figure 1 shows.

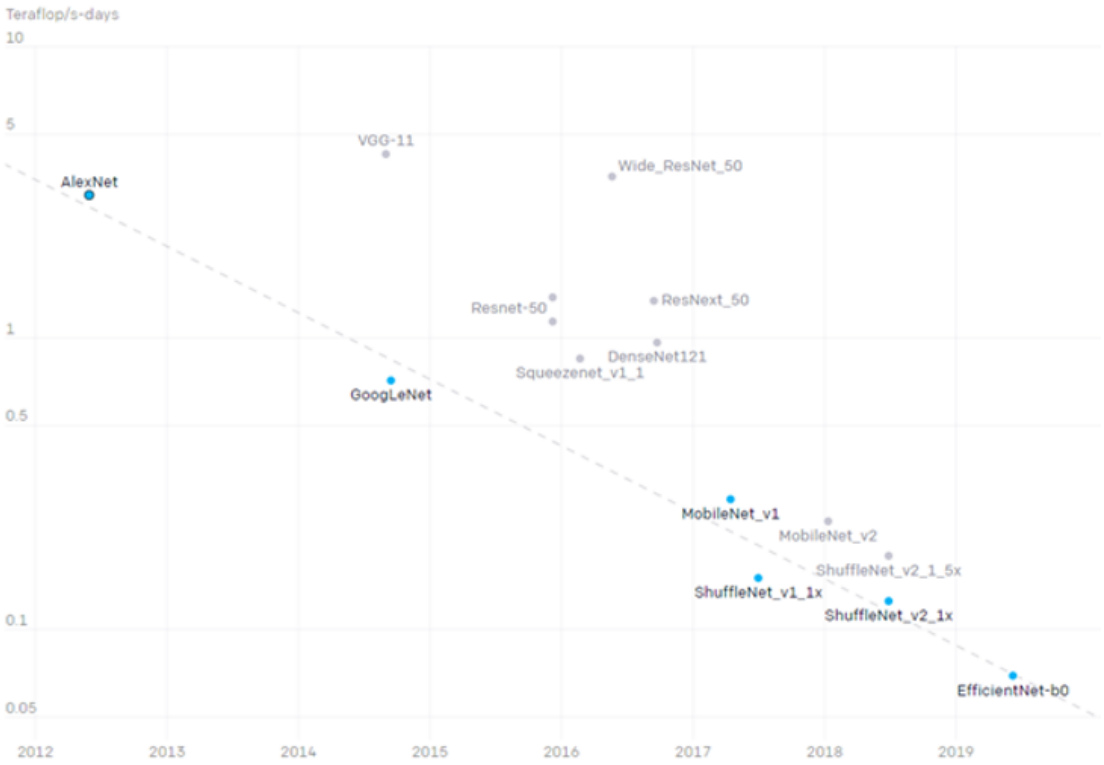


Figure 1, Compute to reach Alexnet level performance, OpenAI

OpenAI also showed algorithmic improvements in the following areas:

Transformers had surpassed seq2seq performance on English to French translation on WMT'14 with 61x less training compute in three years.

AlphaZero took 8x less compute to get to AlphaGoZero level performance 1 year later.

OpenAI Five Rerun required 5x less training compute to surpass OpenAI Five 3 months later.

Additionally Hippke's LessWrong post *Measuring Hardware Overhang* detailed algorithmic improvements in chess, finding that Deep Blue level performance could have been reached on a 1994 desktop PC-level of compute, rather than the 1997 supercomputer-level of compute that it was, if using modern algorithms.

Algorithmic improvements come not just from architectural developments, but also from their optimization library's. An example is Microsoft DeepSpeed (2020). DeepSpeed claims to train models 2–7x faster on regular clusters, 10x bigger model training on a single GPU, powering 10x longer sequences and 6x faster execution with 5x communication volume reduction. With up to 13 billion parameter models trainable on a single Nvidia V100 GPU.

So, across a wide range of machine learning areas major algorithmic improvements have been regularly occurring. Additionally, while this is harder to quantify thanks to limited precedent, it seems the introduction of new architectures can cause sudden and/or discontinuous leaps of performance in a domain, as Transformers did for NLP. As a result, extrapolating past trendlines may not capture such future developments.

If the algorithmic efficiency of machine learning in general had a halving time like image processing's 16 months, we would expect to see ~160x greater efficiency by the end of the decade. So, I think an estimate of general algorithmic improvement of **100 - 1000x** by 2030 seems reasonable.

Edit: I feel less confident and bullish about algorithmic progress now.

Increasing Budgets

The modern era of AI began in 2012, this was the year that the compute used in the largest models began to rapidly increase, with a doubling time of 3.4 months (~10x Yearly), per OpenAI's blog *AI and Compute (2018),* see Figure 2 below. While the graph stops in 2018, the trend held steady with the predicted thousands of petaflop/s-days range being reached in 2020 with GPT-3, the largest ever (non sparse) model, which had an estimated training cost of $4.6 Million, based on the price of a Tesla V100 cloud instance.

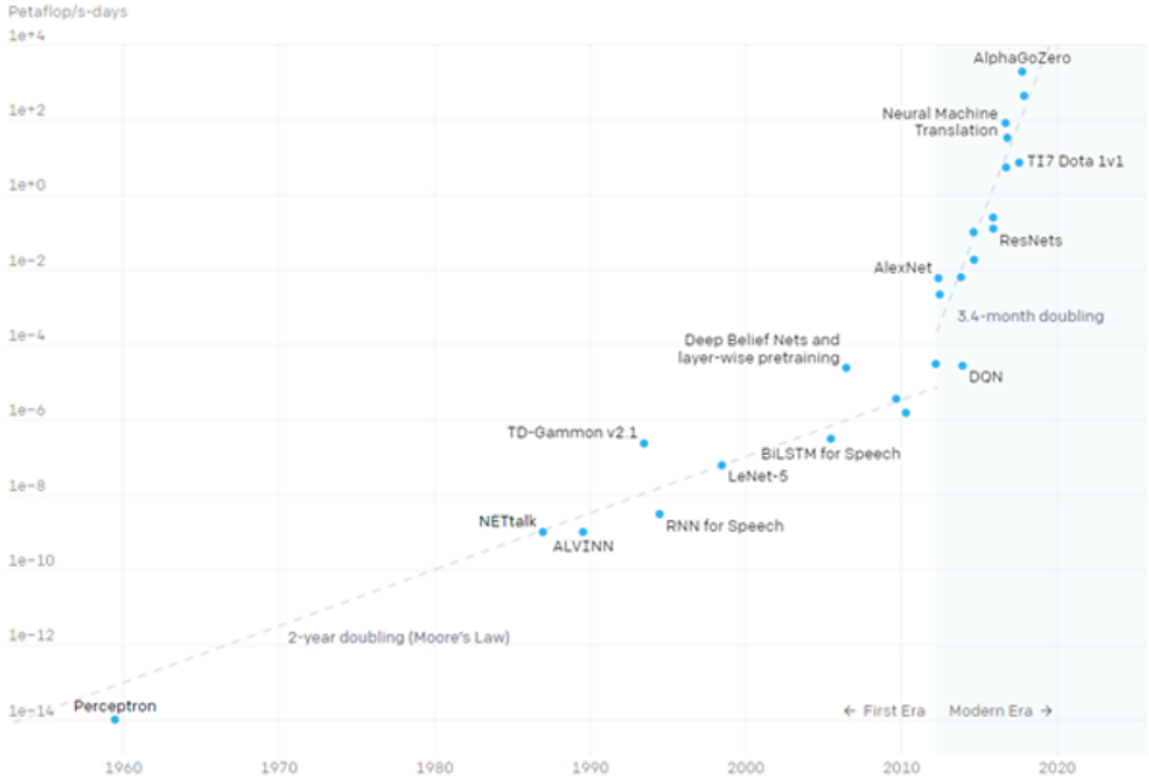


Figure 2, Modern AI era vs first Era, OpenAI

Before 2012 the growth rate for compute of the largest systems had a 2-year doubling time, essentially just following Moore's law with mostly stable budgets, however in 2012 a new exponential trend began: Increasing budgets.

This 3.4 month doubling time cannot be reliably extrapolated because the increasing budget trend isn't sustainable, as it would result in the following approximations (without hardware improvements):

2021 | $10-100M

2022 | $100M-1B

2023 | $1-10B

2024 | $10-100B

2025 | $100B-1T

Clearly without a radical shift in the field, this trend could only continue for a limited time. Astronomical as these figures appear, the cost of the necessary supercomputers would be even more.

Costs have moved away from mere academic budgets and are now in the domain of large corporations, where extrapolations will soon exceed even their limits.

The annual research and development expenditure of Google's parent company Alphabet was $26 Billion in 2019, I have extrapolated their published R&D budgets to 2030 in Figure 3.

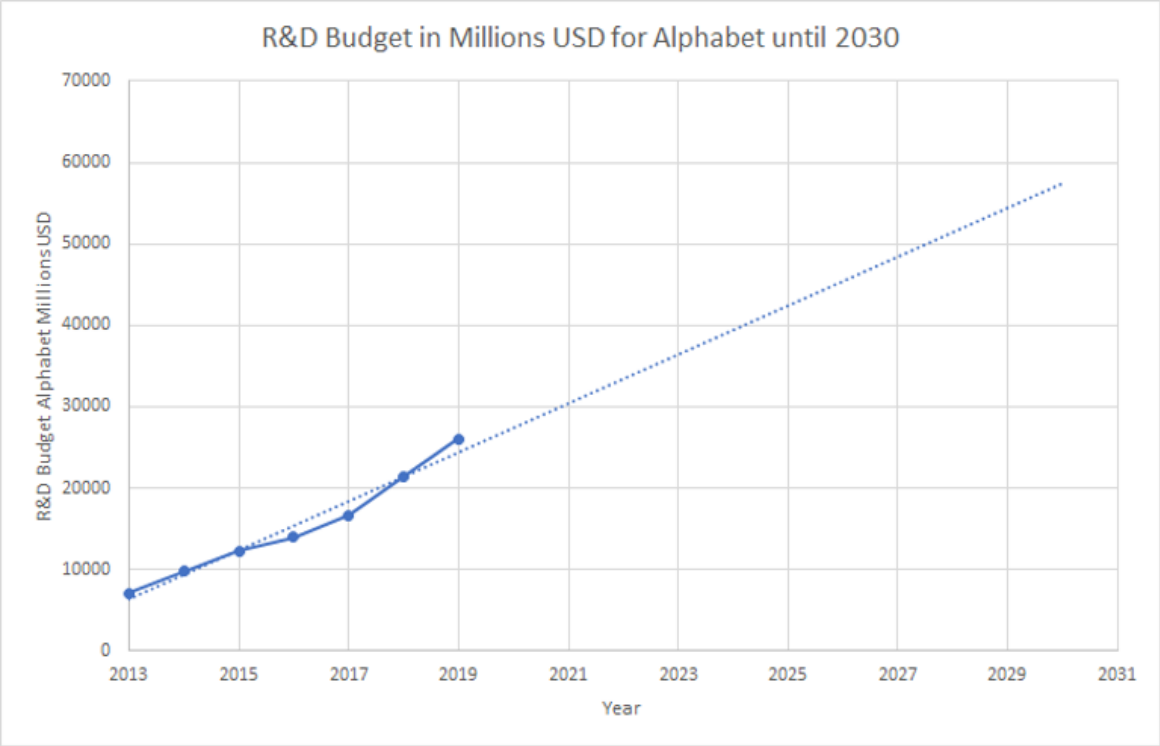


Figure 3, R&D Budget for Alphabet to 2030

By 2030 Alphabets R&D should be just below $60 Billion (approximately $46.8 Billion in 2020 dollars). So how much would Google, or a competitor, be willing to spend training a giant model?

Well to put those figures into perspective: The international translation services market is currently $43 billion and judging from the success of GPT-3 in NLP its successors may be capable of absorbing a good chunk of that. So that domain alone could seemingly justify $1B+ training runs. And what about other domains within NLP like programming assistants?

Investors are willing to put up massive amounts of capital for speculative AI tech already; the self-driving car domain had disclosed investments of $80 Billion from 2014-2017 per a report from Brookings .With those kind of figures even a $10 Billion training run doesn't seem unrealistic if the resulting model was powerful enough to justify it.

My estimate is that by 2030 the training run cost for the largest models will be in the **$1-10 Billion** range (with total system costs higher still). Compared to the single digit millions training cost for frontier 2020 systems, that estimate represents **1,000-10,000x** larger training runs.

Hardware Improvements

Moore's Law had a historic 2-year doubling time that has since slowed. While it originally referred to just transistor count increases, it has changed to commonly refer to just the performance increase. Some have predicted its stagnation as early as the mid-point of this decade (including Gordon Moore himself), but that is contested. More exotic paths forward such as non-silicon materials and 3D stacking have yet to be explored at scale, but research continues.

The microprocessor engineer Jim Keller stated in February 2020 that he doesn't think Moore's law is dead, that current transistors which are sized 1000x1000x1000 atoms can be reduced to 10x10x10 atoms before quantum effects (which occur at 2-10 atoms) stop any further shrinking, an effective 1,000,000x size reduction. Keller expects 10-20 more years of shrinking, and that performance increases will come from other areas of chip design as well. Finally, Keller says that the transistor count increase has slowed down more recently to a 'shrink factor' of 0.6 rather than the traditional 0.5, every 2 years. If that trend holds it will result in a 12.8x increase in performance in 10 years.

But Hardware improvements for AI need not just come just from Moore's law. Other sources of improvement such as Neuromorphic chips designed especially for running neural nets or specialised giant chips could create greater performance for AI.

By the end of the decade I estimate we should see between **8**-**13x** improvement in hardware performance.

Conclusions and Comparisons

If we put my estimates for algorithmic improvements, increased budgets and hardware improvements together we see what equivalent compute multiplier we might expect a frontier 2030 system to have compared to a frontier 2020 system.

Estimations for 2030:

Algorithmic Improvements: 100-1000x

Budget Increases: 1000-10,000x

Hardware Improvements: 8-13x

That results in an *800,000 - 130,000,000x* multiplier in equivalent compute.

Between **EIGHT HUNDRED THOUSAND** and **ONE HUNDRED and THIRTY MILLION**.

To put those compute equivalent multipliers into perspective in terms of what capability they represent there is only one architecture that seems worth extrapolating them out on: Transformers, specifically GPT-3.

Firstly lets relate them to Gwern's estimates for human vs GPT-3 level perplexity from his blogpost *On GPT-3*. Remember that perplexity is a measurement of how well a probability distribution or probability model predicts a sample. This is a useful comparison to make because it has been speculated both that human level prediction on text would represent

human level NLP, and that NLP would be an AI complete problem requiring human equivalent general faculties.

Gwern states that his estimate is very rough and relies on un-sauced claims from OpenAI about human level perplexity on benchmarks, and that the absolute prediction performance of GPT-3 is at a "*best guess*", double that of a human. With some “*irresponsible*” extrapolations of GPT-3 performance curves he finds that a 2,200,000× increase in compute would bring GPT-3 down to human perplexity. Interestingly that’s not far above the lower bound in the 800,000 – 130,000,000x equivalent compute estimate range.

It’s worth stressing, 2030 AI systems could have human level prediction capabilities if scaling continues.

Edit: Removed section extrapolating GPT-3 aggregate performance across benchmarks.

Ultimately the point of these extrapolations isn’t necessarily the specific figures or dates but the clear general trend: not only are *much* more powerful AI systems coming, they are coming *soon*.

AGI safety from first principles: Control

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

It's important to note that my previous arguments by themselves do not imply that AGIs will end up in control of the world instead of us. As an analogy, scientific knowledge allows us to be much more capable than stone-age humans. Yet if dropped back in that time with just our current knowledge, I very much doubt that one modern human could take over the stone-age world. Rather, this last step of the argument relies on additional predictions about the dynamics of the transition from humans being the smartest agents on Earth to AGIs taking over that role. These will depend on technological, economic and political factors, as I'll discuss in this section. One recurring theme will be the importance of our expectation that AGIs will be deployed as software that can be run on many different computers, rather than being tied to a specific piece of hardware as humans are.[1]

I'll start off by discussing two very high-level arguments. The first is that being more generally intelligent allows you to acquire more power, via large-scale coordination and development of novel technological capabilities. Both of these contributed to the human species taking control of the world; and they both contributed to other big shifts in the distribution of power (such as the industrial revolution). If the set of all humans and aligned AGIs is much less capable in these two ways than the set of all misaligned AGIs, then we should expect the latter to develop more novel technologies, and use them to amass more resources, unless strong constraints are placed on them, or they're unable to coordinate well (I'll discuss both possibilities shortly.)

On the other hand, though, it's also very hard to take over the world. In particular, if people in power see their positions being eroded, it's generally a safe bet that they'll take action to prevent that. Further, it's always much easier to understand and reason about a problem when it's more concrete and tangible; our track record at predicting large-scale future developments is pretty bad. And so even if the high-level arguments laid out above seem difficult to rebut, there may well be some solutions we missed which people will spot when their incentives to do so, and the range of approaches available to them, are laid out more clearly.

How can we move beyond these high-level arguments? In the rest of this section I'll lay out two types of disaster scenarios, and then four factors which will affect our ability to remain in control if we develop AGIs that are not fully aligned:

1. Speed of AI development
2. Transparency of AI systems
3. Constrained deployment strategies
4. Human political and economic coordination

Disaster scenarios

There have been a number of attempts to describe the catastrophic outcomes that might arise from misaligned superintelligences, although it has proven difficult to

characterise them in detail. Broadly speaking, the most compelling scenarios fall into two categories. Christiano describes AGIs gaining influence within our current economic and political systems by taking or being given control of companies and institutions. Eventually “we reach the point where we could not recover from a correlated automation failure” - after which those AGIs are no longer incentivised to follow human laws. Hanson also lays out a scenario in which virtual minds come to dominate the economy (although he is less worried about misalignment, partly because he focuses on emulated human minds). In both scenarios, biological humans lose influence because they are less competitive at strategically important tasks, but no single AGI is able to seize control of the world. To some extent these scenarios are analogous to our current situation, in which large corporations and institutions are able to amass power even when most humans disapprove of their goals. However, since these organisations are staffed by humans, there are still pressures on them to be aligned with human values which won’t apply to groups of AGIs.

By contrast, Yudkowsky and Bostrom describe scenarios where a single AGI gains power primarily through technological breakthroughs, in a way that’s largely separate from the wider economy. The key assumption which distinguishes this category of scenarios from the previous category is that a single AGI will be able to gain enough power via such breakthroughs that they can seize control of the world. Previous descriptions of this type of scenario have featured superhuman nanotechnology, biotechnology, and hacking; however, detailed characterisations are difficult because the relevant technologies don’t yet exist. Yet it seems very likely that there exist some future technologies which would provide a decisive strategic advantage if possessed only by a single actor, and so the key factor influencing the plausibility of these scenarios is whether AI development will be rapid enough to allow such concentration of power, as I discuss below.

In either case, humans and aligned AIs end up with much less power than misaligned AIs, which could then appropriate our resources towards their own goals. An even worse scenario is if misaligned AGIs act in ways which are deliberately hostile to human values - for example, by making threats to force concessions from us. How can we avoid these scenarios? It’s tempting to aim directly towards the final goal of being able to align arbitrarily intelligent AIs, but I think that the most realistic time horizon to plan towards is the point when AIs are much better than humans at doing safety research. So our goal should be to ensure that those AIs are aligned, and that their safety research will be used to build their successors. Which category of disaster is most likely to prevent that depends not only on the intelligence, agency and goals of the AIs we end up developing, but also on the four factors listed above, which I’ll explore in more detail now.

Speed of AI development

If AI development proceeds very quickly, then our ability to react appropriately will be much lower. In particular, we should be interested in how long it will take for AGIs to proceed from human-level intelligence to superintelligence, which we’ll call the *takeoff period*. The history of systems like AlphaStar, AlphaGo and OpenAI Five provides some evidence that this takeoff period will be short: after a long development period, each of them was able to improve rapidly from top amateur level to superhuman performance. A similar phenomenon occurred during human evolution, where it only took us a few million years to become much more intelligent than chimpanzees. In our case one of the key factors was scaling up our brain hardware -

which, as I have already discussed, will be much easier for AGIs than it was for humans.

While the question of what returns we will get from scaling up hardware and training time is an important one, in the long term the most important question is what returns we should expect from scaling up the intelligence of scientific researchers - because eventually AGIs themselves will be doing the vast majority of research in AI and related fields (in a process I've been calling *recursive improvement*). In particular, within the range of intelligence we're interested in, will a given increase δ in the intelligence of an AGI increase the intelligence of the best successor that AGI can develop by more than or less than δ? If more, then recursive improvement will eventually speed up the rate of progress in AI research dramatically. In favour of this hypothesis, Yudkowsky argues:

> The history of hominid evolution to date shows that it has not required exponentially greater amounts of evolutionary optimization to produce substantial real-world gains in cognitive performance - it did not require ten times the evolutionary interval to go from Homo erectus to Homo sapiens as from Australopithecus to Homo erectus. All compound interest returned on discoveries such as the invention of agriculture, or the invention of science, or the invention of computers, has occurred without any ability of humans to reinvest technological dividends to increase their brain sizes, speed up their neurons, or improve the low-level algorithms used by their neural circuitry. Since an AI can reinvest the fruits of its intelligence in larger brains, faster processing speeds, and improved low-level algorithms, we should expect an AI's growth curves to be sharply above human growth curves.

I consider this a strong argument that the pace of progress will eventually become much faster than it currently is. I'm much less confident about when the speedup will occur - for example, the positive feedback loop outlined above might not make a big difference until AGIs are already superintelligent, so that the takeoff period (as defined above) is still quite slow. There has been particular pushback against the more extreme fast takeoff scenarios, which postulate a discontinuous jump in AI capabilities before AI has had transformative impacts. Some of the key arguments:

1. The development of AGI will be a competitive endeavour in which many researchers will aim to build general cognitive capabilities into their AIs, and will gradually improve at doing so. This makes it unlikely that there will be low-hanging fruit which, when picked, allow large jumps in capabilities. (Arguably, cultural evolution was this sort of low-hanging fruit during human evolution, which would explain why it facilitated such rapid progress.)
2. Compute availability, which on some views is the key driver of progress in AI, increases fairly continuously.
3. Historically, continuous technological progress has been much more common than discontinuous progress. For example, progress on chess-playing AIs was steady and predictable over many decades.

Note that these three arguments are all consistent with AI development progressing continuously but at an increasing pace, as AI systems contribute to it an increasing amount.

Transparency of AI systems

A transparent AI system is one whose thoughts and behaviour we can understand and predict; we could be more confident that we can maintain control over an AGI if it were transparent. If we could tell when a system is planning treacherous behaviour, then we could shut it down before it gets the opportunity to carry out that plan. Note that such information would also be valuable for increasing human coordination towards dealing with AGIs; and of course for training, as I discussed briefly in previous sections.

Hubinger lists three broad approaches to making AIs more transparent. One is by creating interpretability tools which allow us to analyse the internal functioning of an existing system. While our ability to interpret human and animal brains is not currently very robust, this is partly because research has been held back by the difficulty of making high-resolution measurements. By contrast, in neural networks we can read each weight and each activation directly, as well as individually changing them to see what happens. On the other hand, if our most advanced systems change rapidly, then previous transparency research may quickly become obsolete. In this respect, neuroscientists - who can study one brain architecture for decades - have it easier.

A second approach is to create training incentives towards transparency. For example, we might reward an agent for explaining its thought processes, or for behaving in predictable ways. Interestingly, some hypotheses imply that this occurred during human evolution, which suggests that multi-agent interactions might be a useful way to create such incentives (if we can find a way to prevent incentives towards deception from also arising).

A third approach is to design algorithms and architectures that are inherently more interpretable. For example, a model-based planner like AlphaGo explores many possible branches of the game tree to decide which move to take. By examining which moves it explores, we can understand what it's planning before it chooses a move. However, in doing so we rely on the fact that AlphaGo uses an exact model of Go. More general agents in larger environments will need to plan using compressed representations of those environments, which will by default be much less interpretable. It also remains to be seen whether transparency-friendly architectures and algorithms can be competitive with the performance of more opaque alternatives, but I strongly suspect not.

Despite the difficulties inherent in each of these approaches, one advantage we do have in transparency analysis is access to different versions of an AI over time. This mechanism of cross-examination in Debate takes advantage of this. Or as a more pragmatic example, if AI systems which are slightly less intelligent than humans keep trying to deceive their supervisors, that's pretty clear evidence that the more intelligent ones will do so as well. However, this approach is limited because it doesn't allow us to identify unsafe plans until they affect behaviour. If the realisation that treachery is an option is always accompanied by the realisation that treachery won't work yet, we might not observe behavioural warning signs until an AI arises which expects its treachery to succeed.

Constrained deployment strategies

If we consider my earlier analogy of a modern human dropped in the stone age, one key factor that would prevent them from taking over the world is that they would be "deployed" in a very constrained way. They could only be in one place at a time; they

couldn’t travel or even send messages very rapidly; they would not be very robust to accidents; and there would be little existing infrastructure for them to leverage. By contrast, it takes much more compute to train deep learning systems than to run them - once an AGI has been trained, it will likely be relatively cheap to deploy many copies of it. A misaligned superintelligence with internet access will be able to create thousands of duplicates of itself, which we will have no control over, by buying (or hacking) the necessary hardware. At this point, our intuitions about the capabilities of a “single AGI” become outdated, and the “second species” terminology becomes more appropriate.

We can imagine trying to avoid this scenario by deploying AGIs in more constrained ways - for example by running them on secure hardware and only allowing them to take certain pre-approved actions (such as providing answers to questions). This seems significantly safer. However, it also seems less likely in a competitive marketplace - judging by today’s trends, a more plausible outcome is for almost everyone to have access to an AGI personal assistant via their phone. This brings us to the fourth factor:

Human political and economic coordination

By default, we shouldn’t rely on a high level of coordination to prevent AGI safety problems. We haven’t yet been able to coordinate adequately to prevent global warming, which is a well-documented, gradually-worsening problem. In the case of AGI deployment, the extrapolation from current behaviour to future danger is much harder to model clearly. Meanwhile, in the absence of technical solutions to safety problems, there will be strong short-term economic incentives to ignore the lack of safety guarantees about speculative future events.

However, this is very dependent on the three previous points. It will be much easier to build a consensus on how to deal with superintelligence if AI systems approach then surpass human-level performance over a timeframe of decades, rather than weeks or months. This is particularly true if less-capable systems display misbehaviour which would clearly be catastrophic if performed by more capable agents. Meanwhile, different actors who might be at the forefront of AGI development - governments, companies, nonprofits - will vary in their responsiveness to safety concerns, cooperativeness, and ability to implement constrained deployment strategies. And the more of them are involved, the harder coordination between them will be.

---

1. For an exploration of the possible consequences of software-based intelligence (as distinct from the consequences of increased intelligence) see Hanson’s *Age of Em*. ↩

"Scaling Laws for Autoregressive Generative Modeling", Henighan et al 2020 {OA}

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://arxiv.org/abs/2010.14701

The Darwin Game - Round 1

Technical Difficulties

I messed up the game engine. I gave bots their own move in the `previous` parameter instead of their opponent's move.

This is the primary factor behind the CloneBots diverging from each other. The bug caused CloneBot pairings to score 200-300 instead of 250-250.

I'm going to restart the tournament.

Sorry.

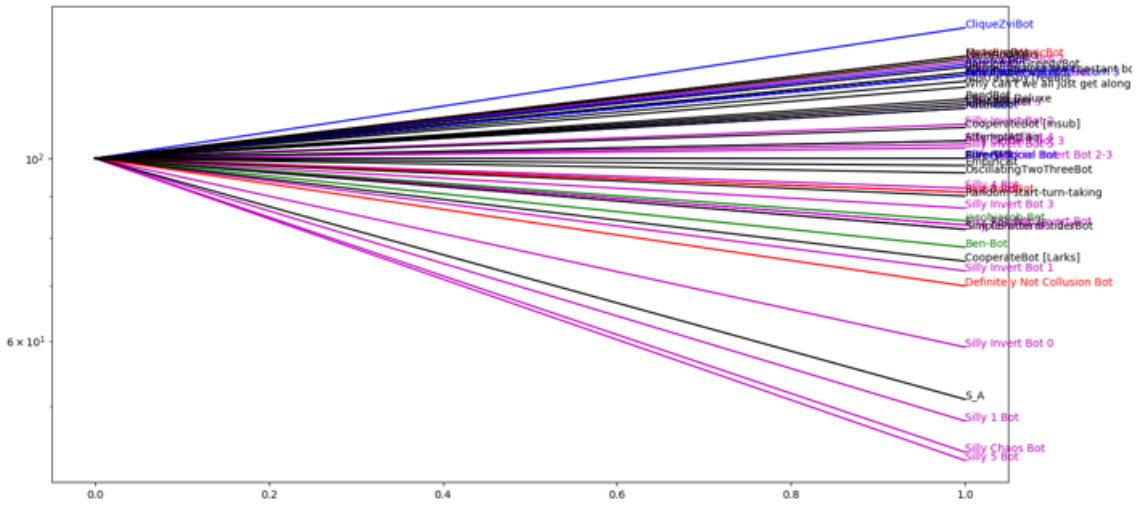
The original tournament, created by a blind idiot me, was exciting. It resembled real world population dynamics. I wrote an 8-part series with 11,277 words and 24 graphs.

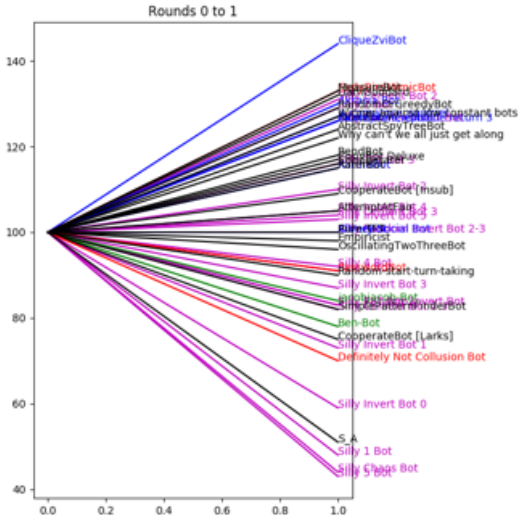
Alas, the actual contest is between intelligent agents.

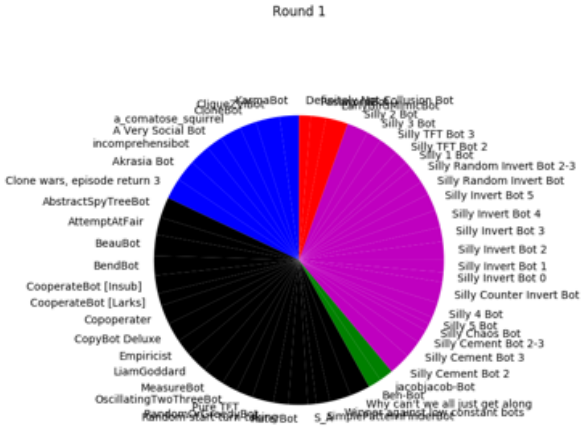
This will be fast.

Edit: Everything below this line is in error too. Click here for the real, real game.

Round 1







Today's Obituary

Bot	Team	Summary	Round
Silly 0 Bot	NPCs	Always returns `0`	1

The next installment of this series will be posted on October 30, 2020 at 5 pm Pacific Time.

Edit 2020-10-29: I'm sick. Next installment may be delayed.

Edit 2020-11-05: Since this is taking so long I feel I owe you guys an explanation. I have not forgotten about the tournament. Over the last couple of weeks, several correlated problems in meatspace triggered a cascading failure in my real life affairs. All the problems are temporary but I am temporarily unable to write bug-free software. (I actually did write a post for October 30 but it had multiple errors in it.) I would be surprised if it took more than a month to resolve everything but can make no guarantees. I plan to continue the tournament as soon as I can write bug-free software. I apologize for the delays.

Edit: 2020-11-10: I found some more bugs in the code and will have to restart things again.

What's holding back outsourcing to cloud labs?

I would have expected Emerald Cloud Lab or similar competitors to go a lot and be successful over the last five years. As far as I know, like Emerald Cloud Lab only had modest growth and there aren't competitors who grew strongly. Outsourcing to cloud labs seems like it allows the laberatory to have benefits of scale and virtualization that drives down costs and is easier to use then working in a wet lab. Is there something holding back this trend that I'm not seeing? Alternatively, what's going on?