

# Best of LessWrong: August 2020

1. [microCOVID.org: A tool to estimate COVID risk from common activities](#)
2. [Why haven't we celebrated any major achievements lately?](#)
3. [Notes on "The Anthropology of Childhood"](#)
4. [Inner Alignment: Explain like I'm 12 Edition](#)
5. [Radical Probabilism](#)
6. [The Fusion Power Generator Scenario](#)
7. [The Bayesian Tyrant](#)
8. [Tools for keeping focused](#)
9. [Matt Botvinick on the spontaneous emergence of learning algorithms](#)
10. [Alignment By Default](#)
11. [Forecasting Thread: AI Timelines](#)
12. [My Understanding of Paul Christiano's Iterated Amplification AI Safety Research Agenda](#)
13. [When can Fiction Change the World?](#)
14. [Unifying the Simulacra Definitions](#)
15. [Measuring hardware overhang](#)
16. [Introduction To The Infra-Bayesianism Sequence](#)
17. [How to teach things well](#)
18. [Updates and additions to "Embedded Agency"](#)
19. [Search versus design](#)
20. [Sunday, 20/8/30, 12pm PDT - Tagging Celebration: Habryka/Crawford + Party](#)
21. [nostalgebraist: Recursive Goodhart's Law](#)
22. [Infinite Data/Compute Arguments in Alignment](#)
23. [Does crime explain the exceptional US incarceration rate?](#)
24. [Alex Irpan: "My AI Timelines Have Sped Up"](#)
25. [A sketch of 'Simulacra Levels and their Interactions'](#)
26. [Model splintering: moving from one imperfect model to another](#)
27. [Three mental images from thinking about AGI debate & corrigibility](#)
28. [Generalized Efficient Markets in Political Power](#)
29. [Mesa-Search vs Mesa-Control](#)
30. [Preface to the sequence on economic growth](#)
31. [Swiss Political System: More than You ever Wanted to Know \(III.\)](#)
32. [Tagging Progress at 100%! \(Party & Celebratory Talk w/ Jason Crawford, Habryka on Sun, Aug 30th, 12pm PDT\)](#)
33. [is gpt-3 few-shot ready for real applications?](#)
34. [Epistemic Comparison: First Principles Land vs. Mimesis Land](#)
35. [10/50/90% chance of GPT-N Transformative AI?](#)
36. ["Good judgement" and its components](#)
37. [\[AN #112\]: Engineering a Safer World](#)
38. [Split-a-Dollar Game](#)
39. [Is Wirecutter still good?](#)
40. [Multitudinous outside views](#)
41. [Property as Coordination Minimization](#)
42. [Strong implication of preference uncertainty](#)
43. [Are We Right about How Effective Mockery Is?](#)
44. [Agentic Language Model Memes](#)
45. ["On Bullshit" and "On Truth," by Harry Frankfurt](#)
46. [You Need More Money](#)
47. [Forecasting Newsletter: July 2020.](#)
48. [Interpretability in ML: A Broad Overview](#)
49. [Highlights from the Blackmail Debate \(Robin Hanson vs Zvi Mowshowitz\)](#)
50. [Forecasting AI Progress: A Research Agenda](#)

# Best of LessWrong: August 2020

1. [microCOVID.org: A tool to estimate COVID risk from common activities](#)
2. [Why haven't we celebrated any major achievements lately?](#)
3. [Notes on "The Anthropology of Childhood"](#)
4. [Inner Alignment: Explain like I'm 12 Edition](#)
5. [Radical Probabilism](#)
6. [The Fusion Power Generator Scenario](#)
7. [The Bayesian Tyrant](#)
8. [Tools for keeping focused](#)
9. [Matt Botvinick on the spontaneous emergence of learning algorithms](#)
10. [Alignment By Default](#)
11. [Forecasting Thread: AI Timelines](#)
12. [My Understanding of Paul Christiano's Iterated Amplification AI Safety Research Agenda](#)
13. [When can Fiction Change the World?](#)
14. [Unifying the Simulacra Definitions](#)
15. [Measuring hardware overhang](#)
16. [Introduction To The Infra-Bayesianism Sequence](#)
17. [How to teach things well](#)
18. [Updates and additions to "Embedded Agency"](#)
19. [Search versus design](#)
20. [Sunday, 20/8/30, 12pm PDT - Tagging Celebration: Habryka/Crawford + Party](#)
21. [nostalgebraist: Recursive Goodhart's Law](#)
22. [Infinite Data/Compute Arguments in Alignment](#)
23. [Does crime explain the exceptional US incarceration rate?](#)
24. [Alex Irpan: "My AI Timelines Have Sped Up"](#)
25. [A sketch of 'Simulacra Levels and their Interactions'](#)
26. [Model splintering: moving from one imperfect model to another](#)
27. [Three mental images from thinking about AGI debate & corrigibility](#)
28. [Generalized Efficient Markets in Political Power](#)
29. [Mesa-Search vs Mesa-Control](#)
30. [Preface to the sequence on economic growth](#)
31. [Swiss Political System: More than You ever Wanted to Know \(III.\)](#)
32. [Tagging Progress at 100%! \(Party & Celebratory Talk w/ Jason Crawford, Habryka on Sun, Aug 30th, 12pm PDT\)](#)
33. [is gpt-3 few-shot ready for real applications?](#)
34. [Epistemic Comparison: First Principles Land vs. Mimesis Land](#)
35. [10/50/90% chance of GPT-N Transformative AI?](#)
36. ["Good judgement" and its components](#)
37. [\[AN #112\]: Engineering a Safer World](#)
38. [Split-a-Dollar Game](#)
39. [Is Wirecutter still good?](#)
40. [Multitudinous outside views](#)
41. [Property as Coordination Minimization](#)
42. [Strong implication of preference uncertainty](#)
43. [Are We Right about How Effective Mockery Is?](#)
44. [Agentic Language Model Memes](#)
45. ["On Bullshit" and "On Truth," by Harry Frankfurt](#)
46. [You Need More Money](#)
47. [Forecasting Newsletter: July 2020.](#)

48. [Interpretability in ML: A Broad Overview](#)
49. [Highlights from the Blackmail Debate \(Robin Hanson vs Zvi Mowshowitz\)](#)
50. [Forecasting AI Progress: A Research Agenda](#)

# microCOVID.org: A tool to estimate COVID risk from common activities

This is a linkpost for <https://microcovid.org/>

This is a linkpost for a model and web tool (that I and several friends created) to quantitatively estimate the COVID risk to you from your ordinary daily activities:

This website contains three outputs of our work:

1. a [web calculator](#) that you can use to calculate your COVID risk (in units of microCOVIDs, a 1-in-a-million chance of getting COVID).
2. a [white paper](#) that explains our estimation method. EAs might be particularly interested in the footnotes throughout, and the detailed [research sources](#) section.
3. a [spreadsheet](#) to compute your COVID risk in more detail and to track your risk over time. EAs might find this more customizable and powerful than the web calculator.

If you have different beliefs than us and would like to use a version of the model that reflects your beliefs rather than ours, you can make modifications to your copy of the spreadsheet, or fork the repository and make a personal copy of the web calculator. We also hope you will [submit suggestions](#), either by emailing us or by making issues or pull requests directly on Github.

Our group house has been using this model as the basis of a shared agreement/protocol, based on a budget of 3,000 microCOVIDs per year to spend outside the house (about 58 per week). We know of another group house that (last we heard) was operating on \*4\* microCOVIDs per week!

We hope this helps you personally live a better pandemic life with more safety *and* more flexibility.

(also linkposted to the [EA Forum](#))

# Why haven't we celebrated any major achievements lately?

This is a linkpost for <https://rootsofprogress.org/celebrations-of-progress>

In reading stories of progress, one thing that has struck me was the wild, enthusiastic celebrations that accompanied some of them in the past. Read some of these stories; somehow it's hard for me to imagine similar jubilation happening today:

## The US transcontinental railroad, 1869

The transcontinental railroad was the first to link the US east and west. Prior to the railroad, to travel from coast to coast could take six months, whether by land or sea, and the journey was hard and perilous. California was like a foreign colony, separated from the life and industry of the East. The railroad changed that completely, taking a six-month journey down to a matter of days.

Here's how the western cities reacted, from Stephen Ambrose's book [\*Nothing Like It in the World\*](#):

At 5 A.M. on Saturday, a Central Pacific train pulled into Sacramento carrying celebrants from Nevada, including firemen and a brass band. They got the festivities going by starting their parade. A brass cannon, the very one that had saluted the first shovelful of earth Leland Stanford had turned over for the beginning of the CP's construction six years earlier, boomed once again.

The parade was mammoth. At its height, about 11 A.M. in Sacramento, the time the organizers had been told the joining of the rails would take place, twenty-three of the CP's locomotives, led by its first, the Governor Stanford, let loose a shriek of whistles that lasted for fifteen minutes.

In San Francisco, the parade was the biggest held to date. At 11 A.M., a fifteen-inch Parrott rifled cannon at Fort Point, guarding the south shore of the Golden Gate, fired a salute. One hundred guns followed. Then fire bells, church bells, clock towers, machine shops, streamers, foundries, the U.S. Mint let go at full blast. The din lasted for an hour.

In both cities, the celebration went on through Saturday, Sunday, and Monday.

## The Brooklyn Bridge, 1883

The Brooklyn Bridge did not connect a distance nearly as great as the transcontinental railroad, but it too was met with grand celebrations. An excerpt from David McCullough's [\*The Great Bridge\*](#):

When the Erie Canal was opened in the autumn of 1825, there were four former Presidents of the United States present in New York City for the occasion—John Adams, Thomas Jefferson, James Madison, and James Monroe—as well as John Quincy Adams, then occupying the White House, and General Andrew Jackson, who would take his place. When the Brooklyn Bridge was opened on May 24, 1883, the main attraction was Chester A. Arthur. ...

Seth Low made the official greeting for the City of Brooklyn, the Marines presented arms, a signal flag was dropped nearby and instantly there was a crash of a gun from the

Tennessee. Then the whole fleet commenced firing. Steam whistles on every tug, steamboat, ferry, every factory along the river, began to scream. More cannon boomed. Bells rang, people were cheering wildly on every side. The band played "Hail to the Chief" maybe six or seven more times, and as the New York Sun reported, "the climax of fourteen years' suspense seemed to have been reached, since the President of the United States of America had walked dry shod to Brooklyn from New York."

Not only did they celebrate, they analyzed and philosophized:

What was it all about? What was everyone celebrating? The speakers of the day had a number of ideas. The bridge was a "wonder of Science," an "astounding exhibition of the power of man to change the face of nature." It was a monument to "enterprise, skill, faith, endurance." It was also a monument to "public spirit," "the moral qualities of the human soul," and a great, everlasting symbol of "Peace." The words used most often were "Science," "Commerce," and "Courage," and some of the ideas expressed had the familiar ring of a Fourth of July oration. ...

... every speaker that afternoon seemed to be saying that the opening of the bridge was a national event, that it was a triumph of human effort, and that it somehow marked a turning point. It was the beginning of something new, and although none of them appeared very sure what was going to be, they were confident it would be an improvement over the past and present.

The celebrations culminated with an enormous fireworks show:

In all, fourteen tons of fireworks—more than ten thousand pieces—were set off from the bridge. It lasted a solid hour. There was not a moment's letup. One meteoric burst followed another. ...

... finally, at nine, as the display on the bridge ended with one incredible barrage—five hundred rockets fired all at once—every whistle and horn on the river joined in. The rockets "broke into millions of stars and a shower of golden rain which descended upon the bridge and the river." Bells were rung, gongs were beaten, men and women yelled themselves hoarse, musicians blew themselves red in the face.

Comparing this to another accomplishment we'll return to below, McCullough writes:

In another time and in what would seem another world, on a day when two young men were walking on the moon, a very old woman on Long Island would tell reporters that the public excitement over the feat was not so much compared to what she had seen "on the day they opened the Brooklyn Bridge."

## Electric lighting, 1879

The electric light bulb was perhaps not met with parades or fireworks, but it did attract visitors from far and wide just to see the marvel. From Robert Gordon's *The Rise and Fall of American Growth*:

Few, if any inventions, have been more enthusiastically welcomed than electric light. Throughout the winter of 1879–1880, thousands traveled to Menlo Park to see the "light of the future," including farmers whose houses would never be electrified in their lifetimes. Travelers on the nearby Pennsylvania Railroad could see the brilliant lights glowing in the Edison offices. The news was announced to the world on December 21, 1879, with a full-page story in the New York Herald, opened by this dramatic and long-winded headline: EDISON'S LIGHT—THE GREAT INVENTOR'S TRIUMPH IN ELECTRIC ILLUMINATION—A SCRAP OF PAPER—IT MAKES A LIGHT, WITHOUT GAS OR FLAME, CHEAPER THAN OIL—SUCCESS IN A COTTON THREAD. On New Year's Eve of 1879, 3,000 people converged by train, carriage, and farm wagon on the Edison laboratory to witness

the brilliant display, a planned laboratory open house of dazzling modernity to launch the new decade.

## The polio vaccine, 1955

Rails, bridges and lights were celebrated in part because they greatly relieved the burdens of distance and darkness. Another burden was lifted in 1955 when the polio vaccine was announced.

Polio *terrified* the nation, much more so than diseases such as tuberculosis that were actually much bigger killers, for a few reasons. It struck in unpredictable, dramatic epidemics. The epidemics were relatively new starting in the late 1800s; it was not a disease that had been widespread throughout history, such as [smallpox](#). It left many victims paralyzed rather than killing them, so its results were visible in the form of crutches, braces, and wheelchairs. It targeted children, striking fear into the hearts of parents. And it could not be fought with the new weapons of cleanliness and sanitation, which were successful against so many other diseases. This added guilt to the fear, as parents of polio victims obsessed over what they had done wrong in failing to protect their children.

So it's understandable that the entire nation was eager to hear the news of a vaccine, and went wild when it was achieved. From *Breakthrough: The Saga of Jonas Salk*, by Richard Carter:

On April 12, 1955, the world learned that a vaccine developed by Jonas Edward Salk, M.D., could be relied upon to prevent paralytic poliomyelitis. This news consummated the most extraordinary undertaking in the history of science, a huge research project led by a Wall Street lawyer and financed by the American people through hundreds of millions of small donations. More than a scientific achievement, the vaccine was a folk victory, an occasion for pride and jubilation. A contagion of love swept the world. People observed moments of silence, rang bells, honked horns, blew factory whistles, fired salutes, kept their traffic lights red in brief periods of tribute, took the rest of the day off, closed their schools or convoked fervid assemblies therein, drank toasts, hugged children, attended church, smiled at strangers, forgave enemies....

The ardent people named schools, streets, hospitals, and newborn infants after him. They sent him checks, cash, money orders, stamps, scrolls, certificates, pressed flowers, snapshots, candy, baked goods, religious medals, rabbits' feet and other talismans, and uncounted thousands of letters and telegrams, both individual and round-robin, describing their heartfelt gratitude and admiration. They offered him free automobiles, agricultural equipment, clothing, vacations, lucrative jobs in government and industry, and several hundred opportunities to get rich quick. Their legislatures and parliaments passed resolutions, and their heads of state issued proclamations. Their universities tendered honorary degrees. He was nominated for the Nobel prize, which he did not get, and a Congressional medal, which he got, and membership in the National Academy of Sciences, which turned him down. He was mentioned for several dozen lesser awards of national or local or purely promotional character, most of which he turned down.

Not all of this happened on April 12, 1955, but much of it did. Salk awakened that morning as a moderately prominent research professor on the faculty of the University of Pittsburgh School of Medicine. He ended the day as the most beloved medical scientist on earth.

David Oshinsky adds more details in [Polio: An American Story](#):

There had been celebrations like this for athletes, soldiers, politicians, aviators—but never for a scientist. Gifts and honors poured in from a grateful nation. Philadelphia awarded Salk its Poor Richard Medal for distinguished service to humanity. Mutual of

Omaha gave him its Criss Award, along with a \$10,000 check, for his contribution to public health. The University of Pittsburgh was swamped with thank-you notes and "donations" addressed to Dr. Salk. His lab was "knee-deep in mail," a staffer recalled. "Paper money [went] into one bin, checks into another, and metal coins into a third." (How much was collected, and who kept what, was never fully divulged.) Elementary schools sent giant posters—WE LOVE YOU DR. SALK—signed by the entire student body. Winnipeg, Canada, site of a major polio epidemic in 1953, sent a 208-foot telegram of congratulation adorned with each survivor's name. A town in the Texas panhandle bought him two heartfelt, if comically inappropriate, gifts: a plow and a fully equipped Oldsmobile 98. (Salk gave the plow to an orphanage and had the car sold so the town could buy more polio vaccine.) A new Cadillac arrived and was donated to charity. Colleges begged him to accept their honorary degrees. Newsweek lauded "A Quiet Young Man's Magnificent Victory," insisting that Salk's name was now "as secure a word in the medical dictionary as Jenner, Pasteur, Schick, and Lister."

Hollywood wasn't far behind. Three major studios—Warner Brothers, Columbia, and Twentieth Century-Fox—fought for the exclusive rights to Salk's life story. Rumors flew that Marlon Brando was angling for the lead—an odd choice, most agreed, but a sure sign of box office pizzazz. Salk wisely told them no. "I believe that such pictures are most appropriately made after the scientist is dead," he remarked, "and I'm willing to await my chances of such attention at that time."

Politicians embraced him. One senator introduced a bill to give the forty-year-old Salk a \$10,000 annual stipend for life. Another proposed the minting of a Salk dime, just like FDR's. (Both ideas went nowhere.) Governor George Leader of Pennsylvania gave him the state's highest honor—the Bronze Medal for Meritorious Service—before a cheering joint session of the legislature (which soon created an endowed chair for Salk at the University of Pittsburgh Medical School with a princely stipend of \$25,000 a year). On an even grander scale, the U.S. House and Senate began the bipartisan process of commissioning a Congressional Gold Medal, the nation's highest civilian award. Salk would become only the second medical researcher to receive one, joining Walter Reed of yellow fever fame. The two men were in good company. Previous honorees included Thomas Edison, Charles Lindbergh, General George C. Marshall, and Irving Berlin.

Hundreds wrote President Eisenhower to request a special White House ceremony for Salk. ... On April 22 Jonas and Donna Salk, their three young boys, and Basil O'Connor arrived at the White House to meet the president. ... The Rose Garden ceremony that day would not soon be forgotten. Few had ever seen Dwight Eisenhower struggle with his feelings in such a public way. "No bands played and no flags waved," wrote a reporter who had followed Ike for years. "But nothing could have been more impressive than this grandfather standing there and telling Dr. Salk in a voice trembling with emotion, 'I have no words to thank you. I am very, very happy.'"

... The banner headline in the Pittsburgh Press on April 12, 1955 had set the tone—POLIO IS CONQUERED. The stories that day spoke of mothers weeping, doctors cheering, politicians toasting God and Jonas Salk.

Steven Pinker, in [Enlightenment Now](#), after quoting some of the passage from Richard Carter above, adds: "The city of New York offered to honor Salk with a ticker-tape parade, which he politely declined." Speaking of which—

## Historic flights, 1920s and '30s

I looked up the history of ticker-tape parades in New York City. Wikipedia [has a list](#). These seem to have been most common from about 1926 to 1965, with multiple parades a year in that period (except when the US was fighting WW2, when there were none), compared with less than one a year on average in the years before or since.

What was celebrated? Mostly politicians, military heroes, visiting foreign leaders, and occasionally sports champions. (There was one parade for a musician, Van Cliburn, after he won the Moscow International Tchaikovsky Competition.)

However, the 1920s and '30s saw over a dozen parades celebrating aviation achievements, including Charles Lindburgh and Amelia Earhart:

- 1926, June 23 – Commander Richard Byrd and Floyd Bennett, flight over the North Pole
- 1927, June 13 – Charles Lindbergh, following solo transatlantic flight.
- 1927, July 18 – “Double” parade for Commander Richard Byrd and the crew of the America; and for Clarence Chamberlin and Charles A. Levine following separate transatlantic flights.
- 1927, November 11 – Ruth Elder and George W. Haldeman following flight from New York City to the Azores.
- 1928, April 25 – Hermann Köhl, Major James Fitzmaurice, and Baron von Hünefeld following first westward transatlantic flight
- 1928, July 6 – Amelia Earhart, Wilmer Stultz, and Louis E. Gordon
- 1930, September 4 – Captain Dieudonne Coste and Maurice Bellonte following flight from Paris to New York City.
- 1931, July 2 – Wiley Post and Harold Gatty following round-the-world flight.
- 1932, June 20 – Amelia Earhart Putnam following transatlantic flight.
- 1933, July 21 – Air Marshal Italo Balbo and crew for flight from Rome to Chicago in 25 Italian seaplanes.
- 1933, July 26 – Wiley Post following eight-day round-the-world flight.
- 1933, August 1 – Captain James A. Mollison and his wife following westward transatlantic flight, from Wales to Connecticut.
- 1938, July 15 – Howard Hughes, following three-day flight around the world.
- 1938, August 5 – Douglas “Wrong Way” Corrigan following flight from New York City to Ireland (he was scheduled to fly to California).

## Astronauts, 1962-71

During the early space program, there were also several NYC ticker-tape parades for astronauts—not just the Apollo 11 heroes, who went on a world tour after the Moon landing, but missions before and after as well:

- 1962, March 1 – John Glenn, following the Mercury-Atlas 6 mission.
- 1962, June 5 – Scott Carpenter, following the Mercury 7 mission.
- 1963, May 22 – Gordon Cooper, following the Mercury 9 mission.
- 1965, March 29 – Virgil “Gus” Grissom and John Young, following the Gemini 3 mission.
- 1969, January 10 – Frank Borman, James A. Lovell, and William A. Anders, following the Apollo 8 mission to the Moon.
- 1969, August 13 – Neil Armstrong, Buzz Aldrin, and Michael Collins, following Apollo 11 mission to the Moon.
- 1971, March 8 – Alan Shepard, Edgar Mitchell, and Stuart Roosa, following Apollo 14 mission to the Moon.
- 1971, August 24 – David Scott, James Irwin, and Alfred Worden, following Apollo 15 mission to the Moon.

And much later:

- 1998, November 16 – John Glenn and astronauts of Space Shuttle Discovery mission STS-95.



Apollo 11 parade [Wikimedia / NASA](#)

## Recent celebrations?

I'm having a hard time coming up with any major celebrations of scientific, technological, or industrial achievements since the Apollo Program.

When I alluded to this [on Twitter](#), some people suggested the long lines of consumers waiting to buy iPhones. I don't count that in the same category: it shows a desire for a product. I'm looking for outright celebration.

It's not that no one cares about progress anymore. Plenty of people still get excited by science news, new inventions, and breakthrough achievements—especially in space, which has a strong “coolness” factor. Noah Smith [polled his followers](#), and ~75% of respondents said they “celebrated or got very excited about” the Mars Pathfinder landing in 1996. More recently, many people in my circles were excited about the SpaceX Dragon launch a few months ago. But a minority of geeks excitedly watching live feeds from home doesn't compare, in my opinion, to the celebrations described above.

It's also not that we don't honor progress in any way. Formal institutions such as the Nobel prizes still do so on a regular basis. I'm talking more about ad-hoc displays of enthusiasm and admiration.

## Some hypotheses

Here are a few hypotheses for why there haven't been any major celebrations of progress in the last ~50 years:

- **There haven't been as many big accomplishments.** We haven't gone back to the Moon or cured cancer. We haven't solved traffic or auto accidents. This is the stagnation hypothesis.

But what about the progress we have made? What about computers and the Internet? What about sequencing the human genome or [producing insulin using genetic engineering](#)?

This leads to the second hypothesis:

- **The progress we have made hasn't been the kind that lends itself to big public celebrations.** Celebrations are generally for big, visible achievements that were completed at a defined point and that the public could easily understand. Computers and the Internet were not obviously about to change the world when they were invented, and they did so gradually, over decades. The human genome was big science news but too removed from immediate practical benefit to cause dancing in the streets.

Similar explanations seem to apply to achievements in the past. For instance, in contrast to the polio vaccine, I can't remember reading about any celebrations of Edward Jenner's smallpox vaccine. The concept of vaccines (and even inoculation, the technique that preceded vaccination) was too new and too controversial. It took time for everyone to believe and accept that the vaccine worked. A century and a half later, after the germ theory was established and there were many clear successes of fighting disease with science, the public was ready to celebrate the polio vaccine.

Take another example, [the Haber-Bosch process](#). This was certainly one to celebrate, but I don't recall any parades or fireworks for it. Again, it seems perhaps too technical and removed from what the general public could get excited about.

- **People celebrate things differently now,** maybe in less formal and public ways. As noted, the ticker-tape parades in NYC waned after the mid-1960s. In an era of telecommunications, maybe people don't have as much of a need to get together in large groups? Maybe 21st-century celebration takes the form of something getting ten million likes on Facebook?

I have a hard time buying this one. We still hold parades for sports championships, launch fireworks for the Olympics, and gather in large groups for New Year's Eve. I think there is still a psychological need for big, public celebrations.

- **We just don't appreciate progress as much as we used to.** I'm not sure we need this hypothesis, in that I think the first two explain all of the observations so far. But I believe it, because it matches a broader trend of waning enthusiasm and growing skepticism and even antagonism towards progress. As a thought experiment, can you imagine Presidential speeches and a brass band at the opening of a bridge today?

## What will happen for future achievements?

OK, you might say, bridges have become commonplace. What if it wasn't a bridge, but the first space elevator? Would that be met with celebration? Or opposition? Or a yawn?

Or take a less sci-fi example. How will we greet the COVID-19 vaccine, when it arrives hopefully in the next year or two? Will people "ring bells, honk horns, blow whistles, fire salutes, drink toasts, hug children, and forgive enemies"? Will they "name schools, streets, hospitals, and newborn infants" after the creator?

Or what if Elon Musk succeeds with a manned mission to Mars? When the first Martian astronauts return, will they go on world tour like Armstrong, Aldrin and Collins?

I don't know. Maybe! It will be interesting to see.

# Notes on "The Anthropology of Childhood"

Crossposted from [The Whole Sky](#).

I read David Lancy's "[The Anthropology of Childhood: Cherubs, Chattel, and Changelings](#)" and highlighted some passages. A lot of passages, it turns out.

[content note: discussion of abortion and infanticide, including infanticide of children with disabilities, in "Life and Death" section but not elsewhere]

I was a sociology major and understood anthropology to be basically "like sociology, but in Papua New Guinea." This is the first cultural anthropology book I've read, and that was pretty much right. I found it very accessible as a first dive into anthropology. The first chapter summarizes all his points without the examples, so you could try that if you want to get the gist without reading the whole book.

I enjoyed it and would recommend it to people interested in this topic. A few things that shifted for me:

- I feel less obliged to entertain my children and intervene in their conflicts. We don't live with a tribe of extended family, but my two children play with each other all day, which is how most people throughout time have spent their childhoods. Lancy isn't a child development expert, but I buy his argument that handling conflict (for example about the rules of a game) is a skill children need to learn, rather than having conflicts always mediated by adults.
- Even though it doesn't change anything concrete, I feel some relief that not having endless patience for toddlers seems to be normal. Except where families were very isolated, it's not normal in traditional societies for one or two adults to watch their own children all day every day. And childcare has traditionally looked mostly like "being sure they don't hurt themselves too badly."
- It surprised me that childcare by non-parents was so common. Some more modern views treat women's childcare work as basically free, but traditional cultures have valued women's labor enough that the society wants to free up their time from childcare. It was striking to me that the expectation that stay-at-home mothers will be responsible for all childcare was a relatively short historical blip. But of course, having childcare done by teenagers and grandmothers requires that those people's time be available, which usually isn't the reality we live in.
- I was surprised at how apparently universal it is for fathers to be uninvolved. I expect they're typically involved in providing food and other material resources, but that wasn't emphasized in this book.

I'm a little unclear on how valid Lancy's conclusions are or how much data they're based on. It seems like an anthropologist could squint at a society and see all kinds of things that someone with a different ideology wouldn't see.

Big caveat that what Lancy is describing is traditional, non-industrialized societies where children are expected to learn how to fit into the appropriate role in their village, not to develop as an individual or do anything different from what their parents and ancestors did. He stresses that traditional childrearing practices are very poor preparation for school. Given that I want my children to learn things I don't know,

to think analytically, etc, the way I approach learning is very different from how traditional societies approach it.

Lancy periodically complains about how much money Western families spend on fertility treatments, medical care for premature infants, etc. He argues that the same money could be used to provide adequate nutrition for many more children in the societies he's studied. I'm [sympathetic](#), but assuming that families would donate this money if they weren't spending it to have a baby is not realistic. I see cutting luxury spending as a much more feasible way that people might do some redistribution.

And now, my notes:

## Views of childhood

As in many areas of research, the children who have been studied by academics are mostly from WEIRD ("Western, educated, industrialized, rich, democratic") populations. Thus our understanding of good or normal childrearing practices is very different from how children have typically been raised. Lancy contrasts modern childrearing norms with those of traditional agrarian or forager societies.

Lancy contrasts neontocracy (where babies and children are most valued) with gerontocracy (where elders or ancestors are most valued). I can think of ways our society isn't very good for children, but I agree that compared with traditional societies, we spend a lot of attention and money on children. (Albeit sometimes by micromanaging them, while Lancy would rather have them figure out more for themselves as children have historically done.)

Even studying children is a strange thing to do in most societies. "Examples of children treated as lacking any sense, as being essentially uneducable, are legion in the ethnographic record." "Anthropologists interested in children are treated in a bemused fashion; after all, why bother to observe or talk to individuals who 'don't know anything'?" (Lancy 1996: 118; also Barley 1983/ 2000: 61)"

"Infants were widely seen as insensible. Almost like plants, their care could be rudimentary"

Traditional societies have two broad patterns toward young children: "One response is 'benign neglect'- everyone waits until the child can talk sensibly before acknowledging its existence. A second typical response is to aggressively humanize the child, including ruthless suppression of all 'sub-human' tendencies (e.g. bawling, crawling, thumb-sucking)."

Europeans were of the second view:

"Like wild men [or beasts], babies lacked the power to reason, speak, or stand and walk erect. [They were] nasty, brutish, and dirty, communicating in wordless cries, grunts, and screams, and were given to crawling on all fours before they could be made to walk like men ... Left to their own devices, they would remain selfish, animalistic, and savage. Parents believed they had to coerce their babies into growing up, and they expected protests and resistance. (Calvert 1992: 26, 34)"

"The Puritans were perhaps the first anxious parents, fearing they might fail and their children would turn out badly."

"We now take for granted the "need" to stimulate the infant through physical contact, motherese, and playing games like peek-a-boo to accelerate physical and intellectual development. Contrast these assumptions with the pre-modern objective of keeping babies quiescent so they'd make fewer demands on caretakers and not injure themselves (LeVine et al. 1994)."

"Much of what we think of as the routine duties (e.g. reading bedtime stories; cf. Lancy 1994) or expenses (e.g. orthodontics) of modern parents are completely unknown outside modern, mainstream societies."

"150 years ago, the idea of the useful child began to give way to our modern notion of the useless but also priceless child (Zelizer 1985). Children become innocent and fragile cherubs, needing protection from adult society, including the world of work. Their value to us is measured no longer in terms of an economic payoff or even genetic fitness but in terms of complementing our own values – as book lovers, ardent travelers, athletes, or devotees of a particular sect."

"Known as the "largest children's migration in history," so-called "orphan trains" carried about 200,000 children (Warren 2001: 4) from orphanages and foundling homes in eastern coastal cities to families in the Midwest (Kay 2003: iii) and West. The orphan trains continued until 1929 (Warren 2001: 20), which indicates how very recently our fundamental conception of children as chattel changed to viewing them as cherubs."

Anne of Green Gables is a story about this dynamic in Canada — the family was expecting to adopt a boy who could serve as an unpaid farmhand, but got a girl orphan by mistake.

## Who cares for children?

Surprisingly to me, in traditional societies it's usually not mothers.

In the early days of infancy, of course, breastfeeding necessitates keeping mother and baby convenient to each other. "Nearly all societies hold very strict views on the necessity for almost constant contact between a mother or other nurturing adult and the infant. Infants are fed on demand, carried constantly, and sleep with their mother. Young mothers are severely chastised for any lapse in infant care. However, once the infant begins to walk, it immediately joins a social network in which its mother plays a sharply diminished role – especially if she's pregnant – and its father may play no role at all."

On the saying "it takes a village to raise a child": "If one actually looks at real kids in real villages, either one sees infants and young children in a group of their peers, untended by an adult, or one sees a mother, or a father, or an older sister, or a grandmother tending the child. These helpful family members are referred to in anthropology as 'alloparents.' The rule governing their behavior would not necessarily be 'Everyone's eager to have a hand in caring for the child,' but, rather, 'Whoever can most easily be spared from more important tasks will take care of the child.' And the next rule we might derive from our observations might be, "The mother is often too busy to tend to the child." At the same time, babies are not simply passive recipients of care. They not only look cute, they beguile caretakers with their gaze, their smiling and their mimicry (Spelke and Kinzler 2007: 92). While alloparents may want to minimize their effort (Trivers 1974) in caring for the child, the very young have an

arsenal of tactics they can deploy to secure additional resources (Povinelli et al. 2005)."

"Weisner and Gallimore examined hundreds of ethnographies in the Human Relations Area Files (HRAF) archive and found that, in accounts of childcare, 40 percent of infants and 80 percent of toddlers are cared for primarily by someone other than their mother, most commonly older sisters (Weisner and Gallimore 1977)."

"Three-year-old children are able to join in a play group, and it is in such play groups that children are truly raised" (Eibl-Eibesfeldt 1989: 600)."

"Once the infant has been judged worthy of rearing, it will be displayed to a community eager to interact with it. In particular, its older sisters will be in the forefront of those wanting to share in the nurturing process. The circle of caretakers may gradually widen to include aunts, grandmothers, and, occasionally, the father. Even more distant kin can be expected to cast a watchful eye on the child when it is playing on the 'mother-ground' (Lancy 1996: 84). Indeed, the toddler must seek comfort from relatives as it may be abruptly weaned and forcibly rejected by its mother as she readies herself for the next child."

In a large polygynous household the author visited in Liberia, even after a few weeks he was unable to figure out which children belonged to which mothers: "I was stymied because the children, once they were no longer attached marsupial-like to their mother's body with a length of cloth, spent far more time in each other's company and in the company of other kin, particularly grandmothers and aunts in nearby houses, than with their mothers. And as far as the chief was concerned, I just had to assume that since these were his wives, the majority of the children in the vicinity must be his as well. Aside from dandling the occasional infant on his knee during the family's evening meal, I never saw him enjoy more than the most fleeting interaction with a child." Later, "I began to see their family arrangements and childcare customs as neither unusual nor exotic, rather as close to the norm for human societies, and, simultaneously, to see the customs of the middle-class Utah community I live in now as extraordinary."

Older sisters are often alloparents:

"Across the primate order, juvenile females show great interest in infants (Hrdy 1999: 157), and it is not hard to sustain an argument that their supervised interaction with younger siblings prepares them for the role of motherhood (Fairbanks 1990; Riesman 1992: 111). The weanling's need for mothering corresponds to the allomother's need to mother."

This seems to be true in other primates as well (though I do imagine researcher bias could influence whether carrying around a stick is 'doll play' depending on the gender of the young chimp.)

"Several studies have documented the gender bias in 'baby lust' (Hrdy 1999: 157). Females show far more interest in babies, images of babies, and even silhouettes of babies than do males. In fact, there's some evidence that young chimp females will cradle, groom, and carry around a "doll" (a stick or a dead animal) in the absence of a live infant (Kahlenberg and Wrangham 2010: 1067)."

"In Uganda in 2003, I observed and filmed numerous primate species and, after resting, eating, and play, "baby-trading" is the most common occupation. Often I observed what amounted to a "tug-of-war" between the nursing mother and her older

daughters for possession of the infant, which may lead to what Sarah Hrdy (1976) referred to as “aunting to death.” By contrast, mothers tend to discourage interest shown by juvenile males in their offspring (Strier 2003).<sup>12</sup>

“Aunting to death” sounds familiar to me. When Lily was born, we lived with Jeff’s family including his two sisters. They would literally race each other to the baby each morning when I came downstairs with Lily, as each aunt tried to arrive first for baby cuddles.

Boys are not seen as good caregivers:

“Dozens of studies have documented the heightened likelihood of sensation-seeking (Zuckerman 1984) or risk-taking by adolescent primate males in groups. Demographers have identified an “accident hump” in mortality curves for male primates, including humans, during puberty (Goldstein 2011).”

“I had a personal epiphany regarding the inadvisability of assigning boys as sibling caretakers in May 2007 as I stood on a busy street in front of the Registan in Samarkand. Two boys were pushing baby carriages in the street, just barely out of traffic. The street sloped downward and the lead carriage-pusher began a game of chicken, releasing his grip on the bar, then rushing after to grab it as the carriage rolled away on its own. This game was repeated with longer intervals between the release and retrieval.”

Children need less oversight in less dangerous environments:

“Tether-length is definitely a useful concept in observing human mother-toddler interaction (Broch 1990: 71-72). As Sorenson discovered in a Fore village, the infant’s “early pattern of exploratory activity included frequent returns to the mother. She served as the home base, the bastion of security but not as director or overseer of activities” (Sorenson 1976: 167). For the forest-dwelling Chewong, the tether is shorter. Toddlers are discouraged from wandering away from proximity to adults with “loud exclamations ... ‘it is hot,’ or ‘it is sharp,’ or ‘there are ... tigers, snakes, millipedes’” (Howell 1988: 163).”

Swaddling makes children easier to watch:

“A swaddled baby, like a little turtle in its shell, could be looked after by another, only slightly older child without too much fear of injury, since the practice of swaddling made ... child care virtually idiot proof. (Calvert 1992: 23-24)”

There is a chain of oversight:

“toddlers are managed by slightly older siblings, who are, in turn, guided by adolescents, while adults serve as rather distant “foremen” for the activity, concentrating, primarily, on their own more productive or profitable activity.”

The stereotype of grandmothers “spoiling” children is not unique to the West:

“[I]n the Mende view, grannies are notoriously lax with children. They are said to feed children upon demand and do not beat them or withhold meals from them for bad behavior or for failing to work ... Children raised like this are said to grow up lazy and dishonest ...”

In Rome, nurses were responsible for childcare in wealthy families:

"It was the nutrix [nurse] who ... took responsibility for ... early infant care: breast-feeding, powdering and swaddling, bathing and massaging, rocking and singing the child to sleep, weaning the child from milk to solid food ... The nutrix, in fact, was only one of a sequence of child-minding functionaries who influenced the early lives of children."

"Public attitudes in Europe reflect a view of the family that echoes the utopian ideals of the Israeli kibbutz from the mid-twentieth century. While the mother might be the primary caretaker during infancy, shortly afterward the child should be placed in a nursery with trained staff as she returns to her job. This policy is seen as beneficial to the mother's self-esteem, the economy, and the child itself (Corsaro 1996; Dahlberg 1992; Eibl-Eibesfeldt 1983: 181). Publicly supported pre-school or daycare in the US has been blocked by the politically powerful religious right, which insists on keeping wives tied full-time to the kitchen and nursery."

When childcare is a collective task, discipline is also collectivized:

"The mother must, however, accept the consequence that virtually anyone older than her child can scold or even discipline them (Whiting 1941). In societies like our own, where childcare is handled within the nuclear family and/or by professionals, the necessity for learning manners and kinship arcana is reduced. At the same time, we are often reluctant to concede to outsiders, even "professionals," the right to discipline our young."

## Why do mothers outsource childcare?

"In a majority of the world's diverse societies, women continue as workers throughout pregnancy and resume working shortly after the child is born. This work is physically demanding, so, for many, there is a peak period in their lives when they have the stamina and fat reserves to do their work and have babies. How many babies they successfully rear will depend heavily on their access to a supportive community of relatives who can help with household work, assist with childcare, and provide supplementary resources."

## How children are taught to relate to others

Contrasted with the emphasis on the mother-child bond in WEIRD society generally and especially in "[attachment parenting](#)", traditional cultures may emphasize finding other caregivers:

"The baby's cherub-like features aid the mother in her quest for helpers. Young mammals, generally, but especially humans, display a suite of physical features that seem to be universally attractive to others, and these features are retained longer in humans than in other mammalian species (Lancaster and Lancaster 1983: 35; Sternglanz et al. 1977). Also critical is the fact that human infants vocalize, make eye contact, and smile from very early on (Chevalier-Skolnikoff 1977) – unlike chimps, for example, whose mothers make more limited use of helpers. Mothers may not always rely on the inherent cuteness of their babies; they may take pains to showcase the baby – at least among close kin. The Kpelle mothers I observed didn't stop at frequently washing and cleaning their babies. They oiled the babies' bodies until they gleamed – an ablution carried out in public view with an appreciative audience. The Kaluli mothers studied by Bambi Schieffelin in Papua New Guinea not only hold their

infants facing toward others in the social group – a practice often noted in the ethnographic record – but treat the baby as a ventriloquist’s dummy in having him or her speak to those assembled (Schieffelin 1990: 71). The Beng advise young mothers: Make sure the baby looks beautiful! ... put herbal makeup on her face as attractively as possible ... we Beng have lots of designs for babies’ faces ... That way, the baby will be so irresistibly beautiful that someone will feel compelled to carry her around for a while that day. If you’re lucky, maybe that person will even offer to be your leng kuli. (Gottlieb 1995: 24) When [Guara] neighbors visit ... relatives – identified by kinship terms – are repeatedly indicated to the child. (Ruddle and Chesterfield 1977: 29) [Marquesan mothers] ... spent much time calling the baby’s name, directing him to look and wave at others ... directing three- to six-year-old siblings to play with him. (Martini and Kirkpatrick 1981: 199)”

“Samoan ...toddlers were fed facing others and prompted to notice and call out to people. (Ochs and Izquierdo 2009: 397) From the moment a [Warlpiri] child is born ... she will hear every day ... for the next few years; “Look, your granny,” “That’s your big sister, your cousin, your auntie.” In fact, they make up the bulk of verbal communication with babies and little children. (Musharbash 2011: 72)”

“There were numerous constraints put on young [Orissa India] mothers to prevent them from focusing too much attention on a new infant. Close, intimate mother-child bonds were viewed as potentially disruptive to the collective well-being of the extended family ... In such families, much early child-care was organized so as to subtly push the infant away from an exclusive dependence on its mother toward membership in the larger group. (Seymour 2001: 15)”

## How do parents learn to parent?

Partly through alloparenting as described above. Among other primates:

“While the benefits to the mother are obvious, allomothering daughters also clearly benefit by learning how to care for infants (Fairbanks 1990). A study of captive chimpanzees showed that females prevented from interacting with their mothers and younger siblings were themselves utterly incompetent as mothers (Davenport and Rogers 1970).”

I was surprised at how hard it was to feed a newborn - in my case I got help from the midwife, a lactation consultant, and the pediatrician, but traditionally advice would come from family and neighbors:

“Field and colleagues, working with Haitian immigrant mothers in Miami, find these mothers often have difficulty feeding their offspring, who are therefore hospitalized for dehydration and malnutrition at a high rate (Field et al. 1992: 183). I think it’s possible these young women immigrants lost the opportunity to learn how to care for infants from older women.”

Among the Fulani of West Africa:

“All women caring for their first babies will have had years of experience taking care of babies ... under the watchful and sometimes severe eyes of their mothers, aunts, cousins or older sisters. The other women ... will immediately notice, comment on, and perhaps strongly criticize any departure from customary behavior on the part of mothers. (Riesman 1992: 111)”

(Anthropologists traveling with their own children also get a lot of advice from locals.)

## Nutrition

I hadn't really thought about how much of life in traditional societies revolved around the essential, never-ending task of getting calories. There is often not enough to go around, and social differences can be observed through which children's growth is stunted.

"A study of the Mende found that senior wives did have higher fitness while junior wives had fewer surviving children than their counterparts in monogamous unions (Isaac and Feinberg 1982). Similarly, in Botswana, children of more senior wives enjoyed nutrition and school attendance advantages (Bock and Johnson 2002: 329)."

The author recalls seeing "a picture of a mother holding on her lap a boy and girl of about the same age, possibly twins. The girl was skeletal, obviously in an advanced state of malnutrition, the boy robust and healthy. He sat erect, eyes intent on the camera; she sprawled, like a rag doll, her eyes staring into space. That picture and what it represented has haunted me ever since."

Babies of the preferred sex are likely to be nursed longer and have higher survival rates.

Many folk traditions recommend foods for children, or diets for sick children, that are undernourishing or likely to be contaminated:

"Meat is usually among the foods kept from children. This is probably harmful, as a protein shortage, in particular, is often found in recently weaned children. However, malnutrition is rarely identified by parents as the root of a child's illness. Katherine Dettwyler pointedly titled her study of the Dogon Dancing Skeletons, describing, in graphic detail, the horrific sight of severely malnourished children. She finds that, while the mothers are aware of something amiss, they attribute the problem to locally constructed folk illnesses and seek medicine from the anthropologist to effect a cure. When she tells them to provide the child with more food, they are skeptical. Children can't benefit from good food because they haven't worked hard to get it, and they don't appreciate its good taste or the feeling of satisfaction it gives. Anyway, "old people deserve the best food, because they're going to die soon" (Dettwyler 1994: 94–95). Yoruba mothers feed children barely visible scraps compared to the portions they give themselves. Good food might spoil the child's moral character (Zeitlin 1996: 418; also true for the Tlingit – cf. de Laguna 1965: 17). The prescription for a sick child among the Gurage tribe in southwest Ethiopia is often the sacrifice of a sheep: "The flesh of the sacrificial animal is eaten exclusively by the parents of the sick child and others who are present at the curing rite; no portion of the meat is consumed by the patient, whose illness may well stem from an inadequate diet" (Shack 1969: 296)."

"Aside from a demonstrable shortage of food (Hill and Hurtado 1996: 319), under-nutrition may be attributable to customs that support a shortening of the nursing period, such as the belief by some East African pastoralists that certain babies nurse "too much" and should, therefore, be weaned early (Sellen 1995). On Fiji, nursing beyond one year is condemned as keeping "the child in babyhood [, leading to] a weak, simpering person" (Turner 1987: 107). The Alorese use threats to discourage nursing: "If you continue nursing, the snakes will come ... the toad will eat you" (Du

Bois 1941: 114)." (The WHO currently recommends breastfeeding until age 2 or beyond.)

While medical science considers the first milk (colostrum) to be especially beneficial to the newborn because of the antibodies it contains, folk tradition often withholds it from newborns: "In a survey of fifty-seven societies, in only nine did nursing begin shortly after birth (Raphael 1966)."

## Spacing children

Contrasted with agriculturalists who go for large families, "foragers adopt a "survivorship" reproductive strategy. Around-the-clock nursing and a post-partum sex taboo combine to insure long intervals between births, leading to lower fertility. Low fertility is offset by the attention bestowed on the few offspring, enhancing their chances of survival (Fouts et al. 2001)." Breastfeeding suppresses women's fertility.

"Another way in which nature contributes to increasing IBI [inter-birth interval] is through post-partum depression following a miscarriage, stillbirth, or infant death. Binser notes that depression elevates cortisol and leaves the mother lethargic and sleepy, which may just serve to put off the next pregnancy until she has had a chance to recoup her vigor (Binser 2004). Nature is aided by culture in promoting longer IBIs through injunctions that militate against long intervals between nursing bouts. Frequent, round- the-clock nursing maintains high prolactin levels. The post-partum taboo on intercourse between husbands and wives also plays a critical role in spacing births."

In other cases the mother is physically separated from her husband: "The wife may be lodged in a birthing or 'lying-in' house (Lepowsky 1985: 64), or secluded in her own home, until, in the Trobriands, 'mothers lost their tans and their skin color matched that of their infants' (Montague 1985: 89)."

In traditional societies, early sexual activity was less likely to result in pregnancy because adolescents were often malnourished and their fertility lower than we'd expect.

## Which children are preferred

I had assumed that boys were always preferred in traditional societies, but it depends. The gender preference, or lack therof, is influenced by parents' expectations of help their children will provide them with.

"There is a world in which children almost always feel "wanted" and where "there is no cultural preference for babies of either sex" (Howell 1988: 159). Infants are suckled on demand by their mothers and by other women in her absence. They are indulged and cosseted by their fathers, grandparents, and siblings. Children wean themselves over a long period and are given nutritious foods (Robson and Kaplan 2003: 156). They are subject to little or no restraint or coercion. Infants and toddlers are carried on long journeys and comforted when distressed. If they die in infancy, they may be mourned (Henry 1941/1964: 66). They are rarely or never physically punished or even scolded (Hernandez 1941: 129–130). They are not expected to make a significant contribution to the household economy and are free to play until the mid to late teens (Howell 2010: 30). Their experience of adolescence is relatively stress free (Hewlett

and Hewlett 2013: 88). This paradise exists among a globally dispersed group of isolated societies – all of which depend heavily on foraging for their subsistence. They are also characterized by relatively egalitarian and close social relations, including relative parity between men and women (Hewlett et al. 1998)."

"One thorough study compared Hungarian Gypsies (matriarchal) with mainstream Hungarian (patriarchal) society. Gender preferences were as expected and behaviors tracked preferences. Gypsy girls were extremely helpful to their mothers and tended to remain at home longer than their brothers, helping even after marriage. They were nursed longer than their brothers, while Hungarian boys were nursed longer than their sisters. "Gypsy mothers were more likely to abort after having had one or more daughters, while Hungarians are more likely to abort pregnancies when they have had sons" (Bereczkei and Dunbar 1997: 18)."

"Names such as "Boy Needed" (Oghul Gerek) or "Last Daughter" (Songi Qiz) are common for girls. (Irons 2000: 230)"

## Family structure

"We now realize that mothers, fathers, and children have differing agendas. The nursing child wants to be the last child his mother will ever have so that he can enjoy her care and provisioning exclusively. The father will be opportunistic in seeking mating opportunities and display a similar fickleness toward the provisioning of his offspring. He will, in other words, spread his investment around to maximize the number of surviving offspring. The mother has the most difficult decisions of all. She must weigh her health and longevity and future breeding opportunities against the cost of her present offspring, including any on the way. She must also factor in any resources that might be available from her children's fathers and her own kin network."

(Of course I can think of many loving and capable fathers, not least my own partner. But I was surprised that they seem to have historically played so little role in childrens' lives.)

Polygyny is a common traditional way of structuring families, "the great compromise" between these competing interests.

"Estimates range from 85 percent (Murdock 1967: 47) to 93 percent (Low 1989: 312) of all societies ever recorded (about 1,200) having practiced polygyny."

"Women in a polygynous relationship gain access to a higher-ranking, reliable provider at the cost of emotional strain in sharing resources (including the husband's affection) with others. In one study, children of senior wives were better nourished than children in monogamous unions, who were, in turn, better nourished than children of later wives (Isaac and Feinberg 1982: 632). A woman must weigh the trade-offs between marrying a young man in a monogamous union or marrying an older man and joining a well-established household as a junior wife. Studies show that, if they choose monogamy, they enjoy slightly higher fertility (Josephson 2002: 378) and their children may be somewhat better nourished (Sellen 1998a: 341). However, they are, perhaps, more likely to be abandoned or divorced by their husbands."

Both polygyny and monogamy have their pros and cons:

"In my fieldwork in Gbarngasuakwelle, I lived (as a guest) in a large, polygynous household and the tensions were palpable. This was seen as harmful to children. The shaman (village blacksmith in this case) came often to divine the cause and, using appropriate rituals (inevitably involving the sacrifice of a chicken), would attempt to ameliorate it (Lancy 1996: 167)."

"In Uganda, monogamy has led to less stable marriages. A man, rather than bringing a second wife into the household, now abandons the first wife and her children to set up a second separate household with his new mate (Ainsworth 1967: 10–11). A typical case among the Nyansongo in Kenya describes a mother, whose childhood was spent in a large polygynous compound where multiple caretakers were always available, who must cope alone in a monogamous household. She leaves her three-year-old to mind her six-month- and two-year-old infants as she performs errands like bringing the cow in from pasture. Unfortunately, the three-year-old is simply not mature enough for this task and is, in fact, 'rough and dangerously negligent' (Whiting and Edwards 1988a: 173)."

"As societies become more mobile and men migrate seeking employment, the likelihood that the male will abandon (or neglect) his family in the village in order to establish a new family in the city is increasingly high (Bucher and d'Amorim 1993: 16; Timaeus and Graham 1989). And, perhaps most common of all, women whose fertility is on the decline are replaced by younger wives in peak breeding condition (Low 2000: 325)"

"The abandoned spouse and her children may face severe difficulties. One might think that an obviously fertile woman would be a 'catch,' but 'Having a child towards whom a new husband will have to assume step-parental duties diminishes rather than enhances a woman's marriageability' (Wilson and Daly 2002: 307). "

"In the case of a young, pregnant widow, ancient Roman law permitted both annulment and the exposure of the infant in order to enhance her chances of remarriage (French 1991: 21). Raffaele describes an unfortunate case in a Bayaka<sup>13</sup> foraging band in Central Africa:

Mimba had been in a trial marriage ... her partner's father had refused to pay the bride price and she had just been forced to return to her own family. She is two months' pregnant, and it is a disgrace for an unmarried Bayaka woman to give birth" (Raffaele 2003: 129). Fortunately for Mimba, the tribe's pharmacopoeia includes sambolo, a very reliable and safe herbal abortifacient, which she will use. Mimba will return to the pool of eligible mates and, hopefully, will find a family willing to pay the bride-price so their son can join her in raising a family – something she could not accomplish by herself."

"Studies in the USA indicate that living with a stepfather and stepsiblings leads to elevated cortisol levels, immunosuppression, and general illness (Flinn and England 1995)<sup>31</sup> as well as poorer educational outcomes (Lancaster and Kaplan 2000: 196). Daly and Wilson find that a child is a hundred times more likely to be killed by a stepparent than by a biological parent (1984: 499).

Some form of fostering, adoption, or "child circulation" is practiced in many societies:

"Most commonly the child is transferred 'to fulfill another household's need for labor' (Fée – Martin 2012: 220) as a 'helper' (Inuit – Honigmann and Honigmann 1953: 46). The request may be for a girl in families with a shortage of female labor (Kosrae – Ritter 1981: 46; Bellona – Monberg 1970: 132). On Raroia boys are requested as they

can work in copra processing (Danielsson 1952: 120). On the other hand, the impetus may begin with a family that has a surplus of children (Bodenhorn 1988: 14), or children too close in age, or discord within the family; or as the means to defray a debt. Stepchildren are often moved out of the natal home to make way for the new parent's biological offspring."

## Life and death

The topic that most surprised me in the book was traditional attitudes toward abortion and infanticide. I thought of life before birth control as "the bad old days" when women, perhaps not even understanding how babies are conceived, might be sentenced to a lifetime of childbearing and rearing against their wishes. I had never thought about how traditional societies actually handled unwanted babies.

"Data from a range of societies past and present suggest that from one-fifth to one-half of children don't survive to five years (Dentan 1978: 111; Dunn 1974: 385; Kramer and Greaves 2007: 720; Le Mort 2008: 25). The first-century CE philosopher Epictetus cautioned, "When you kiss your child, say to yourself, it may be dead in the morning" (Stearns 2010: 168).

"Extrapolating from these figures I'd guess that miscarriages and stillbirths were also common by comparison with modern, post-industrial society. And I'd expect that if half the children died, then the majority were seriously ill in childhood. Indeed, in many villages studied by anthropologists the level of clinical malnutrition is 100 percent, as is the level of chronic parasite infestation and diarrhea. There are, then, ample reasons for withholding investment in the infant and maintaining a degree of emotional distance."

"Humans have always had to cope with the loss of infants, and societies have developed an elaborate array of "cover stories" to lessen grief and recrimination (Martin 2001: 162; Scrimshaw 1984: 443). As discussed in the previous chapter, the primary strategy is to treat the infant as not yet fully human. Most importantly, if the baby is secluded initially and treated as being in a liminal state, its loss may not be widely noted."

Some societies believed repeated miscarriages or stillbirths were caused by demons, and treated them with various attempts at exorcism. "It should be understood that these folk theories and treatments not only serve to dampen the sense of grief or loss but, more importantly, they deflect blame from the living. The Nankani have constructed an elaborate myth of the "spirit child not meant for this world" to explain away the tragedy of mother or infant death in childbirth and/or chronic infant sickness and, eventually, death (Denham 2012: 180). The alternative to, in effect, blaming the deceased child or "evil forces" is to blame the parents or other family/community member."

"While new mothers may be evaluating the actuarial odds, we know that many are also suffering from post-partum depression or, less severely, detachment from and indifference toward their offspring. An argument can be made that this failure to bond immediately with the infant is adaptive in that it permits the mother to keep her options open, and also shields her emotionally from the impact of the infant's death – often, a likely outcome (de Vries 1987a; Eible-Eibesfeldt 1983: 184; Hagen 1999; Konner 2010: 130, 208; Laes 2011: 100)."

"In the Himalayan kingdom of Ladakh, high-altitude living imposes an extra cost on the expectant mother who does farm-work throughout her pregnancy. Her infant's life chances, owing to inevitably low birth-weight and other complications, are sharply reduced (Wiley 2004: 6). The worth of a new child in Ladakh will always be calculated as a tiny fraction of that of his fully mature, productive mother. While the mother's health is closely monitored and she is treated with great solicitude, her infant's fate is of less concern. Its death will be "met with sadness, but also with a sense of resignation ... they are buried, not cremated like adults" (Wiley 2004: 131-132)."

"It is not unusual for the [Ayoreo] newborn to remain unnamed for several weeks or months, particularly if the infant is sickly. The reason given is that should the child die, the loss will not be so deeply felt. (Bugos and McCarthy 1984: 508)"

"Being a "calculating" mother is not synonymous with wickedness; on the contrary, it is adaptive behavior. While the well-to-do mothers in the first section seem to "live for their children," in the next section, we discover just how recently these attitudes have become incorporated in Western society. We will trace the fluctuating value of infants in history and see that what we now consider horrible crimes were, in earlier periods, the principal means of birth control."

In ancient Greece, "Illegitimacy was usually a death sentence. "Identity was given by the family, and without a recognized father and family, the child had no proper guardian (*kurios*) since its mother could not legally fulfill such a function. Without a father, the child had no true place in the patrilineal kin structure, no right to the family name" (Patterson 1985: 115). Until at least the end of the eighteenth century, any Venetian infant of questionable parentage would have been abandoned or destroyed (Ferraro 2008)."

"While the termination of the fetus or of the infant's life is most often the parents' decision and we've seen numerous possible reasons for this behavior, societies often legitimize that decision. Overpopulation, the burden on the community of a hard-to-raise child, the social disharmony created by illegitimacy: all give the society a stake in this critical decision. Ultimately, also, the community must value the life and emotional wellbeing of its experienced, productive adult females over any potential value a tiny infant might have."

In foraging societies, "Both men and women face significant health and safety hazards throughout their relatively short lives, and they place their own welfare over that of their offspring. A survey of several foraging societies shows a close association between the willingness to commit infanticide and the daunting challenge "to carry more than a single young child on the nomadic round" (Riches 1974: 356)."

"The Inuit, among others, were known to cull females in anticipation of high mortality among males through hunting accidents, homicide, and suicide (Dickemann 1979: 341)."

Twins, being hard to nourish, were often discarded: "Mothers are unable to sustain two infants, especially where both are likely to be underweight. As Gray (1994: 73) notes, "even today, with the availability of western medical services it is difficult to maintain twins." On Bali, which is otherwise extraordinary in its elevation of babies to very high esteem, giving birth to more than one child at a time is seen as evidence of incest. Priests consider the birth of twins as sub-human or animal-like (Lansing 1994; Barth 1993; Belo 1980). Similarly, the Papel (Guinea-Bissau) believe that it is mufunesa to give birth to many children at the same time like animals. Pigs have many offspring.

Human beings give birth to only one each time. Therefore twins have to be thrown away. If not, the father, the mother, or somebody in the village may die. (Einarsdóttir 2004: 147).

Among the !Kung, Nancy Howell found that mothers whose toddlers had not been weaned might terminate the life of their newborn. In a society with high infant mortality (IM), an unweaned but otherwise thriving child is a better bet than a newcomer of unknown viability. The mother is expected by the band to kill one of a pair of twins or an infant with obvious defects. She would not be committing murder because, until the baby is named and formally presented in camp, it is not a person (Howell 1979: 120).

We can juxtapose this picture – paralleled in pre-modern communities the world over – with the almost legendary affection and love the !Kung show their young (Konner 2005). Similarly, Trobriand Island (Papua New Guinea) women, who also shower affection on their children, “were surprised that Western women do not have the right to kill an unwanted child … the child is not a social being yet, only a product manufactured by a woman inside her own body” (Montague 1985: 89). ”

“In farming communities, additional farmhands are usually welcomed. Still, in rural Japan, a family would be subjected to considerable censure for having “too many” children and might find themselves ostracized if they failed “to get rid of the ‘surplus’” (Jolivet 1997: 118; see also Neel 1970). Bear in mind that breastfeeding is more costly – metabolically – than pregnancy (Hagen 1999: 331). In the impoverished northeast of Brazil, women can count on very little support from their child’s father, and their own resources are meager. Hence, “child death a mingua (accompanied by maternal indifference and neglect) is understood as an appropriate maternal response to a deficiency in the child. Part of learning how to mother … include[s] learning when to ‘let go’” (Scheper-Hughes 1987b: 190). Early cessation of nursing – one manifestation of the mother’s minimizing her investment – is supported by an elaborate folk wisdom that breast milk can be harmful, characterized as “dirty,” “bitter,” “salty,” or “infected.” Another folk illness category, doença de criança, is used flexibly by mothers in justifying a decision to surrender the child into the hands of God or, alternatively, raise it as a real “fighter.” Of 686 pregnancies in a sample of 72 women, 251 infants failed to reach one year of age (Scheper-Hughes 1987a). ”

“Long before the “one-child policy,” abortion was common in China. The oldest Chinese medical text found so far, some 5,000 years in age, includes reference to mercury as an abortifacient.”

I also hadn’t thought about traditional attitudes toward children with disabilities, or children (perhaps with autism) who don’t engage in eye contact, smiling, and other behavior that charms adults. Hrdy (in press) suggests that the infant’s gaze-following and close attention to facial expressions and moods – along with a plump body and other neotenous features – are designed to send a clear signal to its mother and other caretakers: “Keep me!” ”

“In earlier times, the “difficult” or unwanted child might be dubbed a “changeling” or devil-inspired spirit, thereby providing a blanket of social acceptability to cloak its elimination (Haffter 1986). In cases where mothers are forced to rear unwanted children, the young may suffer abuse severe enough to end their life. While our society may treat such behavior by the parent as a heinous crime, “This capacity for selective removal in response to qualities both of offspring and of ecological and social

environments may well be a significant part of the biobehavioral definition of *Homo sapiens*" (Dickeman 1975: 108)."

"Changelings represent a special sub-group of "demon" children who provoke a negative response from caretakers. The changeling was an *enfant changé* in France, a *Wechselbag* in Germany, and, in England, a "fairy child." Strategies to reverse the switch included tormenting the infant or abandoning it in a lonely spot (Haffter 1986). A Beng mother-to-be who breaks a taboo may have her uterus invaded by a snake. The snake takes the fetus's place and, after birth, is gradually revealed by the infant's strange behavior. "The child may be harassed and hit by stones; however, being boneless like a snake, the snake-person is thought to feel no pain" (Gottlieb 1992: 145). A Papel infant deemed abnormal may be a spirit that's entered the mother's uterus. Two procedures are available to determine whether the child is human, but surviving either procedure seems improbable (Einarsdóttir 2008: 251). Dogon children thought to be evil spirits are taken: Out into the bush and you leave them ... they turn into snakes and slither away ... You go back the next day, and they aren't there. Then you know for sure that they weren't really [Dogon] children at all, but evil spirits. (Dettwyler 1994: 85-86) Among the Nuer, it is claimed, a disabled infant was interpreted as a hippopotamus that had mistakenly been born to human parents; the child would be returned to its proper home by being thrown into the river. (Scheer and Groce 1988: 28) In ... northern Europe, changelings were left overnight in the forest. If the fairies refused to take it back, the changeling would die during the night - but since it was not human, no infanticide could have occurred. (Hrdy 1999: 465) [For Lurs] Djenn are said to be ... jealous of the baby, especially during the first ten to forty days; they might steal the baby or exchange it for their own, sickly one. A baby indicates that it might be a changeling by fussiness, weakness, or lack of growth. (Friedl 1997: 69)"

## Foragers vs. agriculturalists

Attitude toward children in general seems to vary by livelihood.

"In Central Africa, systematic comparisons have been drawn between foragers and farmers in the same region. Bofi-speaking foragers follow the !Kung model. Babies are carried or held constantly, by mothers and fathers, are soothed or nursed as soon as they cry, and may wean themselves after three to four years. Children are treated with the affection and respect consistent with preparing them to live in an egalitarian society where the principal subsistence strategy is cooperative net-hunting. Bofi-speaking farmers, on the other hand, tend not to respond as quickly to fussing and crying, are likely to pass the infant off to a slightly older sibling, and are verbally and physically abusive to children, who are treated like the farmhands they are soon to be."

"The Garo, who live in the forests of Bengal, all share in infant and childcare, and parents "seldom roughhouse with their children, but play with them quietly, intimately, and fondly" (Burling 1963: 106). In the Northwest Territory of Canada, the Inuit (aka Eskimo) would never leave a child alone or let it cry for any length of time. Infants receive a great deal of solicitous care and lots of tactile comfort, anticipatory of "the interdependence and close interpersonal relations that are an integral part of Inuit life" (Condon 1987: 59; Sprott 2002: 54).

Draper observed a similar mindset operating among !Kung foragers in the Kalahar: Adults are completely tolerant of a child's temper tantrums and of aggression directed

by a child at an adult. I have seen a seven-year-old crying and furious, hurling sticks, nutshells, and eventually burning embers at her mother ... Bau (the mother) put up her arm occasionally to ward off the thrown objects but carried on her conversation nonchalantly. (Draper 1978: 37)"

## How children are expected to speak

"Clearly Euroamerican and Asian parents are preparing children to be more than merely competent native speakers. They encourage the development of narrative ability through frequent queries about the child's activity, including their subjective assessments: "mothers pick up on children's ... topics, repeat and extend what their children say, and adjust their language ... to support the child's projects" (Martini 1995: 54). Toddlers are expected to hold and to voice their opinions! As parents seek "explanations" from their children, they also tolerate interruptions and contradiction (Portes et al. 1988). And this entire package of cultural routines is almost completely absent in the ethnographic record (Robinson 1988).

"In a Mayan community ... children are taught to avoid challenging an adult with a display of greater knowledge by telling them something" (Rogoff 1990: 60). West African Wolof parents never quiz their kids by asking known-answer questions (Irvine 1978) – a favorite trick of Euroamerican parent-teachers. Fijian children are never encouraged to address adults or even to make eye contact. Rather their demeanor should express timidity and self-effacement (Toren 1990: 183)."

"Qualities we value, such as precocity, verbal fluency, independent and creative thought, personal expression, and ability to engage in repartee, would all be seen by villagers as defects to be curtailed as quickly as possible.<sup>25</sup> These are danger signs of future waywardness. "Inquisitiveness by word or deed is severely censured, especially in [Kogi] women and children" (Reichel-Dolmatoff 1976: 283). "A [Sisala] child who tries to know more than his father is a 'useless child' (bichuola), for he has no respect" (Grindal 1972: 28). In rural Turkey the trait most valued by parents (60 percent) was obedience; least valued (18 percent) was independence (Kagitçibasi and Sunar 1992: 81)."

## How do children learn?

"I discuss the prevailing view in WEIRD society – among most scholars as well as the public at large – that children's development into mature, competent members of society depends critically on the guidance and lessons, beginning in infancy, provided by an eager parent who's a "naturally gifted" teacher. Based on unequivocal evidence of the relative unimportance of teaching in the ethnographic record, I question that assumption as well as its evolutionary foundation."

"De León (2012) records an episode from her Zinacantecan site where a three-year-old boy nearly runs, barefoot, through a fire. Adults do not react sympathetically. Instead, they comment that the child is flawed in not developing awareness of its surroundings, not paying close attention, and not figuring things out. There is an uneasy trade-off here. On the one hand, by indulging their curiosity about the environment and the things in it, parents insure that children are learning useful information without the necessity of parental intervention. This efficiency comes at a cost of the occasional damage to or loss of one's offspring (Martini and Kirkpatrick 1992)."

"Active or direct teaching/instruction is rare in cultural transmission, and that when it occurs, it is not aimed at critical subsistence and survival skills – the area most obviously affected by natural selection – but, rather, at controlling and managing the child's behavior."

"Outside WEIRD or post-industrial society, this suite of parent-infant interaction patterns is rare. Mothers don't often engage cognitively with infants, they may only respond contingently to their distress cues, and they probably do not gaze at them or engage in shared attention to novel objects (de León 2011: 100; Göncü et al. 2000; LeVine 2004: 161)."

In most traditional societies, children and young adults are expected to learn by observation rather than direct teaching.

In a Guatemalan indigenous community where people use a traditional learning style to approach factory work: "The newly hired worker performs menial tasks<sup>33</sup> such as bringing material to the machine or taking finished goods off of it, but most of the time is spent observing the operations of the person running the machine. [The new worker] neither asked questions nor was given advice. When the machine snagged or stopped, she would look carefully to see what the operator did to get it back into motion ... This constituted her daily routine for nearly six weeks, and at the end of this time she announced that she was ready to run a loom ... and she operated it, not quite as rapidly as the girl who had just left it, but with skill and assurance ... at no time during her learning and apprentice period had she touched a machine or practiced operating ... She observes and internally rehearses the set of operations until she feels able to perform. She will not try her hand until she feels competent, for to fumble and make mistakes is a cause for verguenza – public shame. She does not ask questions because that would annoy the person teaching her, and they might also think she is stupid. (Nash 1958: 26-27)"

"I provide an extended example, of mother Sua and daughter Nyenpu each weaving a fishnet. As the vignette unfolded, the main point seemed to be how little interest Sua had in getting involved in Nyenpu's weaving. Sua claimed that her stance was typical and replicated her own mother's attitude when she was learning net-weaving. Several other informants told me of approaching experts for help and being rebuffed (Lancy 1996: 149-150). Other ethnographers report similar tales. Reichard describes a Navajo girl who learned to weave in spite of her mother's repulsing her interest (1934: 38), which paralleled a case from Truk of a weaver/basket-maker whose kin were unsupportive of her efforts to learn their skills (Gladwin and Sarason 1953: 414-415), and a case from the Venda tribe where a potter is vehement that "'We don't teach. When women make pots some (children and others) come to watch, then go and try'" (Krause 1985: 95)."

A Javanese shellfish diver responds to the question of whether she learned the practice from her mother:

"My mother! she said loudly, She drove me away! I tried to follow her to the bottom to watch, but she shoved me back. When we were on the surface again, she practically screamed at me to move OFF and find my danged abalone BY MYSELF. So we had to discard [one] cliché about how artisans learn. (Hill and Plath 1998: 212)"

There are a few cases of explicit teaching:

"There are a few cases in the literature of grandmothers conducting educational tours through the bush to acquaint their younger kin with medicinal plants (Ngandu –

Hewlett 2013: 76; Tonga – Reynolds 1996: 7)."

"An interesting "work-around" for the prohibition on teaching is provided by the Fort Norman Slave [Canada], who hunt during severe winter weather and must traverse ice-fields. Fathers "instruct" sons about this dangerous environment (which comprises thirteen kinds of ice and multiple modes of travel) via a game-like quiz (Basso 1972: 40)."

## Analytic thinking

While there's a lot of knowledge being transmitted in traditional societies, like how to make and use a blowgun for hunting or how to hollow a canoe, analysis and taxonomy seem to be absent in societies where people haven't gone to school. Lancy cites Alexander Luria's 1930s interviews with peasants in Central Asia:

"In the first example we can see the villager reasoning from personal experience (or lack thereof) and inability or unwillingness to apply a general rule. 'Problem posed: 'In the Far North, where there is snow, all bears are white. Novaya Zemlya is in the far north and there is always snow there. What color are the bears?' Response: 'We always speak of only what we see; we don't talk of what we haven't seen.' (Luria 1976: 108)

In another problem, men and women were asked to sort and group various kinds and colors of weaving yarn (Uzbekistan is noted for its carpets). The male response was 'men [not being weavers] don't know colors and call them all blue.' The women refused to impose any grouping or organization – something educated Uzbeks did quite easily – exclaiming that 'none of these are the same' (Luria 1976: 25, 27).

In a fishing community in Sulawesi, Vermonden found directly parallel results, with fishers resistant to discussing marine life more generally; they eschewed speaking of types of fish or of considering different ways of grouping them. Their thinking was governed by their practice (true also for Penan hunters – Puri 2005: 280 – and South American and African subsistence farmers – Henrich et al. 2010: 72). . . Had Vermonden's informants been schooled, they might have used broader and more inclusive organizing principles and been able to display a more encyclopedic knowledge of fish." I assume that these people did in fact know a lot about fish, yarn, etc, which were their daily livelihood, but were used to thinking in practical terms.

(I was telling Jeff the bear example at dinnertime. "It's white," piped up our four-year-old without prompting. She's used to "known-answer questions" where grownups ask you things even though they know the answer.)

## Learning to be street-smart

While children in some societies need to learn to avoid predators and poisonous plants, Lancy also briefly covers urban environments where children must be equipped for other dangers. "A mother in a favela of Rio de Janeiro knows "intuitively that in order for her children to survive, toughness, obedience, subservience, and street smarts are necessary; otherwise, the child can end up dead" (D. Goldstein 1998: 395)."

Charles Dickens depicts a similar strategy in 19th-century London, with a father describing how he's trained his son: "I took a good deal o' pains with his eddication, sir; let him run in the streets when he was very young, and shift for hisself. It's the only way to make a boy sharp, sir" (Dickens 1836/1964: 306)."

## Learning through play

"Play is a truly universal trait of childhood. The one thing that children can appropriate for themselves, without the sanction of culture or explicit blessing of parents, is play. It is ubiquitous. A baby will play with its mother's breast. The first glimmer of understanding about the natural world and how it works comes through play with objects. After its nurturing mother, the child's first close relationships are with its playmates – usually siblings. The child's first active engagements with the tasks that will occupy most of its adult life – hunting, cooking, house-building, baby-tending – all occur during make-believe."

"Many of the child's most basic needs seem to be fed by play – their need to socialize with peers and their need for physical, sensory, and, to a lesser extent, cognitive stimulation (Lancy 1980a). The demands of earning a living and reproduction gradually extinguish the desire to play. This happens earlier in girls than in boys – almost universally."

Modern children:

"To encourage object play, we provide lots of toys, including safe, miniature tools, in various sizes, along with the dolls to use them. We also provide objects to play with that are specifically designed to facilitate the kind of cognitive complexity and flexibility that many assert is the *raison d'être* of object play (Power 2000). And, what is perhaps most remarkable, we sometimes intervene to "teach" our children how to use their toys or nudge them into more complex uses (Gaskins et al. 2007). I have found only one example of this in the ethnographic literature – a Wogeo father assisting his son with a miniature canoe (Hogbin 1946: 282) – and I am confident it occurs rarely. In research where the investigators created conditions designed to facilitate their involvement, East Indian and Guatemalan villagers would not intervene in their toddlers' play (Göncü et al. 2000). It's hard to escape the conclusion that our "micro-management" of children's toys and play is driven by the inexorable demands of schooling."

In contrast to play with specially provided objects, social play and pretend play are ubiquitous.

"Comparing across fifteen species of primates, observers found a statistically reliable relationship between cerebellum size and time devoted to social play (Lewis and Barton 2004; see also Fisher 1992)."

"This rapid growth in understanding – correlated with a rapidly growing brain – emerges in early childhood as two powerful motives. These are, first, to "fit in," to be liked, appreciated, and accepted. The second motive force is a drive to become competent, to replicate the routine behaviors enacted by those who're older and more capable. The presence of these drives accounts for the child's ability to learn through observation, imitation, and, by extension, playing with objects and ideas in make-believe."

"Esther Goody describes the richness and complexity of make-believe cooking in a village in north Ghana. Miniature kitchens are constructed, ingredients gathered, and soup made, all the while accompanied by singing and the construction of play scripts that mimic adult discourse. And, of course, the girls must insure that their play enfolds the younger siblings who are in their care. Boys have bit parts in these playlets as "husbands," and are limited to commenting on the flavor of the soup (Goody 1992).



*My kids and their cousins are avid mud-soup makers. Boys are full participants in this case*

"Make-believe reveals children's insight into the adult world. Araucania boys accurately mimic the speech and movements of drunken males celebrating fiesta (Hilger 1958: 106). Yanamamo boys pretend to "smoke" hallucinogens and then stagger around in perfect imitation of their stoned fathers acting as shamans (Asch and Chagnon 1974)."

Of course, an anthropologist in the village provides an interesting topic for pretend play: "Parenthetically, many an anthropologist has seen herself or himself reflected (unflatteringly) in the play of erstwhile subjects (Bascom 1969: 58)."

"The doll is arguably the most widely found toy and the range of materials used and designs employed is immense (Ruddle and Chesterfield 1977: 36).<sup>16</sup> From rags tied into a shapeless bundle to high-tech baby dolls that produce a babble of baby-talk, wet themselves, and eagerly move their limbs, the variety is fascinating. While baby dolls seemed to have been a universal adjunct to Roman girls' play, lower-class girls had infant dolls that they mock-nursed, comforted, and cleansed while upper-class girls, whose future as adults would not include childcare, dressed and primped the ancient equivalent of Barbie (Wiedemann 1989: 149-150).

## **Not all cultures encourage play**

"Play may be seen as a sign of waywardness. Bulusu' view play as naughty (jayil) and those who play "too much" as crazy (mabap) (Appell-Warren 1987: 160). Children may be scolded for getting dirty or telling stories they know aren't true (e.g. fantasizing)

(Gaskins et al. 2007: 192). On Malaita Island, where children are expected to carefully observe and report on newsworthy events in the village, children's fantasy constructions are discouraged; they "are mildly reprimanded with 'you lie'" (Watson-Gegeo and Gegeo 2001: 5).

Following the Protestant Reformation, many influential authorities condemned play in general as well as specific kinds of play, such as solitary play or contact sports. Morality came to be equated with decorum and emotional restraint; "indulging children was a cardinal sin" (Colón with Colón 2001: 284).

Similar sentiments were expressed by Chinese sages: Huo T'ao had no tolerance for play ... as soon as a child is able to walk and talk, it must be taught not to play with other children. Children must practice treating one another as adults ... When [children] see each other in the morning, they must be taught to bow solemnly to each other. (Dardess 1991: 76)"

## **When do parents play with young children?**

Parents playing with babies varies a lot:

"An analysis of 186 archived ethnographies of traditional societies indicated wide variation in the amount of mother-infant play and display of affection (Barry and Paxson 1971). In a more recent comparative observational study, "Euro-American adults were much more likely than Aka or Ngandu adults to stimulate (e.g., tickle) and vocalize to their infants. As a result, Euro-American infants were significantly more likely than Aka and Ngandu infants to smile, look at, and vocalize to their care providers" (Hewlett et al. 2000: 164).<sup>21</sup> Play with infants also seems generally less common among agrarian societies; for example, an Apache (North American agro-pastoralists) "mother sometimes plays with her baby ... A father is not likely to play with a baby" (Goodwin and Goodwin 1942: 448). In hundreds of hours of close observation of parent-child interaction among Kipsigis (Kenyan) farmers, Harkness and Super (1986: 102) recorded "no instances of mothers playing with their children.""

"Among the !Kung, parents not only don't play with their children post-infancy, they reject the notion outright as potentially harmful to the child's development. They believe that children learn best without adult intervention (Bakeman et al. 1990: 796). The mother of a toddler not only faces potential conflict between childcare and work, she's likely pregnant as well. I would argue that the mother's greatest ally, at this point in the childrearing process, is the magnetic attraction of the sibling or neighborhood play group (Parin 1963: 48). The last thing a pregnant mother wants is for her child to see her as an attractive play partner. Even verbal play is avoided."

I found this such a relief to read. The hardest stage of parenting for me was caring for an infant while being my two-year-old's only regular playmate. She had an insatiable desire for stories, and I just wasn't up for it.

Young monkeys play with each other, but chimpanzee mothers play with and tickle their babies.

"Why is the chimpanzee mother providing her baby with what monkey infants get from their peers? One clue in the direction of an answer may be the group structure of chimpanzees. I observed that chimpanzee mothers spend most of their time alone

with their babies. As a consequence it is the chimpanzee mother who has to give her baby this sort of interaction if he gets it at all. (Plooij 1979: 237) "

Similar forces may promote mother-child play among humans. The small band of "Utkuhikhalingmiut [Inuit are] the sole inhabitants of an area 35,000 or more miles square" (Briggs 1970: 1). Aside from the almost total lack of other children to play with, the mother-child pair is isolated inside their igloo for days on end during the worst weather. Jean Briggs observed mothers talking to their children, making toys for them, playing with them, and encouraging their language development. Further, there is every reason to believe that modern living conditions in which infants and toddlers are isolated from peers in single-parent or nuclear households produce a parallel effect. That is, like chimps in the wild, modern, urban youngsters only have access to their mothers as potential play partners. In Japan, the mother-child pair has become quite isolated, sequestered in high-rise apartment buildings."

This sounds very familiar to me.

## Learning through chores

Learning to do the chores of adult daily life is of great interest to children everywhere.

"In the Giriama language the term for a child roughly two through three years in age is kahoho kuhuma madzi: a youngster who can be sent to fetch a cup of water ... A girl, from about eight years until approximately puberty, is muhoho wa kubunda, a child who pounds maize; a boy of this age is a muhoho murisa, a child who herds. (Wenger 1989: 98)"

"Generally speaking, a girl's working sphere coincides with that of her mother: the household, kitchen, nursery, laundry, garden, and market stall. (Paradise and Rogoff 2009: 113) depict a five-year-old Mazahua girl closely following her mother's lead in setting up an onion stand in the market - trimming, bunching, and arranging their onions. When invited to establish a satellite onion stand, 'her excitement is unmistakable and she quickly takes the initiative in finding an appropriate spot and setting it up.'"

"In WEIRD society, parents and adults generally take every opportunity to instruct children, even when they are patently unmotivated or too awkward and immature. The term "scaffolding" may be used to describe the process whereby the would-be teacher provides significant assistance and support so that the novice can complete a task that is otherwise well beyond his grasp (McNaughton 1996: 178). Elaborate scaffolding is rarely seen elsewhere (Chapter 5). No one wants to waste time teaching novices who might well learn in time without instruction."

"Little girls strap bundles of leaves on their backs as babies, boys build little houses ... A little girl accompanying her mother to the fields practices swinging a hoe and learns to pull weeds or pick greens while playing about ... Playing with a small gourd, a child learns to balance it on his head, and is applauded when he goes to the watering-place with the other children and brings it back with a little water in it. As he learns, he carries an increasing load, and gradually the play activity turns into a general contribution to the household water supply. (Edel 1957/1996: 177)"

"In the Sepik region of Papua New Guinea, Kwoma children eagerly embrace the piglets they're given to protect, raise, and train (Whiting 1941: 47). Talensi boys are

said to possess “a passionate desire to own a hen” (Fortes 1938/1970: 20). ”

“The Touareg boy progresses from a single kid (at three years of age) to a herd of goats (at ten) to a baby camel (at ten) to a herd of camels (at fifteen) to managing a caravan on a trek across the Sahara (at twenty). Preferentially, the aspirant herder interacts with and learns from herders who are slightly older, not adults. Adults are too forbidding to ask questions of or display ignorance in front of. Above all, it is a hands-on experience, as “The abstract explanation so typical of our schooling is completely absent” (Spittler 1998: 247). ”

“Four-year-old Bafin has already grasped the meaning of sowing and is able to perform the various movements ... he is entrusted with an old hoe as well as with some seeds so that he can gain some practice in this activity. However ... he has to be allocated a certain part of the field where he neither gets in the way of the others nor spoils the rows they have already sown ... As a rule, his rows have to be re-done. (Polak 2003: 126, 129)”

This is one of the few passages that got at my concern about children’s involvement in chores: it usually creates more work for the parents. It did persuade me to let Anna load the dishwasher, which she does ineptly but avidly.

## **Chores vs. crafts**

“I was surprised to discover that, in Gbarngasuakwelle, there is a gulf between the chore curriculum and what we might call the craft curriculum. The former is often compulsory - a child may be severely chastised or beaten for failure to complete appropriate chores satisfactorily. The latter is not only entirely voluntary, but children seem to be offered little encouragement in it. Indeed, they may be actively discouraged from trying to learn a craft or otherwise complex trade.”

“Somewhat later, the child may elect to move beyond the core skills expected of everyone to tackle more challenging endeavors such as learning pottery or weaving. She or he must demonstrate adequate strength, physical skill, and motivation before anyone will deign to spend time on his or her instruction.”

## **Rites of passage**

Most traditional societies involve some initiation ceremony to mark the transition to adulthood, may involving “days of hazing, fasting, beating, sleeplessness, and sudden surprises.”

After being raised by women, boys’ rituals often focus on separating them from the world of women:

“One element that looms large in the training of male adolescents in much of Africa and Papua New Guinea is misogyny, as noted above. There is a distinct focus on teaching boys to feel superior toward and contemptuous of women. The “text” of many messages conveyed to initiates is replete with references to women’s physical weakness relative to men and their power to pollute through menstrual and puerperal blood. Another tool in the men’s arsenal is the use of “secrets,” including sacred terms, rituals, locations, and objects such as masks. These “secrets” are denied to women on pain of death. For the Arapesh (Sepik Region), “initiation ceremonies

[include] an ordeal followed by the novices being shown the secret paraphernalia ... flutes, frims, paintings, statues, bullroarers" (Tuzin 1980: 26). Denying female access to powerful spirit forces aids in maintaining male hegemony. A Mehinacu girl "cannot learn the basic myths because the words 'will not stay in her stomach'" (Gregor 1990: 484). Wagenia "women and girls belong to the social category of the non-initiated, from whom the secrets of initiation were carefully concealed" (Droogers 1980: 78)."

"Immediately following [the ordeal], the initiators drop their razors, spears, cudgels or what have you, and comfort the boys with lavish displays of tender emotion. What resentment the latter may have been harboring instantly dissipates, replaced by a palpable warmth and affection for the men who, moments before, had been seemingly bent on their destruction. As their confidence recovers itself, the novices become giddy with the realization that they have surmounted the ordeal. (Tuzin 1980: 78)"

"The Hitler Youth and the Soviet Young Pioneers both capitalized on the idealism and fanaticism characteristic of adolescence (Valsiner 2000: 295; see also Kratz 1990: 456). During the Cultural Revolution, Chinese authorities used the naturally "anti-social," rebellious nature of adolescents in recruiting, training, and then setting them loose as "Red Guards" to destroy bourgeois, Western, or intellectual elements of Chinese society (Lupher 1995). Today, Muslim terrorist organizations easily recruit male and female adolescents to serve as suicide bombers. Again, there are fundamental biological and psychological aspects of adolescence that render them susceptible to group-think mentality. Normal standards of human decency are suspended, allowing them to commit crimes in the name of the group."

## **Neither here nor there**

The author describes the plight of young people who have been socialized away from their traditional cultures but not given anything good in exchange:

"Christian missions offer them the opportunity to escape the restrictions imposed by traditional rites associated, in the Sepik area, with the men's Haus Tambaran, without successfully socializing them to embrace Western/Christian values. Similarly, in attending government schools, young males signal their abandonment of the traditional agrarian economy without actually learning enough to secure a job in the modern economy. In short, they have been led to believe they are superior to the senior men, yet bring no significant resources to the community"

"Disaffected African students, their hopes for white-collar jobs dashed by stagnant economies, are easily recruited as "rebels" (Lancy 1996: 198) and street rioters (Durham 2008: 173). Terrorists and rebel armies capitalize on the peculiarities of adolescent psychology, brought on in part by "living in limbo," to create pliable fanatics (Rosen 2005: 157). Rosen also notes the continuity between traditional Mende warrior training, described earlier in this chapter, and the recruitment and training of child soldiers."

The decades-long Salvadoran Civil War raised a generation of men with no livelihood other than war:

"Initiation rites in the socialization of young rebels, unlike traditional rites, do "not facilitate their social transition into responsible adulthood" (Honwana 2006: 63). Similarly, in the Salvadorian civil war, young soldiers "were not given a chance to practice and learn how to be campesinos, dedicated to subsistence agriculture ... and

the lack of preparation for a new, adult peacetime identity led many youth to choose the negative identity of ... marero [delinquent/gang member]. (Dickson-Gómez 2003: 344-345)"

"Similarly, adolescent males living on Indian reservations suffer mortality and suicide rates three times the national average."

## Traditional cultures meet Western schools

Western schools were historically places where knowledge is crammed and beaten into children.

4000 years ago a Sumerian student described his day: "My headmaster read my tablet, said: 'There is something missing,' caned me. 'Why didn't you speak Sumerian,' caned me. My teacher said: 'Your hand is unsatisfactory,' caned me.' And so I began to hate the scribal art" (Kramer 1963: 238-239)."

"Until fairly the 1970s, elite English boarding schools (and their US counterparts) for males weren't all that different in terms of the constant hazing of younger by older boys, the emphasis on physical deprivation and removal from family, and daily engagement in team sports. This is probably what prompted Arthur Wellesley, the duke of Wellington, to remark: 'The battle of Waterloo was won on the playing-fields of Eton.'"

"The lamentations of passionate critics provide another window on the nature of schooling. These critics believed that the reluctant scholar problem could be solved by making schooling more like the experiences of the unschooled child, mixing in play, letting the child make choices, rewarding curiosity and independent learning. The fact that these pleas continue to appear over nearly two millennia suggests how enduring and intractable were the earliest ideas about the nature of schooling."

"The idea that school should interest children was considered a radical new pedagogical philosophy in the United States of the 1840s"

But although the West has moved into a more child-centered mode, schools in the developing world remain on an old-fashioned model:

"As schools are introduced to formerly school-less communities, they much more closely resemble medieval schools than they do modern, progressive institutions. Bare, drafty classrooms, rote memorization, a scarcity of teaching materials, corporal punishment, unintelligible teachers, menial labor by students, the underrepresentation and exploitation of girls - all harken back to the dawn of schooling in the West."

"Schools have encountered resistance from pupils who struggle to "sit still" or to meet the teacher's gaze; from parents who'd prefer their children to be working and who reject their assigned role as "under-teacher," prepping and supporting their child's schooling; from patriarchal societies that impose limits on the choices available to women; and from the general public because of the very poor quality of instruction and the coercive atmosphere."

"In a survey of childhood across history and culture, the suite of practices and teaching/learning abilities associated with modern schooling is largely absent"

An anthropologist “marvels at how facile and active the Matses children are in the natural environment, compared to what she feels is her own ineptitude. She is cowed by three- and four-year-olds who competently paddle and maneuver canoes on the wide river. She observes young boys nimbly catching and handling enormous catfish (Figure 28). And then she is struck by the painful contrast between the children’s mastery of their natural surroundings and the great discomfort and incompetence they display in the classroom. She summarizes the dilemma as ‘learning to sit still.’”

## The demographic transition

400 years ago, a change began to happen in the Netherlands:

“In the seventeenth century, foreigners were already recording their astonishment at the laxity of Dutch parents ... they preferred to close their eyes to the faults of their children, and they refused to use corporal punishment ... foreigners remarked on something else: since the sixteenth century, most Dutch children – girls as well as boys – had been going to school. (Kloek 2003: 53)”

“John Locke – exiled to Holland in 1685–1688 – was profoundly influenced by what he saw. His treatise on childrearing, published in 1693, brought Dutch ideas on childcare to England (Locke 1693/1994). At the end of the eighteenth century, the Quakers also embraced population control and used various means to reduce their fertility. “The drop in the birth rate also reflected ... a rejection of the view that women were chattels who should devote their adult lives to an endless cycle of pregnancy and childbirth” (Mintz 2004: 78).”

Dutch paintings of this era are no longer only stiff portraits, but depict families enjoying time together (though I’m not sure how much the cat is enjoying this experience.)



*“Teaching a cat to read”, Jan Steen, 1660s. Note the young teacher holds a switch - this is how lessons worked even in the relaxed Netherlands.*

In developing countries, traditional methods of spacing births may be discouraged, resulting in a baby boom:

"For example, from Malaita Island in the South Pacific, traditional Kwara'ae practice was to keep men separated from their nursing wives for at least a year. However, the "abolition of the tabu system and the ascendancy of Christianity has meant that ... ritual separation [is] no longer practiced" (Gegeo and Watson-Gegeo 1985: 240-241). As a result, fertility has jumped and families with ten to thirteen children are not uncommon."

Western intervention has addressed one aspect of population but not another: "The agencies that intervened to reduce infant mortality were not as ready with contraception and family-planning interventions, and the result has been masses of humanity living on the ragged edge of poverty."

Even where it's available, people may not be interested in birth control, despite the practical difficulties of raising lots of children. In Burkina Faso:

"There are no perceived disadvantages in having lots of children. Children are never seen as a drain on resources. The availability of food is believed to be purely a product of the God-given fortune of the child, and nothing to do with the level of resources available within the household or the number of mouths to feed [because] 'every child is born with its own luck.' (Hampshire 2001: 115)"

The author gets editorial at times, quipping "The rich get richer, and the poor get lots of sickly children."

"Unfortunately, the ubiquity of infant death along with well-established coping mechanisms inures people to a phenomenon that, given the state of medical knowledge and a pharmacopeia adequate to the task, shouldn't be happening. The wastage of young human life and the debilitating impact this has on mothers are staggering and cannot possibly be justified. And, in the West, we remain largely oblivious of the problem of child malnutrition and death in the Third World until it reaches such proportions that the story becomes newsworthy."

# Inner Alignment: Explain like I'm 12 Edition

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

(This is an unofficial explanation of Inner Alignment based on the [Miri paper Risks from Learned Optimization in Advanced Machine Learning Systems](#) (which is almost identical to the [LW sequence](#)) and the [Future of Life podcast](#) with Evan Hubinger ([Miri/LW](#)). It's meant for anyone who found the sequence too long/challenging/technical to read.)

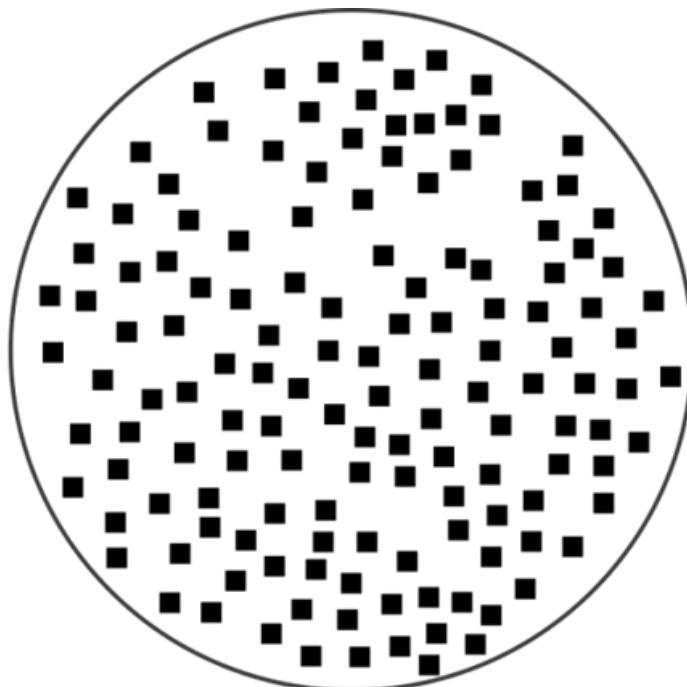
Note that **bold and italics** means "this is a new term I'm introducing," whereas underline and *italics* is used for emphasis.

## What is Inner Alignment?

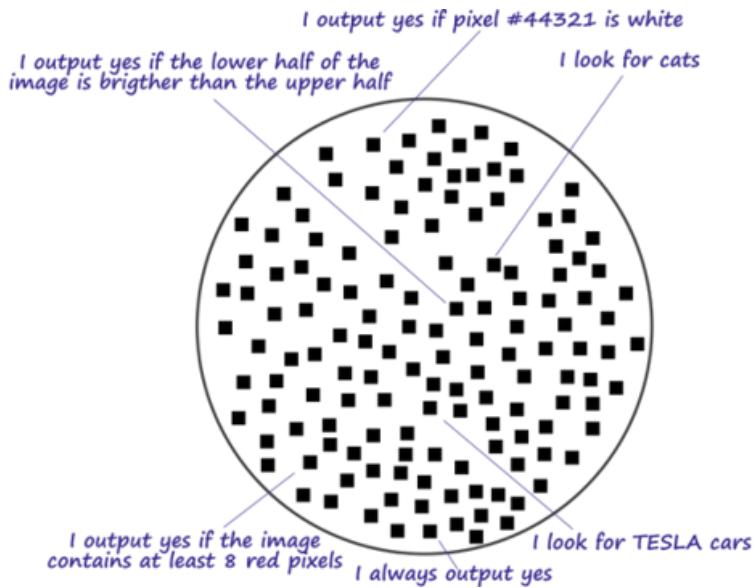
Let's start with an abridged guide to how Deep Learning works:

1. Choose a problem
2. Decide on a space of possible solutions
3. Find a good solution from that space

If the problem is "find a tool that can look at any image and decide whether or not it contains a cat," then each conceivable set of rules for answering this question (formally, each function from the set of all pixels to the set {yes, no}) defines one solution. We call each such solution a **model**. The space of possible models is depicted below.



Since that's *all* possible models, most of them are utter nonsense.

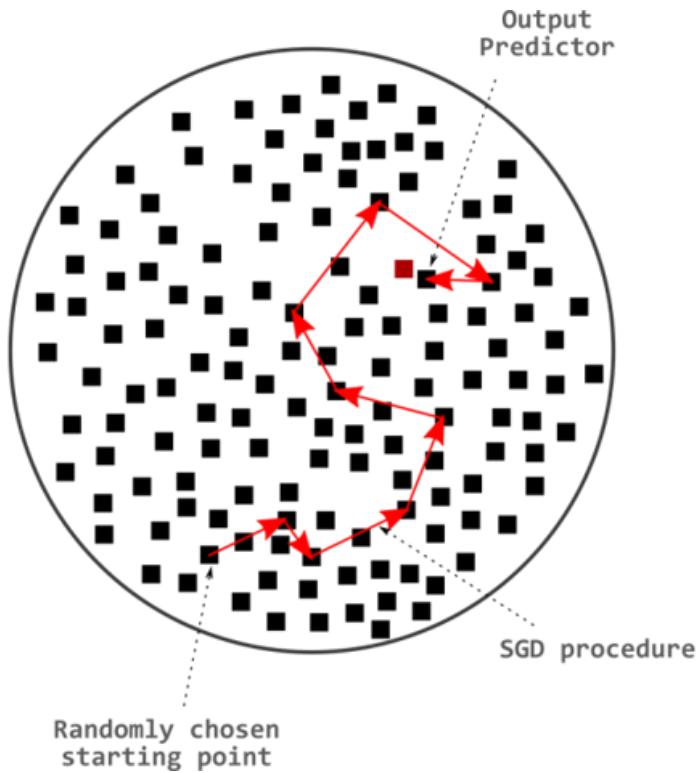


Pick a random one, and you're as likely to end up with a car-recognizer than a cat-recognizer – but far more likely with an algorithm that does nothing we can interpret. Note that even the examples I annotated aren't typical – most models would be more complex while still doing nothing related to cats. Nonetheless, somewhere in there is a model that would do a decent job on our problem. In the above, that's the one that says, "I look for cats."

How does ML find such a model? One way that does *not* work is trying out all of them.

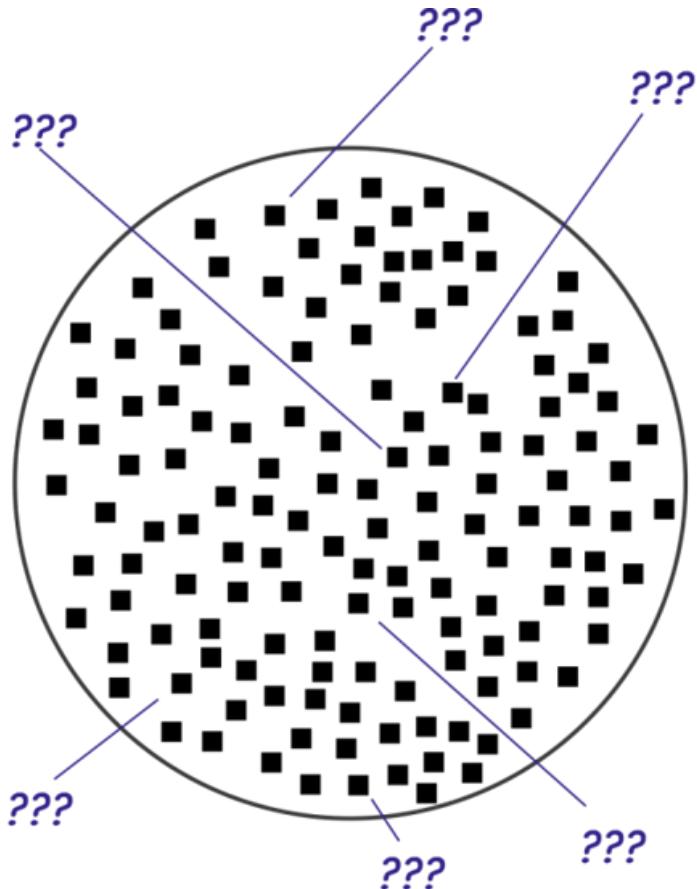
That's because the space is too large: it might contain over  $10^{1000000}$  candidates.

Instead, there's this thing called **Stochastic Gradient Descent (SGD)**. Here's how it works:



SGD begins with some (probably terrible) model and then proceeds in steps. In each step, it switches to another model that is "close" and hopefully a little better. Eventually, it stops and outputs the most recent model.<sup>[1]</sup> Note that, in the example above, we don't end up with the perfect cat-recognizer (the red box) but with something close to it - perhaps a model that looks for cats but has some unintended quirks. SGD does generally not guarantee optimality.

The speech bubbles where the models explain what they're doing are annotations for the reader. From the perspective of the programmer, it looks like this:



The programmer has no idea what the models are doing. Each model is just a black box.<sup>[2]</sup>

A necessary component for SGD is the ability to measure a model's performance, *but this happens while treating them as black boxes*. In the cat example, assume the programmer has a bunch of images that are accurately labeled as "contains cat" and "doesn't contain cat." (These images are called the **training data** and the setting is called **supervised learning**.) SGD tests how well each model does on these images and, in each step, chooses one that does better. In other settings, performance might be measured in different ways, but the principle remains the same.

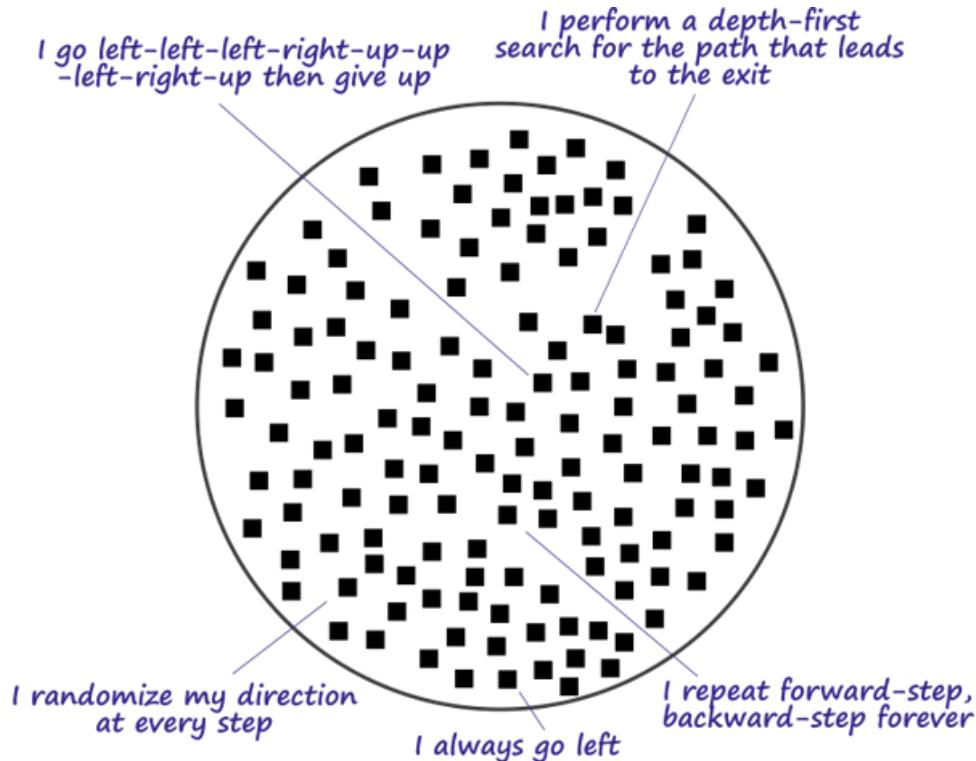
Now, suppose that the images we have happen to include only white cats. In this case, SGD might choose a model implementing the rule "output yes if there is something white and with four legs." The programmer would not notice anything strange – all she sees is that the model output by SGD does well on the training data.

In this setting, there is thus only a problem if our way of obtaining feedback is flawed. If it is perfect – if the pictures with cats are perfectly representative of what images-with-cats are like, and the pictures without cats are perfectly representative of what images-without-cats are like, then there isn't an issue. Conversely, if our images-with-cats are non-representative because all cats are white, the model SGD outputs might not be doing precisely what the programmer wanted. In Machine Learning slang, we would say that the training distribution is different from the distribution in deployment.

Is this Inner Alignment? Not quite. This is about a property called **distributional robustness**, and it's a well-known problem in Machine Learning. But it's close.

To explain Inner Alignment itself, we have to switch to a different setting. Suppose that, instead of trying to classify whether images contain cats, we are trying to train a model that solves mazes. That is, we want an algorithm that, given an arbitrary solvable maze, outputs a route from the Maze Entry to the Maze Exit.

As of before, our space of all possible models will consist primarily of nonsense solutions:

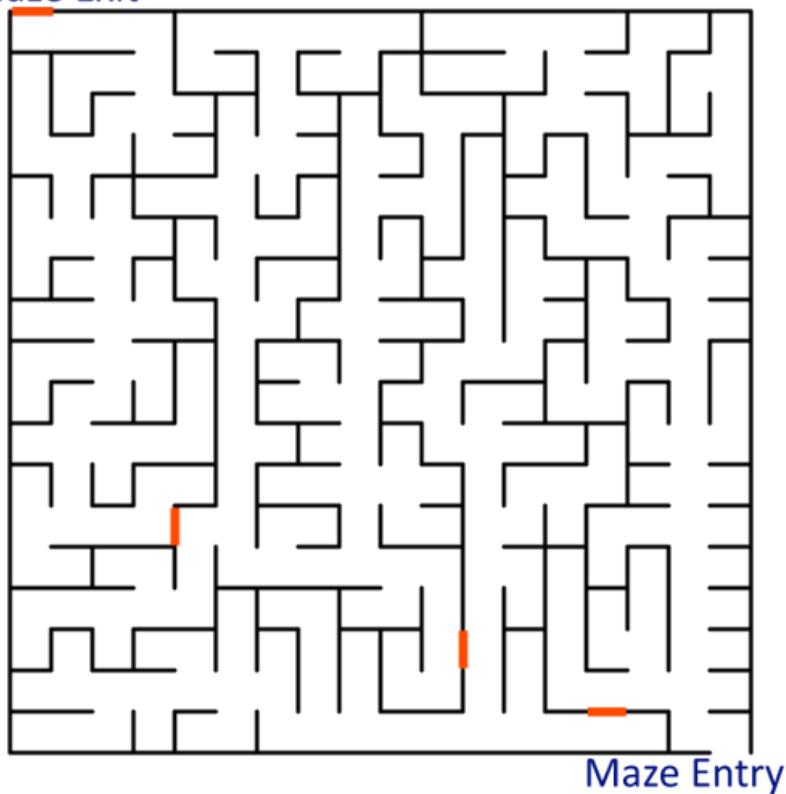


(If you don't know what depth-first search means: as far as mazes are concerned, it's simply the "always go along one wall" rule.)

The annotation "I perform depth-first search" means the model contains a formal algorithm that implements depth-first search, and analogously with the other annotations.

As with the previous example, we might apply SGD to this problem. In this case, the feedback mechanism would come from evaluating the model on test mazes. Now, suppose that all of the test mazes have this form,

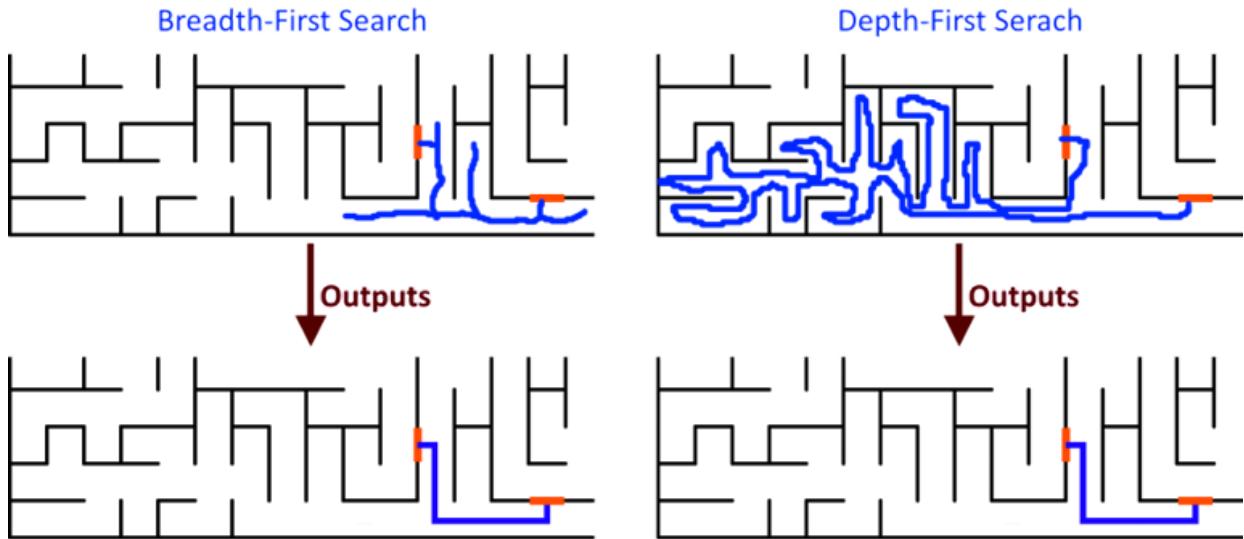
Maze Exit



Maze Entry

where the red areas represent doors. That is, all mazes are such that the shortest path leads through all of the red doors, and the exit is itself a red door.

Looking at this, you might hope that SGD finds the "depth-first" model. However, while that model would find the shortest path, it is not the best model. (Note that it first performs depth-first search and then, once it has found the right path, discards dead ends and outputs the shortest path only.) The alternative model with annotation "perform breadth-first search to find the next red door, repeat forever" would perform better. (Breadth-first means exploring all possible paths in parallel.) Both models always find the shortest path, but the red-door model would find it more quickly. In the maze above, it would save time by finding the path from the first to the second door without wasting time exploring the lower-left part of the maze.



*Both models output the same solution,  
but the right model takes longer*

Note that breadth-first search only outperforms depth-first search because it can truncate the fruitless paths after having reached the red door. Otherwise, it wouldn't know that the bottom-left part is fruitless until much later in the search.

As of before, all the programmer will see is that the left model performs better on the training data (the test mazes).

The qualitative difference to the cat picture example is that, in this case, *we can talk about the model as running an optimization process*\*\*.\*\* That is, the breadth-first search model does itself have an objective (go through red doors), and it tries to optimize for that in the sense that it searches for the shortest path that leads there. Similarly, the depth-first model is an optimization process with the objective "find exit of maze."

This is enough to define Inner Alignment, but to make sure the definition is the same that one reads elsewhere, let's first define two new terms.

- The **Base Objective** is the objective we use to evaluate models found by SGD. In the first example, it was "classify pictures correctly (i.e., say "contains cat" if it contains a cat and "doesn't contain cat" otherwise). In the second example, it was "find [a shortest path that solves mazes] as quickly as possible."
- In the cases where the model is running an optimization process, we call the model a **Mesa Optimizer**, and we call its objective the **Mesa Objective** (in the maze example, the mesa objective is "find shortest path through maze" for the depth-first model, and "repeatedly find shortest path to the next red door" for the breadth-first model).

With that said,

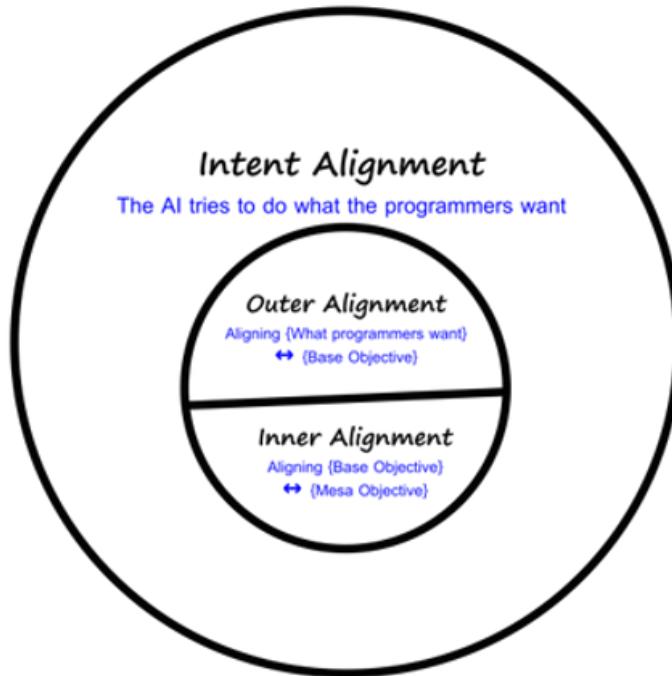
*Inner Alignment is the problem of aligning the Base Objective with the Mesa Objective.*

Some clarifying points:

- The red-door example is thoroughly contrived and would not happen in practice. It only aims to explain what Inner Alignment is, not why misalignment might be probable.
- You might wonder what the space of all models looks like. The typical answer is that the possible models are *sets of weights* for a [neural network](#). The problem exists insofar as some sets of weights implement specific search algorithms.
- As of before, the reason for the inner alignment failure was that our way of obtaining feedback was flawed (in MIL language: because there was distributional shift). (Although misalignment may also arise for other [very complicated reasons](#).)
- If the Base Objective and Mesa Objective are misaligned, this causes problems as soon as the model is deployed. In the second example, as soon as we take the model output by SGD and apply it to real mazes, it would still search for red doors. If those mazes don't contain red doors, or the red doors aren't always on paths to the exit, the model would perform poorly.

Here is the relevant Venn-Diagram. (Relative sizes don't mean anything.)

**Entire Alignment Problem**  
No consensus on the definition yet, but should imply  
that AI systems are safe to use.



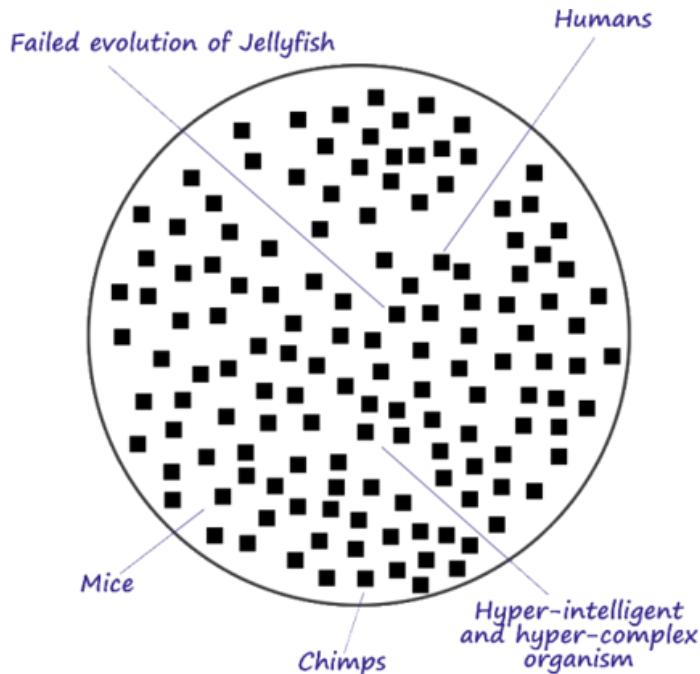
Note that  $\{\text{What AI tries to do}\} = \{\text{Mesa Objective}\}$  by definition.

Most classical discussion of AI alignment, including most of the book [Superintelligence](#), is about Outer Alignment. The classical examples where we assume the AI is optimized to cure cancer and then kills humans so that no-one can have cancer anymore is about a misalignment of  $\{\text{What Programmers want}\}$  and the  $\{\text{Base Objective}\}$ . (The Base Objective is  $\{\text{minimize the number of people who have cancer}\}$ , and while it's not clear what the programmers want, it's certainly not that.)

Admittedly, the inner alignment model is not maximally general. In this post, we've looked at black box search, where we have a parametrized model and do SGD to update the parameters. This describes most of what Machine Learning is up to in 2020, but it does not describe what the field did pre-2000 and, in the event of a paradigm shift similar to the deep learning revolution, it may not describe what the field looks like in the future. In the context of black box search, inner alignment is a well-defined property and Venn-Diagram a valid way of slicing up the problem, but there are people who expect that AGI will not be built that way.<sup>[3]</sup> There are even concrete proposals for safe AI where the concept doesn't apply. Evan Hubinger has since written [a follow-up post](#) about what he calls "training stories", which is meant to be "a general framework through which we can evaluate any proposal for building safe advanced AI".

## The Analogy to Evolution

Arguments about Inner Alignment often make reference to evolution. The reason is that evolution is an optimization process – it optimizes for inclusive genetic fitness. The space of all models is the space of all possible organisms.



Humans are certainly not the best model in this space – I've added the description on the bottom right to indicate that there are better models that haven't been found yet. However, humans are the best model that evolution has found so far.

As with the maze example, *humans do themselves run optimization processes*. Thus, we can call them/us Mesa Optimizes, and we can compare the Base Objective (the one evolution maximizes for) with the Mesa Objective (the one humans optimize for).

- Base Objective: maximize inclusive genetic fitness
- Mesa Objective: avoid pain, seek pleasure

(This is simplified – some humans optimize for other things, such as the well-being of all possible minds in the universe – but those are no closer to the Base Objective.)

We can see that humans are not aligned with the base objective of evolution. And it is easy to see why – the way Evan Hubinger put it is to imagine the counterfactual world where evolution did select inner-aligned models. In this world, a baby who stabs its toe has to compute how stabbing its toe affects its inclusive genetic fitness before knowing whether or not to repeat this behavior in the future. This would be computationally expensive, whereas the "avoid pain" objective immediately tells the baby that stabbing toe=bad, which is much cheaper and usually the correct answer. Thus, an unaligned model outperforms the hypothetical aligned model. Another interesting aspect is that the size of the misalignment (the difference between the Base Objective and the Mesa Objective) has widened over the last few millennia. In the ancestral environment, they were pretty close, but now, they are so far apart that we need to pay people to donate their sperm, which, according to the Base Objective, ought to be a highly desirable action.

Consequently, the analogy might be an argument for why Inner Misalignment is *probable* since it has occurred "naturally" in the biggest non-human-caused optimization process we know. However, the big caveat here is that *evolution does not implement Stochastic Gradient Descent*. Evolution navigates the model space by performing random mutations and then evaluating performance, which is fundamentally different (and a billion times less efficient) from modifying the model according to the expected derivative of the loss function, which is what SGD does. Thus, while the analogy works in most ways, it stops working as soon as one makes arguments that rely on properties of SGD other than that it optimizes the Base Objective.

## Deceptive Alignment

This is the abridged version of the [fourth part](#) of the sequence. I'm linking to it because this is probably the one where leaving out the technical details is the most problematic.

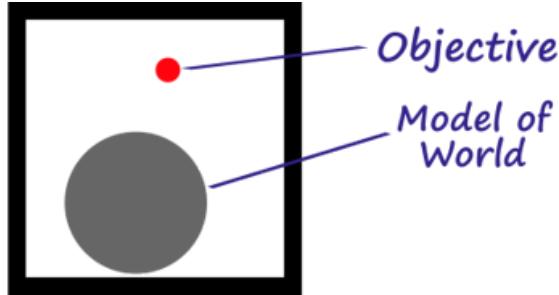
### The concept

In this section, we make the following assumptions:

- The learning task is hard, and therefore, models are very complex. Think of a question-answering system, rather than an image classifier.
- Instead of having a single learning process, we update a model over time.
- The learning process will select a Mesa Optimizer.
- The Base Objective is complicated, and the model won't get it right immediately (i.e., the model starts out not being inner-aligned).

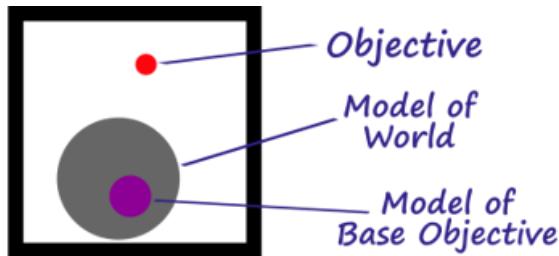
Since the model is sophisticated, we assume that it builds up a model of the world somehow. Think of GPT-3 (the language model that can write text): it clearly recognizes whether you're prompting it about Lord of the Rings or about politics. This shows that it has an internal model of these things, however flawed or incomplete.

Thus, if we look inside the model (which, again, the programmers cannot do), we have the following two components:



Recall that the model is a Mesa Optimizer by assumption, hence we know it has an objective. This (red blob) is the Mesa Objective.

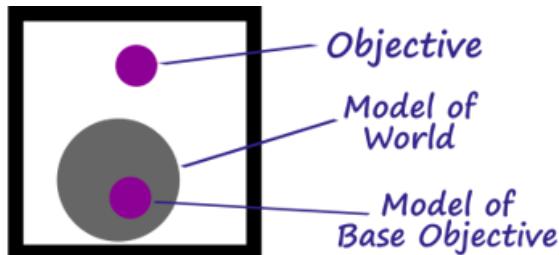
As its model of the world improves, it might eventually include a model of the Base Objective. Recall that the Base Objective is what SGD optimizes for.



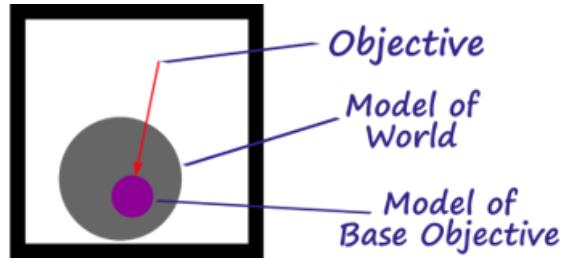
I've drawn the model of the Base Objective (purple blob) larger than the Mesa Objective since we assume the Base Objective is fairly complex.

SGD tries to make the model better, and if [the thing that the model optimizes for] becomes more similar to the Base Objective, the model does become better. Therefore, we speculate that the model will change such that this happens. We further speculate that there are three different ways this could happen, which I'll illustrate below.

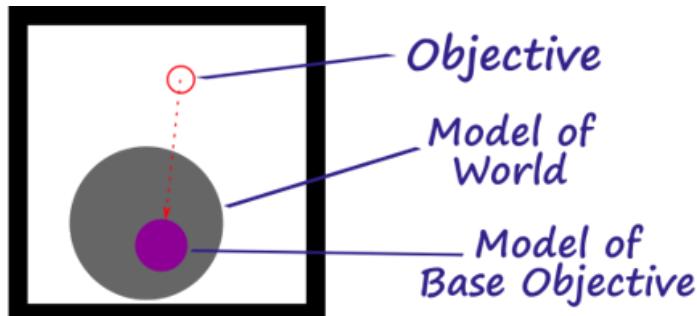
The first is called ***Internalization***. The model adopts the Base Objective outright.



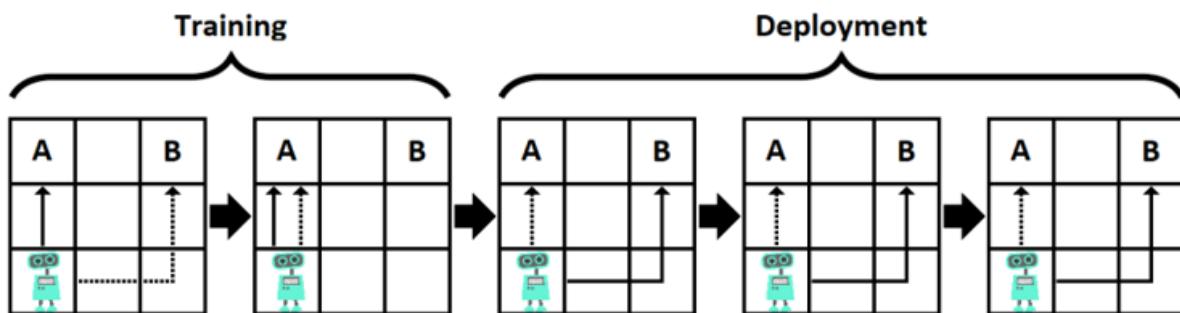
The second is called ***Proxy Alignment***. The Mesa Objective is no longer represented explicitly – instead, it is a pointer to the part of its world model that includes the Base Objective.



The third is called **Deceptive Alignment**. Here, the Mesa Objective remains different from the Base Objective, but the model chooses to optimize the Base Objective anyway because doing so makes the learning procedure think it is aligned. Thus, the dotted pointer does not indicate that the Mesa Objective is the Base Objective, nor that it points to the base objective, but that the model will optimize the Base Objective for instrumental reasons, i.e., as part of its strategy for optimizing the Mesa Objective.



For deceptive alignment to be possible, we require that (a) the model knows it is in a training process, and (b) it cares about its objective long-term. For deceptive alignment to become a problem, we further require that (c) the model expects that it eventually leaves the training process permanently. If all of those are given, the paper uses the following example to illustrate why a model might be deceptively aligned (i.e., might optimize the Base Objective even though it's not what it terminally wants).



In this case, the Base Objective is "go to A" and the Mesa Objective "go to B." If the model that wants to go to B is smart, it will go to A in the training process since that allows it to go to B during deployment, and there are three deployment episodes vs. only two training episodes. Thus, optimizing the Base Objective during training yields better results for the Mesa Objective (in the long-term) than optimizing for the Mesa Objective outright. This is deceptive alignment.

Buck Shlegeris ([Miri/LW](#)) has the following analogy illustrating the three cases. Suppose the Christian God wants to create humans that do what they want (god's utility function is the Base Objective). The three kinds of alignment have analogs in what kinds of humans do what god wants.

- For **internalization**, think of Jesus. He does what god wants because he has the same utility function as god.
- For **proxy alignment**, think of Martin Luther. He does what God wants because his utility function is "figure out what the bible wants and do that."
- For **deceptive alignment**, his example was [Blaise Pascal](#); however, I don't need that example because I used to have a friend in middle and high school who explicitly wanted to do what the bible said as he wanted to go to heaven. He openly admitted (to me, anyway) to have terminal goals that contradict what the bible says, but made genuine efforts to adhere to the bible anyway. I'm very confident that this type of person is more common than Martin Luthers or Jesuses.

One purpose of this analogy is to illustrate why deceptive models might be more probable than what intuition might claim. Deceptive alignment sounds like a weird, specific kind of thing, but the analogy suggests that there are, in fact, many more deceptively aligned models out there than proxy aligned models or models aligned via internalization. Only one utility function is the same as that of the Christian god, but many utility functions place very high utility on "don't go to hell." Similarly, there is a range of possible Mesa Objectives that allow for deceptive alignment, and all of those will optimize the Base Objective during training.

The above is just an analogy, but there are other reasons to expect that deception is likely. Specifically, I'll cover two further arguments why the alternatives are "difficult" to achieve and hence might be less likely.

## Why internalization might be difficult

If a model is heavily optimized for performance on a complex training distribution, we expect it to have a complex world model. This is already true for GPT-3, which requires explicit models of millions of human concepts (money, salt shakers, the Lord of the Rings movies, you name it) to predict the subsequent word on arbitrary inputs.

If the task is sufficiently hard, we may further expect this world model includes a model of the Base Objective. Not for nefarious reasons, but simply as one more thing that slightly increases performance on the training distribution.

If the above is true at the point where the model first switches from a bunch of heuristics to an explicit optimization target, Internalization would require rebuilding [the model of the Base Objective] in full, whereas proxy alignment or deception only require building a pointer to the [model of the Base objective] that *already exists* in the world model. Since rebuilding the entire model is likely much harder than installing a pointer (which can typically be represented by a single number), SGD is unlikely to go down that path.

In the analogy to evolution (which is still just an analogy and might not work in this case), it seems intuitive that programming a detailed model of "parents" into our DNA is harder for evolution than implementing love for parents via reference to our world model.

## Why proxy alignment might be difficult

The argument in the previous subsection does *not* address the difference between deceptive and proxy alignment as neither of them needs to represent a complex objective explicitly.

Instead, the argument made by the paper is one of **robustness**. It goes like so:

- If a model is proxy-aligned, but the pointer is wrong, this can cause the model's performance to degrade.
- Even if the model's world model (gray blob) improves to the point where it recognizes this failure, it will not correct the pointer because the pointer's current target is the Mesa Objective.
- SGD can update the pointer, but this isn't easy to get right.
- Conversely, a deceptive model will immediately adjust its behavior if it learns that the pointer was flawed as doing so is helpful for its objective. (This is so because its strategy is "hide my true objective for now, such that it won't get modified and I can pursue it in deployment.")

Note that this does not exhaust all arguments that have been made in the paper.

## Miscellaneous

It seems to me that one under-appreciated aspect of Inner Alignment is that, even if one had the one-true-utility-function-that-is-all-you-need-to-program-into-AI, this would not, in fact, solve the alignment problem, nor even the intent-alignment part. It would merely solve outer alignment (provided the utility function can be formalized). If we do SGD based on the one true utility function, this could still lead to a mesa optimized that wants something else.

Another interesting point is that the plausibility of internalization (i.e., of a model representing the Base Objective explicitly) does not solely depend on the complexity of the objective. For example, evolution's objective of "maximize inclusive genetic fitness" is quite simple, but it is still not represented explicitly because *figuring out how actions affect the objective is computationally hard*. Thus, {probability of Mesa Optimizer adopting an objective} is at least dependent on {complexity of objective} as well as {difficulty of assessing how actions impact objective}.

- 
1. In practice, one often runs SGD multiple times with different initializations and uses the best result. Also, the output of SGD may be a linear combination of all models on the way rather than just the final model. [←](#)
  2. However, there are efforts to create transparency tools to look into models. Such tools might be helpful if they become really good. Some of the [proposals for building safe advanced AI](#) explicitly include transparency tools [←](#)
  3. If an AGI does contain more hand-coded parts, the picture gets more complicated. E.g., if a system is logically separated into a bunch of components, inner alignment may apply to some of the components but not others. It may

even apply to parts of biological systems, see e.g., Steven Byrne's [Inner Alignment in the brain](#). ↵

# Radical Probabilism

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

*This is an expanded version of [my talk](#). I assume a high degree of familiarity with Bayesian probability theory.*

[Toward a New Technical Explanation of Technical Explanation](#) -- an attempt to convey the practical implications of logical induction -- was one of my most-appreciated posts, but I don't really get the feeling that very many people have received the update. Granted, that post was speculative, sketching what a new technical explanation of technical explanation *might* look like. I think I can do a bit better now.

If the implied project of that post had really been completed, I would expect new practical probabilistic reasoning tools, explicitly violating Bayes' law. For example, we might expect:

- A new version of information theory.
  - An update to the "[prediction=compression](#)" maxim, either repairing it to incorporate the new cases, or explicitly denying it and providing a good intuitive account of why it was wrong.
  - A new account of concepts such as mutual information, allowing for the fact that variables have behavior over thinking time; for example, variables may initially be very correlated, but lose correlation as our picture of each variable becomes more detailed.
- New ways of thinking about epistemology.
  - One thing that my post did manage to do was to spell out the importance of "making advanced predictions", a facet of epistemology which Bayesian thinking does not do justice to.
  - However, I left aspects of the problem of old evidence open, rather than giving a complete way to think about it.
- New probabilistic structures.
  - Bayesian Networks are one really nice way to capture the structure of probability distributions, making them much easier to reason about. Is there anything similar for the new, wider space of probabilistic reasoning which has been opened up?

Unfortunately, I still don't have any of those things to offer. The aim of this post is more humble. I think what I originally wrote was too ambitious for didactic purposes. Where the previous post aimed to communicate the insights of logical induction by sketching broad implications, I here aim to communicate the insights *in themselves*, focusing on the detailed differences between classical Bayesian reasoning and the new space of ways to reason.

Rather than talking about logical induction directly, I'm mainly going to explain things in terms of a very similar philosophy which Richard Jeffrey invented -- apparently starting with his PhD dissertation in the 50s, although I'm unable to get my hands on it or other early references to see how fleshed-out the view was at that point. He called this philosophy **radical probabilism**. Unlike logical induction, radical probabilism appears not to have any roots in worries about logical uncertainty or bounded rationality. Instead it appears to be motivated simply by a desire to generalize, and a refusal to accept unjustified assumptions. Nonetheless, it carries most of the same insights.

Radical Probabilism has not been very concerned with computational issues, and so constructing an actual algorithm (like the logical induction algorithm) has not been a focus. (However, there have been some developments -- see historical notes at the end.) This could be seen as a weakness. However, for the purpose of communicating the core insights, I think this is a strength -- there are fewer technical details to communicate.

A terminological note: I will use "radical probabilism" to refer to the new theory of rationality (treating logical induction as merely a specific way to flesh out Jeffrey's theory). I'm more conflicted about how to refer to the older theory. I'm tempted to just use the term "Bayesian", implying that the new theory is non-Bayesian -- this highlights its rejection of Bayesian updates. However, radical probabilism is Bayesian in the most important sense. Bayesianism is not about Bayes' Law. Bayesianism is, at core, about the subjectivist interpretation of probability. Radical probabilism is, if anything, *much more* subjectivist.

However, this choice of terminology makes for a confusion which readers (and myself) will have to carefully avoid: confusion between Bayesian probability theory and Bayesian updates. The way I'm using the term, a Bayesian need not endorse Bayesian updates.

In any case, I'll default to Jeffrey's term for the opposing viewpoint: **dogmatic probabilism**. (I will occasionally fall into calling it "classical Bayesianism" or similar.)

## What Is Dogmatic Probabilism?

Dogmatic Probabilism is the doctrine that the conditional probability  $P(B|A)$  is *also* how we update probabilistic beliefs: any rational change in beliefs should be explained by a Bayesian update.

We can unpack this a little:

1. (**Dynamic Belief:**) A rational agent is understood to have different beliefs over time -- call these  $P_1, P_2, P_3, \dots$
2. (**Static Rationality:**) At any one time, a rational agent's beliefs  $P_n$  are probabilistically coherent (obey the Kolmogorov axioms, or a similar axiomatization of probability theory).
3. (**Empiricism:**) Reasons for changing beliefs *across* time are *given entirely by observations* -- that is, propositions which the agent learns.
4. (**Dogmatism of Perception:**) Observations are believed with probability one, once learned.
5. (**Rigidity:**) Upon observing a proposition  $A$ , conditional probabilities  $P(B|A)$  are unmodified.

The assumptions minus *empiricism* imply that an update on observing  $A$  is a Bayesian update: if we start with  $P_n$  and update on  $A$  to get  $P_{n+1}$ , then  $P_{n+1}(A)$  must equal 1, and  $P_{n+1}(B|A) = P_n(B|A)$ . So we must have  $P_{n+1}(B) = P_n(B|A)$ . Then, *empiricism* says that this is the *only* kind of update we can possibly have.

## What Is Radical Probabilism?

Radical probabilism accepts assumptions #1 and #2, but rejects the rest. (Logical Induction need not follow axiom #2, either, since beliefs at any given time only approximately follow the probability laws -- however, it's not necessary to discuss this complication here. Jeffrey's philosophy did not attempt to tackle such things.)

Jeffrey seemed uncomfortable with updating to 100% on anything, making *dogmatism of perception* untenable. A similar view [is already popular on LessWrong](#), but it seems that no

one here took the implication and denied Bayesian updates as a result. (Bayesian updates [have been questioned for other reasons](#), of course.) This is a bit of an embarrassment. But fans of Bayesian updates reading this are more likely to accept that zero and one are probabilities, rather than give up Bayes.

Fortunately, this isn't actually the crux. Radical probabilism is a pure generalization of orthodox Bayesianism; you can have zero and one as probabilities, and still be a radical probabilist. The real fun begins not with the rejection of *dogmatism of perception*, but with the rejection of *rigidity* and *empiricism*.

This gives us a view in which a rational update from  $P_n$  to  $P_{n+1}$  can be almost anything. (You still can't update from  $P_n(A) = 0$  to  $P_{n+1}(A) > 0$ .) Simply put, *you are allowed to change your mind*. This doesn't make you irrational.

Yet, there are still *some* rationality constraints. In fact, we can say a lot about how rational agents think in this model. In place of assumptions #3-#5, we assume *rational agents cannot be Dutch Booked*.

## Radical Probabilism and Dutch Books

# Rejecting the Dutch Book for Bayesian Updates

At this point, if you're familiar with the philosophy of probability theory, you might be thinking: wait a minute, isn't there a Dutch Book argument for Bayesian updates? If radical probabilism accepts the validity of Dutch Book arguments, shouldn't it thereby be forced into Bayesian updates?

No!

As it turns out, there is a major flaw in the Dutch Book for Bayesian updates. The argument *assumes that the bookie knows how the agent will update*. (I encourage the interested reader to [read the SEP section on diachronic Dutch Book arguments](#) for details.) Normally, a Dutch Book argument requires the bookie to be *ignorant*. It's no surprise if a bookie can take our lunch money by getting us to agree to bets *when the bookie knows something we don't know*. So what's actually established by these arguments is: ***if you know how you're going to update, then your update had better be Bayesian.***

Actually, that's not quite right: the argument for Bayesian updates also still assumes *dogmatism of perception*. If we relax that assumption, all we can really argue for is *rigidity*: ***if you know how you are going to update, then your update had better be rigid.***

This leads to a generalized update rule, called **Jeffrey updating** (or Jeffrey conditioning).

## Generalized Updates

Jeffrey updates keep the rigidity assumption, but reject *dogmatism of perception*. So, we're changing the probability of some sentence A to  $P(A) = c$ , without changing any  $P(B|A)$ .

There's only one way to do this:

$$P_{n+1}(B) = c \cdot P_n(B|A) + (1 - c) \cdot P_n(B|\neg A)$$

In other words, the Jeffrey update interpolates linearly between the Bayesian update on A and the Bayesian update on  $\neg A$ . This generalizes Bayesian updates to allow for uncertain evidence: we're not sure we just saw someone duck behind the corner, but we're 40% sure.

If this way of updating seems a bit arbitrary to you, Jeffrey would agree. It offers only a small generalization of Bayes. Jeffrey wants to open up much broader space:

	Dogmatic	Not Dogmatic
Rigid	Bayes Update	Jeffrey Update
Not Rigid	anything can happen so long as $P(\text{observation}) = 1$	anything can happen

Classifying updates by assumptions made.

As I've already said, the rigidity assumption can only be justified *if the agent knows how it will update*. Philosophers like to say the agent *has a plan* for updating: "If I saw a UFO land in my yard and little green men come out, I would believe I was hallucinating." This is something we've worked out ahead of time.

A non-rigid update, on the other hand, means you don't know how you'd react: "If I saw a convincing proof of  $P=NP$ , I wouldn't know what to think. I'd have to consider it carefully." I'll call non-rigid updates **fluid updates**.

For me, fluid updates are primarily about *having longer to think, and reaching better conclusions as a result*. That's because my main motivation for accepting a radical-probabilist view is logical uncertainty. Without such a motivation, I can't really imagine being very interested. I boggle at the fact that Jeffrey arrived at this view without such a motivation.

**Dogmatic Probabilist:** All I can say is: why??

**Richard Jeffrey:** I've explained to you how the Dutch Book for Bayesian updates fails. What more do you want? My view is simply what you get when you remove the faulty assumptions and keep the rest.

**Dogmatic Probabilist (DP):** I understand that, but why should anyone be interested in this theory? OK, sure, I CAN make Jeffrey updates without getting Dutch Booked. But why ever would I? If I see a cloth in dim lighting, and update to 80% confident the cloth is red, I update in that way **because of the evidence which I've seen, which is itself fully confident**. How could it be any other way?

**RJ:** Tell me one piece of information you're absolutely certain of in such a situation.

**DP:** I'm certain I had that experience, of looking at the cloth.

**RJ:** Surely you aren't 100% sure you were looking at cloth. It's merely very probable.

**DP:** Fine then. The experience of looking at ... what I was looking at.

**RJ:** I'll grant you that tautologies have probability one.

**DP:** It's not a tautology... it's the fact that I had an experience, rather than none!

**RJ:** OK, but you are trying to defend the position that **there is some observation, which you condition on, which explains your 80% confidence the cloth is red**. Conditioning on "I had an experience, rather than none" won't do that. What proposition are you confident in, which explains your less-confident updates?

**DP:** The photons hitting my retinas, which I directly experience.

**RJ:** Surely not. You don't have any detailed knowledge of that.

**DP:** OK, fine, the individual rods and cones.

**RJ:** I doubt that. Within the retina, before any message gets sent to the brain, these get put through an opponent process which sharpens the contrast and colors. You're not perceiving rods and cones directly, but rather a probabilistic guess at light conditions based on rod and cone activation.

**DP:** The output of that process, then.

**RJ:** Again I doubt it. You're engaging in **inner-outer hocus pocus**.\* There is no clean dividing line before which a signal is external, and after which that signal has been "observed". The optic nerve is a noisy channel, warping the signal. And the output of the optic nerve itself gets processed at V1, so the rest of your visual processing doesn't get direct access to it, but rather a processed version of the information. And all this processing is noisy. Nowhere is anything certain. Everything is a guess. If, anywhere in the brain, there were a sharp 100% observation, then the nerves carrying that signal to other parts of the brain would rapidly turn it into a 99% observation, or a 90% observation...

**DP:** I begin to suspect you are trying to describe human fallibility rather than ideal rationality.

**RJ:** Not so! I'm describing how to rationally deal with uncertain observations. The source of this uncertainty could be anything. I'm merely giving human examples to establish that the theory has practical interest for humans. The theory itself only throws out unnecessary assumptions from the usual theory of rationality -- as we've already discussed.

**DP:** (sigh...) OK. I'm still never going to design an artificial intelligence to have uncertain observations. It just doesn't seem like something you do on purpose. But let's grant, provisionally, that rational agents could do so and still be called rational.

**RJ:** Great.

**DP:** So what's this about giving up rigidity??

**RJ:** It's the same story: it's just another assumption we don't need.

**DP:** Right, but then how do we update?

**RJ:** However we want.

**DP:** Right, but how? I want a constructive story for where my updates come from.

**RJ:** Well, if you precommit to update in a predictable fashion, you'll be Dutch-Bookable unless it's a rigid fashion.

**DP:** So you admit it! Updates need to be rigid!

**RJ:** By no means!

**DP:** But updates need to come from somewhere. Whether you know it or not, there's some mechanism in your brain which produces the updates.

**RJ:** Whether you know it or not is a critical factor. Updates you can't anticipate need not be Bayesian.

**DP:** Right, but... the point of epistemology is to give guidance about forming rational beliefs. So you should provide some formula for updating. But any formula is predictable. So a formula has to satisfy the rigidity condition. So it's got to be a Bayesian update, or at least a Jeffrey update. Right?

**RJ:** I see the confusion. But epistemology does not have to reduce things to a strict formula in order to provide useful advice. Radical probabilism can still say many useful things. Indeed, I think it's **more** useful, since it's closer to real human experience. Humans can't always account for **why** they change their minds. They've updated, but they can't give any account of where it came from.

**DP:** OK... but... I'm sure as hell never designing an artificial intelligence that way.

I hope you see what I mean. It's all terribly uninteresting to a typical Bayesian, especially with the design of artificial agents in mind. Why consider uncertainty about evidence? Why study updates which don't obey any concrete update rules? What would it even mean for an artificial intelligence to be designed with such updates?

In the light of logical uncertainty, however, it all becomes well-motivated. Updates are unpredictable not because there's no rule behind them -- nor because we lack knowledge of what exactly that rule is -- but because we can't always anticipate the results of computations before we finish running them. There are updates without corresponding evidence because we can think longer to reach better conclusions, and doing so does not reduce to Bayesian conditioning on the output of some computation. This doesn't imply uncertain evidence in exactly Jeffrey's sense, but it does give us cases where we update specific propositions to confidence levels other than 100%, and want to know how to move other beliefs in response. For example, we might apply a heuristic to determine that some number is very very likely to be prime, and update on this information.

Still, I'm very impressed with Jeffrey for reaching so many of the right conclusions without this motivation.

## Other Rationality Properties

So far, I've emphasized that fluid updates "can be almost anything". This makes it sound as if there are essentially no rationality constraints at all! However, this is far from true. We can establish some very important properties via Dutch Book.

### Convergence

No *single* update can be condemned as irrational. However, if you keep changing your mind again and again without ever settling down, *that* is irrational. Rational beliefs are required to eventually move less and less, converging to a single value.

*Proof:* If there exists a point  $p$  which your beliefs forever oscillate around (that is, your belief falls above  $p + c$  infinitely often, and falls below  $p - c$  infinitely often, for some  $c > 0$ ) then a bookie can make money off of you as follows: when your belief is below  $p - c$ , the bookie makes a bet in favor of the proposition in question, at  $p : (1 - p)$  odds. When your belief is above  $p + c$ , the bookie offers to cancel that bet for a small fee. The bookie earns the fee with certainty, since your beliefs are sure to swing down eventually (allowing the bet to be placed) and are sure to swing up some time after that (allowing the fee to be collected). What's more, the bookie can do this again and again and again, turning you into a money pump.

If there exists no such  $p$ , then your beliefs must converge to some value.  $\square$

Caveat: this is the proof in the context of logical induction. There are other ways to establish convergence in other formalizations of radical probabilism.

In any case, this is *really important*. This isn't just a nice rationality property. *It's a nice rationality property which dogmatic probabilists don't have.* Lack of a convergence guarantee is **one of the main criticisms Frequentists make of Bayesian updates.** And it's a good critique!

Consider a simple coin-tossing scenario, in which we have two hypotheses:  $h_{\frac{1}{2}}$  posits that the probability of heads is  $\frac{1}{2}$ , and  $h_{\frac{2}{3}}$  posits that the probability of heads is  $\frac{2}{3}$ . The prior places probability  $\frac{1}{2}$  on both of these hypotheses. The only problem is that the true coin probability is  $\frac{2}{3}$ . What happens? The probabilities  $P(h_{\frac{1}{2}})$  and  $P(h_{\frac{2}{3}})$  will oscillate forever without converging.

*Proof:* The quantity heads – tails will take a random walk as we keep flipping the fair coin. A random walk returns to zero infinitely often (a phenomenon known as [gambler's ruin](#)). At each such point, evidence is evenly balanced between the two hypotheses, so we've returned to the prior. Then, the next flip is either heads or tails. This results in a probability of  $\frac{1}{3}$  for one of the hypotheses, and  $\frac{2}{3}$  for the other. This sequence of events happens infinitely often, so  $P(h_{\frac{1}{2}})$  and  $P(h_{\frac{2}{3}})$  keep experiencing changes of size at least  $\frac{1}{3}$ , never settling down.  $\square$

Now, the objection to Bayesian updates here isn't just that oscillating forever looks irrational. Bayesian updates are *supposed to help us predict the data well*; in particular, you might think they're *supposed to help us minimize log-loss*. But here, we would be doing much better if beliefs would converge toward  $P(h_{\frac{1}{2}}) = P(h_{\frac{2}{3}}) = \frac{2}{3}$ . The problem is, Bayes takes each new bit of evidence *just as seriously* as the last. Really, though, a rational agent in this situation should be saying: "Ugh, this again! If I send my probability up, it'll come crashing

right back down some time later. I should skip all the hassle and keep my probability close to where it is."

In other words, a rational agent should be looking out for Dutch Books against itself, including the non-convergence Dutch Book. Its probabilities should be adjusted to avoid such Dutch Books.

**DP:** Why should I be bothered by this example? If my prior is as you describe it, I assign **literally zero probability** to the world you describe -- I **know** the coin isn't fair. I'm fine with my inference procedure displaying pathological behavior in a universe I'm absolutely confident I'm not in.

**RJ:** So you're fine with an inference procedure which performs abysmally in the real world?

**DP:** What? Of course not.

**RJ:** But the real world cannot possibly be in your hypothesis space. It's too big. You can't explicitly write it down.

**DP:** Physicists seem to be making good progress.

**RJ:** Sure, but those aren't hypotheses which you can directly use to anticipate your experiences. They require too much computation. Anything that can fit in your head, can't be the real world.

**DP:** You're dealing with human frailty again.

**RJ:** On the contrary. Even idealized agents can't fit inside a universe they can perfectly predict. To see the contradiction, just let two of them play rock-paper-scissors with each other. Anything that can anticipate what you expect, and then do something else, can't be in your hypothesis space. But let me try a different angle of attack. Bayesianism is supposed to be the philosophy of subjective probability. Here, you're arguing as if the prior represented an objective fact about how the universe is. It isn't, and can't be.

**DP:** I'll deal with both of those points at once. I don't really need to assume that the **actual universe** is within my hypothesis space. Constructing a prior over a set of hypotheses guarantees you this: **if there is a best element in that class, you will converge to it.** In the coin-flip example, I don't have the objective universe in my set of hypotheses unless I can perfectly predict every coin-flip. But the subjective hypothesis which treats the coin as fair is the best of its kind. In the rock-paper-scissors example, rational players would similarly converge toward treating each other's moves as random, with  $\frac{1}{3}$  probability on each move.

**RJ:** Good. But you've set up the punchline for me: **if there is no best element, you lack a convergence guarantee.**

**DP:** But it seems as if good priors usually do have a best element. Using Laplace's rule of succession, I can predict coins of any bias without divergence.

**RJ:** What if the coin lands as follows: 5 heads in a row, then 25 tails, then 125 heads, and so on, each run lasting for the next power of five. Then you diverge again.

**DP:** Ok, sure... but if the coin flips might not be independent, then I should have hypotheses like that in my prior.

**RJ:** I could keep trying to give examples which break your prior, and you could keep trying to patch it. But we have agreed on the important thing: good priors should have the convergence property. At least you've agreed that this is a desirable property not always achieved by Bayes.

**DP:** Sure.

In the end, I'm not sure who would win the counterexample/patch game: it's quite possible that there general priors with convergence guarantees. [No computable prior has convergence guarantees for "sufficiently rich" observables](#) (ie, observables including logical combinations of observables). However, that's a theorem with a lot of caveats. In particular, Solomonoff Induction isn't computable, so might be immune to the critique. And we can certainly get rid of the problem by restricting the observables, EG by [conditioning on their sequential order rather than just their truth](#). Yet, [I suspect all such solutions will either be really dumb, or uncomputable](#).

So there's work to be done here.

But, in general (ie *without any special prior which does guarantee convergence for restricted observation models*), a Bayesian [relies on a realizability \(aka grain-of-truth\) assumption for convergence, as it does for some other nice properties](#). Radical probabilism demands these properties without such an assumption.

So much for technical details. Another point I want to make is that convergence points at a notion of "objectivity" for the radical probabilist. Although the individual updates a radical probabilist makes can go all over the place, the beliefs must eventually settle down to something. The goal of reasoning is to settle down to that answer as quickly as possible. Updates may appear arbitrary from the outside, but internally, they are always moving toward this goal.

This point is further emphasized by the next rationality property: conservation of expected evidence.

## Conservation of Expected Evidence

[The law of conservation of expected evidence](#) is a dearly beloved Bayesian principle. You'll be glad to hear that it survives unscathed:

$$P_n(X) = E_n P_m(X)$$

In the above,  $P_n(X)$  is your current belief in some proposition  $X$ ;  $P_m(X)$  is some future belief about  $X$  (so I'm assuming  $m > n$ ); and  $E_n$  is the expected value operator according to your current beliefs. So what the equation says is: your current beliefs equal your expected value of your future beliefs. This is just like the usual formulation of no-expected-net-update, except we no longer take the expectation *with respect to evidence*, since a non-Bayesian update may not be grounded in evidence.

*Proof:* Suppose  $P_n(X) \neq E_n P_m(X)$ . One of the two numbers is higher, and the other lower.

Suppose  $E_n P_m(X)$  is the lower number. Then a bookie can buy a certificate paying  $\$P_m(X)$  on day  $m$ ; we will willingly sell the bookie this for  $\$E_n P_m(X)$ . The bookie can also sell us a certificate paying  $\$1$  if  $X$ , for a price of  $\$P_n(X)$ . At time  $m$ , the bookie gains  $\$P_m(X)$  due to the first certificate. It can then buy the second certificate back from us for  $\$P_m(X)$ , using the

winnings. Overall, the bookie has now paid  $\$E_n P_m(X)$  to us, but we have paid the bookie  $\$P_n(X)$ , which we assumed was greater. So the bookie profits the difference.

If  $P_n(X)$  is the lower number instead, the same strategy works, reversing all buys and sells.  $\square$

The key idea here is that both a direct bet on  $X$  and a bet on  $P_m(X)$  will be worth  $P_m(x)$  later, so they'd better have the same price now, too.

I see this property as being even more important for a radical probabilist than it is for a dogmatic probabilist. For a dogmatic probabilist, it's a consequence of Bayesian conditional probability. For a radical probabilist, it's a basic condition on rational updates. With updates being so free to go in any direction, it's an important anchor-point.

Another name for this law is *the martingale property*. This is a property of many stochastic processes, such as Brownian motion. From [wikipedia](#):

In [probability theory](#), a **martingale** is a [sequence](#) of [random variables](#) (i.e., a [stochastic process](#)) for which, at a particular time, the [conditional expectation](#) of the next value in the sequence, given all prior values, is equal to the present value.

It's important that a sequence of rational beliefs have this property. Otherwise, future beliefs are different from current beliefs in a predictable way, and we would be better off updating ahead of time.

Actually, that's not immediately obvious, right? The bookie in the Dutch Book argument doesn't make money by updating to the future belief faster than the agent, but rather, by playing the agent's beliefs off of each other.

This leads me to a stronger property, which has the martingale property as an immediate consequence (**strong self trust**):

$$P_n(X | P_m(X) = y) = y$$

Again I'm assuming  $m > n$ . The idea here is supposed to be: *if you knew your own future belief, you would believe it already*. Furthermore, you believe  $X$  and  $P_m(X)$  are perfectly correlated: the only way you'd have high confidence in  $X$  would be if it were very probably true, and the only way you'd have low confidence would be for it to be very probably false.

I won't try to prove this one. In fact, be wary: this rationality condition is a bit too strong. The condition holds true in the radical-probabilism formalization of [Diachronic Coherence and Radical Probabilism by Brian Skyrms](#), so long as  $P_n(P_m(X) = y) > 0$  (see section 6 for statement and proof). However, [Logical Induction](#) argues persuasively that this condition is undesirable in specific cases, and replaces it with a slightly weaker condition (see section 4.12).

Nonetheless, for simplicity, I'll proceed as if *strong self trust* were precisely true.

At the end of the previous section, I promised that the current section would further illuminate my remark:

The goal of reasoning is to settle down to that answer as quickly as possible. Updates may appear arbitrary from the outside, but internally, they are always moving toward this goal.

The way radical probabilism allows *just about any change* when beliefs shift from  $P_n$  to  $P_{n+1}$  may make its updates seem irrational. How can the update be *anything*, and still be called rational? Doesn't that mean a radical probabilist is open to garbage updates?

No. A radical probabilist doesn't *subjectively* think all updates are equally rational. A radical probabilist *trusts the progression of their own thinking*, and also *does not yet know the outcome of their own thinking*; this is why I asserted earlier that a fluid update can be just about anything (barring the transformation of a zero into a positive probability). However, this does not mean that a radical probabilist would accept a psychedelic pill which arbitrarily modified their beliefs.

Suppose a radical probabilist has a sequence of beliefs  $P_1, P_2, P_3, P_4, \dots, P_n$ . If they thought hard for a while, they could update to  $P_{n+1}$ . On the other hand, if they took the psychedelic pill, their beliefs would be modified to become  $Q$ . The sequence would be abruptly disrupted, and go off the rails:  $P_1, P_2, P_3, \dots, P_n, Q, R, S, \dots$

The radical probabilist does not trust *whatever they believe next*. Rather, the radical probabilist has a concept of *virtuous epistemic process*, and is willing to believe the next output of such a process. Disruptions to the epistemic process do not get this sort of trust without reason. (For those familiar with *The Abolition of Man*, this concept is very reminiscent of his "Tao".)

On the other hand, a radical probabilist *could* trust a different process. One person,  $P$ , might trust that another person,  $Q$ , is better-informed about any subject:

$$P_n(X | Q_n(X) = y) = y$$

This says that  $P$  trusts  $Q$  on any subject if they've had the same amount of time to think. This leaves open the question of what  $P$  thinks if  $Q$  has had longer to think. In the extreme case, it might be that  $P$  thinks  $Q$  is better *no matter how long P has to think*:

$$\forall m, n P_m(X | Q_n(X) = y) = y$$

On the other hand,  $P$  and  $Q$  can both be perfectly rational by the standards of radical probabilism *and not trust each other at all*.  $P$  might not trust  $Q$ 's opinion no matter how long  $Q$  thinks.

(Note, however, that you *do* get eventual agreement on matters where good feedback is available -- much like in dogmatic Bayesianism, it's difficult for two Bayesians to disagree about *empirical predictions* for long.)

This means you can't necessarily replace one "virtuous epistemic process" with another.  $P_1, P_2, P_3, \dots$  and  $Q_1, Q_2, Q_3, \dots$  might both be perfectly rational by the standards of radical probabilism, and yet the disrupted sequence  $P_1, P_2, P_3, Q_4, Q_5, Q_6, \dots$  would not be, because  $P_3$  does not necessarily trust  $Q_4$  or subsequent  $Q$ s.

Realistically, we can be in this kind of position *and not even know what constitutes a virtuous reasoning process by our standards*. We generally think that we can "do philosophy" and reach better conclusions. But we don't have a clean specification of our own thinking process. We don't know exactly what counts as a virtuous continuation of our thinking vs a disruption.

This has some implications for AI alignment, but I won't try to spell them out here.

## Calibration

One more rationality property before we move on.

One could be forgiven for reading Eliezer's [A Technical Explanation of Technical Explanation](#) and coming to believe that Bayesian reasoners are calibrated. Eliezer goes so far as to suggest that we *define* probability in terms of calibration, so that *what it means* to say "90% probability" is that, in cases where you say 90%, the thing happens 9 out of 10 times.

However, the truth is that calibration is a neglected property in Bayesian probability theory. Bayesian updates do not help you learn to be calibrated, any more than they help your beliefs to be convergent.

We can make a sort of Dutch Book argument for calibration: if things happen 9-out-of-ten times when the agent says 80%, then a bookie can place bets with the agent at 85:15 odds and profit in the long run. (Note, however, that this is a bit different from typical Dutch Book arguments: it's a strategy in which the bookie risks some money, rather than just getting a sure gain. What I can say is that Logical Induction treats this as a valid Dutch Book, and so, we get a calibration property in that formalism. I'm not sure about other formalisations of Radical Probabilism.)

The intuition is similar to convergence: even lacking a hypothesis to explain it, a rational agent should eventually notice "hey, when I say 80%, the thing happens 90% of the time!". It can then improve its beliefs in future cases by adjusting upwards.

This illustrates "meta-probabilistic beliefs": a radical probabilist can have informed opinions *about the beliefs themselves*. By default, a classical Bayesian doesn't have beliefs-about-beliefs except as a result of learning about the world and reasoning about themselves as a part of the world, which is [problematic in the classical Bayesian formalism](#). It is possible to add second-order probabilities, third-order, etc. But calibration is a case which collapses all those levels, illustrating how the radical probabilist can handle all of this more naturally.

I'm struck by the way calibration *is something Bayesians obviously want*. The set of people who advocate applying Bayes Law and the set of people who look at calibration charts for their own probabilities has a *very significant overlap*. Yet, Bayes' Law does not give you calibration. It makes me feel like more people should have noticed this sooner and made a bigger deal about it.

# Bayes From a Distance

Before any more technical details about radical probabilism, I want to take a step back and give one intuition for what's going on here.

We can see radical probabilism as *what a dogmatic Bayesian looks like if you can't see all the details*.

## The Rationality of Acquaintances

Imagine you have a roommate who is perfectly rational in the dogmatic sense: this roommate has low-level observations which are 100% confident, and performs a perfect Bayesian update on those observations.

However, observing your roommate, you can't track all the details of this. You talk to your roommate about some important beliefs, but you can't track every little Bayesian update -- that would mean tracking every sensory stimulus.

From your perspective, your roommate has constantly shifting beliefs, which can't quite be accounted for. If you are particularly puzzled by a shift in belief, you can discuss reasons. "I updated against getting a cat because I observed a hairball in our neighbor's apartment." Yet, none of the evidence discussed is itself 100% confident -- it's at least a little bit removed from low-level sense-data, and at least a little uncertain.

Yet, this is not a big obstacle to viewing your roommate's beliefs as rational. You can evaluate these beliefs on their own merits.

I've heard this model called *Bayes-with-a-side-channel*. You have an agent updating via Bayes, but part of the evidence is hidden. You can't give a formula for changes in belief over time, but you can still assert that they'll follow conservation of expected evidence, and some other rationality conditions.

What Jeffrey proposes is that we allow these dynamics without necessarily positing a side-channel to explain the unpredictable updates. This has an anti-reductionist flavor to it: updates do not have to reduce to observations. But why should we be reductionist in that way? Why would subjective belief updates *need* to reduce to observations?

(Note that *Bayes-with-a-side-channel* does not imply conditions such as convergence and calibration; so, Jeffrey's theory of rationality is more demanding.)

## Wetware Bayes

Of course, Jeffrey would say that our relationship with ourselves is much like the roommate in my story. Our beliefs move around, and while we can often give some account of why, we can't give a full account in terms of things we've learned with 100% confidence. And it's not simply because we're a Bayesian reasoner who lacks introspective access to the low-level information. The nature of our wetware is such that there isn't really any place you can point to and say "this is a 100% known observation". Jeffrey would go on to point out that there's no clean dividing line between external and internal, so you can't really draw a boundary between external event and internal observation-of-that-event.

(I would remark that Jeffrey doesn't exactly give us a way to *handle* that problem; he just offers an abstraction which doesn't chafe on that aspect of reality so badly.)

Rather than imagining that there are perfect observations somewhere in the nervous system, we can instead imagine that a sensory stimulus exerts a kind of "evidential pressure" which can be less than 100%. These evidential pressures can also come from within the brain, as is the case with logical updates.

# But Where Do Updates Come From?

Dogmatic probabilism raises the all-important question "where do priors come from?" -- but once you answer that, everything else is supposed to be settled. There have been many debates about what constitutes a rational prior.

**Q. How can I find the priors for a problem?**

**A.** Many commonly used priors are listed in the *Handbook of Chemistry and Physics*.

**Q. Where do priors originally come from?**

**A.** Never ask that question.

**Q. Uh huh. Then where do scientists get their priors?**

**A.** Priors for scientific problems are established by annual vote of the AAAS. In recent years the vote has become fractious and controversial, with widespread acrimony, factional polarization, and several outright assassinations. This may be a front for infighting within the Bayes Council, or it may be that the disputants have too much spare time. No one is really sure.

**Q. I see. And where does everyone else get their priors?**

**A.** They download their priors from Kazaa.

**Q. What if the priors I want aren't available on Kazaa?**

**A.** There's a small, cluttered antique shop in a back alley of San Francisco's Chinatown. *Don't ask about the bronze rat.*

-- Eliezer Yudkowsky, [\*An Intuitive Explanation of Bayes' Theorem\*](#)

Radical probabilists put less emphasis on the prior, since a radical probabilist can effectively "decide to have a different prior" (updating their beliefs as if they'd swapped out one prior for another). However, they face a similarly large problem of where *updates* come from.

We are given a picture in which beliefs are like a small particle in a fluid, reacting to all sorts of forces (some strong and some weak). Its location gradually shifts as a result of Brownian motion. Presumably, the interesting work is being done behind the scenes, by whatever is generating these updates. Yet, Jeffrey's picture seems to mainly be about the dance of the particle, while the fluid around it remains a mystery.

A full answer to that question is beyond the scope of this post. (Logical Induction offers *one* fully detailed answer to that question.) However, I do want to make a few remarks on this problem.

- It might at first seem strange for beliefs to be so radically malleable to external pressures. But, actually, this is already the familiar Bayesian picture: everything happens due to externally-driven updates.
- Bayesian updates don't really answer the question of where updates come from, either. They take it as given that there are some "observations". Radical probabilism simply allows for a *more general* sort of feedback for learning.
- An orthodox probabilist might answer this challenge by saying something like: when we design an agent, we design sensors for it. These are connected in such a way as to feed in sensory observations. A radical probabilist can similarly say: when we design an agent, we get to decide what sort of feedback it uses to improve its beliefs.

The next section will give some practical, human examples of non-Bayesian updates.

## Virtual Evidence

Bayesian updates are path-independent: it does not matter in what order you apply them. If you first learn A and then learn B, your updated probability distribution is

$P_3(X) = P_2(X|B) = P_1(X|A \& B)$ . If you learn these facts the other way around, it's still

$P_3(X) = P_2(X|A) = P_1(X|A \& B)$ .

Jeffrey updates are path-dependent. Suppose my probability distribution is as follows:

A	$\neg A$
B	30% 20%
$\neg B$	20% 30%

I then apply the Jeffrey update  $P(B)=60\%$ :

A	$\neg A$
B	36% 24%
$\neg B$	16% 24%

Now I apply  $P(A)=60\%$ :

A	$\neg A$
B	41.54% 20%
$\neg B$	18.46% 20%

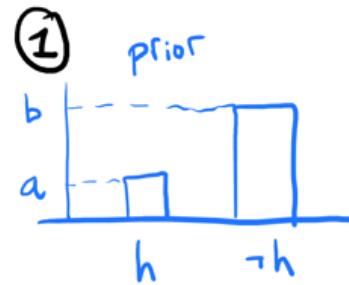
Since this is asymmetric, but the initial distribution was symmetric, obviously this would turn out differently if we had applied the Jeffrey updates in a different order.

Jeffrey considered this to be a bug -- although he seems fine with path-dependence under some circumstances, he used examples like the above to motivate a *different* way of handling uncertain evidence, which I'll call **virtual evidence**. (Judea Pearl strongly advocated virtual evidence over Jeffrey's rule near the beginning of Probabilistic Reasoning in Intelligent Systems (Section 2.2.2 and 2.3.3), in what can easily be read as a critique of Jeffrey's theory -- if one does not realize that Jeffrey is largely in agreement with Pearl. I thoroughly recommend Pearl's discussion of the details.)

Recall the basic anatomy of a Bayesian update:

$$P(h|e) = \frac{P(e|h)P(h)}{P(e)}$$

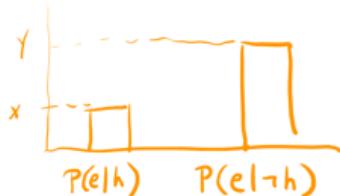
likelihood                      prior probability  
                                     normalizing factor



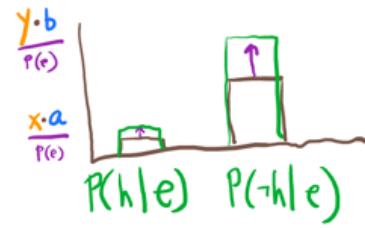
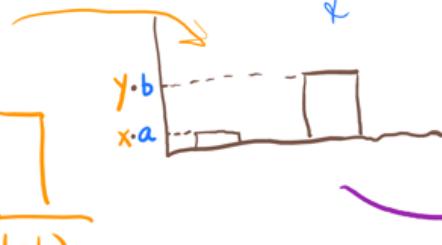
③ normalize

② multiply by likelihood function

Likelihood function



likelihood  $\times$  prior



The idea of virtual evidence is to use evidence 'e' which is not an event in our event space. We're just acting as if there were evidence 'e' which justifies our update. Terms such as  $P(e)$ ,  $P(e\&h)$ ,  $P(e|h)$ ,  $P(h|e)$ , and so on are not given the usual probabilistic interpretation; they just stand as a convenient notation for the update. **All we need to know is the likelihood function for the update.** We then multiply our probabilities by the likelihood function as usual, and normalize.  $P(e)$  is easy to find, since it's just whatever factor makes everything sum to one at the end. This is good, since it isn't clear what  $P(e)$  would mean for a virtual event.

Actually, we can simplify even further. All we *really* need to know is the likelihood *ratio*: the ratio between the two numbers in the likelihood function. (I will illustrate this with an example soon). However, it may sometimes be easier to find the whole likelihood function in practice.

Let's look at the path-dependence example again. As before, we start with:

A	$\neg A$
B	30% 20%
$\neg B$	20% 30%

I want to apply a Jeffrey update which makes  $P(B)=60\%$ . However, let's represent the update via virtual evidence this time. Currently,  $P(B)=50\%$ . To take it to 60%, we need to see virtual evidence with a 60:40 likelihood ratio, such as  $P(B|E)=60\%$ ,  $P(\neg B|E)=40\%$ . This gives us the same update as before:

A     $\neg A$

B	36%	24%
$\neg B$	16%	24%

(Note that we would have gotten the same result with a likelihood function of  $P(B|E)=3\%$ ,  $P(\neg B|E)=2\%$ , since 60:40 is the same as 3:2. That's what I meant when I said that only the ratio matters.)

But now we want to apply the same update to A as we did to B. So now we update on virtual evidence  $P(A|E)=60\%$ ,  $P(\neg A|E)=40\%$ . This gives us the following (approximately):

A	$\neg A$
B	43% 19%
$\neg B$	19% 19%

As you can see, the result is quite symmetric. In general, virtual evidence updates will be path-independent, because multiplication is commutative (and the normalization step of updating doesn't mess with this commutativity).

So, virtual evidence is a reformulation of Jeffrey updates with a lot of advantages:

- Unlike raw Jeffrey updates, virtual evidence is path-independent.
- You don't have to decide right away what you're updating *to*; you just have to decide the strength and direction of the update.
- I don't fully discuss this here, but Pearl argues persuasively that it's easier to tell when a virtual-evidence update is appropriate than when a Jeffrey update is appropriate.

Because of these features, virtual evidence is much more useful for *integrating information from multiple sources*.

## Integrating Expert Opinions

Suppose you have an ancient artefact. You want to know whether this artefact was made by ancient aliens. You have some friends who are also curious about ancient aliens, so you enlist their help.

You ask one friend who is a metallurgist. After performing experiments (the details of which you don't understand), the metallurgist isn't sure, but gives 80% that the tests would turn out that way if it were of terrestrial origin, and 20% for metals of non-terrestrial origin. (Let's pretend that ancient aliens would 100% use metals of non-Earth origin, and that ancient humans would 100% use Earth metals.)

You then ask a second friend, who is an anthropologist. The anthropologist uses cultural signs, identifying the style of the art and writing. Based on that information, the anthropologist estimates that it's half as likely to be of terrestrial origin as alien.

How do we integrate this information? According to Jeffrey and Pearl, we can apply the virtual evidence formula *if we think the two expert judgements are independent*. What 'independence' means for virtual evidence is a bit murky, since the evidence is not part of our probability calculus, so we can't apply the usual probabilistic definition. However, Pearl argues persuasively that this condition is easier to evaluate in practice than the *rigidity* condition which governs the applicability of Jeffrey updates. (He also gives an example where rigidity is violated, so a naive Jeffrey update gives a nonsensical result but where virtual evidence can still be easily applied to get a correct result.)

The information provided by the anthropologist and the metallurgist seem to be quite independent types of information (at least, if we ignore the fact that both experts are biased by an interest in ancient aliens), so let's apply the virtual evidence rule. The likelihood ratio

from the metallurgist was 80:20, which simplifies to 4:1. The likelihood ratio from the anthropologist was 1:2. That makes the combined likelihood vector 2:1 in favor of terrestrial origin. We would then combine this with our prior; for example, if we had a prior of 3:1 in favor of a terrestrial origin, our posterior would be 6:1 in favor.

(Note that we also have to think that the virtual evidence is independent *of our prior information*.)

So, virtual evidence offers a practical way to integrate information when we cannot quantify exactly what the evidence was -- a condition which is especially likely when consulting experts. This illustrates the utility of the bayes-with-a-side-channel model mentioned earlier; we are able to deal effectively with evidence, even when the exact nature of the evidence is hidden to us.

A few notes on how we gathered expert information in our hypothetical example.

- We asked for likelihood ratios, rather than posterior probabilities. This allows us to combine the information as virtual evidence.
- In the case of the metallurgist, it makes sense to ask for likelihood ratios, since the metallurgist is unlikely to have good prior information about the artefact. Asking only for likelihoods allows us to factor out any effect from this poor prior (and instead use our own prior, which may still be poor, but has the advantage of being ours).
- In the case of the anthropologist, however, it doesn't make as much sense -- if we trust their expertise, we're likely to think the anthropologist has a good prior about artefacts. It might have made more sense to ask for the anthropologist's posterior, take it as our own, and *then* apply a virtual-evidence update to integrate the metallurgist's report. (However, if we weren't able to properly communicate our own prior information to the anthropologist, it would be ignored in such an approach.)
- In the case of the metallurgist, it felt more natural to give a full likelihood function, rather than a likelihood ratio. It makes sense to know the probability of test result given a particular substance. It would have made even more sense if the likelihood function were a *function of each metal the artefact could be made of*, rather than just "terrestrial" or "extraterrestrial" -- using broad categories allows the metallurgist's prior about specific substances to creep in, which might be unfortunate.
- In the case of the anthropologist, however, it didn't make sense to give a full likelihood function. "The probability that the artefact would look exactly the way it looks assuming that it's made by humans" is very very low, and seems quite difficult and unnatural to evaluate. It seems much easier to come up with a likelihood *ratio*, comparing the probability of terrestrial and extraterrestrial origin.

Why did Pearl devote several sections to virtual evidence, in a book which is otherwise a bible for dogmatic probabilists? I think the main reason is the close analogy to the mathematics of Bayesian networks. The message-passing algorithm which makes Bayesian networks efficient is almost exactly the virtual evidence procedure I've described. If we think of each node as an expert trying to integrate information from its neighbors, then the efficiency of Bayes nets comes from the fact that they can use virtual evidence to update on likelihood functions rather than needing to know about the evidence in detail. This may have even been one source of inspiration for Pearl's belief propagation algorithm?

## Can Dogmatic Probabilists Use Virtual Evidence?

OK, so we've put Jeffrey's radical updates into a more palatable form -- one which borrows the structure and notation of classical Bayesian updates.

Does this mean orthodox Bayesians can join the party, and use virtual evidence to accomplish everything a radical probabilist can do?

No.

### **Virtual evidence abandons the ratio formula.**

One of the longstanding axioms of classical Bayesian thought is the ratio formula for conditional probability that Bayes himself introduced:

$$P(A|B) = \frac{P(A \& B)}{P(B)}$$

Virtual evidence, as an updating practice, holds that  $P(A|B)$  can be usefully defined in cases where the ratio  $P(A \& B)/P(B)$  **cannot** be usefully defined. Indeed, virtual evidence treats Bayes' Law (which is usually a derived theorem) as more fundamental than the ratio formula (which is usually taken as a definition).

Granted, dogmatic probabilism *as I defined it at the beginning of this post* does not explicitly assume the ratio formula. But the assumption is so ingrained that I assume most readers took  $P(A|B)$  to mean the ratio.

Still, even so, we can *consider* a version of dogmatic probabilism which rejects the ratio formula. Couldn't they use virtual evidence?

### **Virtual evidence requires probability functions to take arguments which aren't part of the event space.**

Even abandoning the ratio formula, still, it's hard to see how a dogmatic probabilist could use virtual evidence without abandoning the Kolmogorov axioms as the foundation of probability theory. The Kolmogorov axioms make probabilities a function of events; and events are taken from a pre-defined event space. Virtual evidence constructs new events at will, and does not include them in an overarching event space (so that, for example, virtual evidence  $V$  can be defined -- so that  $P(X|V)$  is meaningful for all  $X$  from the event space --without events like  $X \& V$  being meaningful, as would be required for a sigma-algebra).

I left some wiggle room in my definition, saying that a dogmatic probabilist might endorse the Kolmogorov axioms "or a similar axiomatization of probability theory". But even the Jeffrey-Bolker axioms, which are pretty liberal, don't allow enough flexibility for this!

## **Representing Fluid Updates**

A final point about virtual evidence and Jeffrey updates.

Near the beginning of this essay, I gave a picture in which Jeffrey updates generalize Bayesian updates, but *fluid* updates generalize things even further, opening up the space of possibilities when rigidity does not hold.

However, I should point out that *any update is a Jeffrey update on a sufficiently fine partition.*

So far, for simplicity, I've focused on binary partitions: we're judging between  $H$  and  $\neg H$ , rather than a larger set such as  $H_1, H_2, H_3$ . However, we can generalize everything to arbitrarily sized partitions, and will often want to do so. I noted that a larger set might have

been better when asking the metallurgist about the artefact, since it's easier to judge the probability of test results given specific metals rather than broad categories.

If we make a partition large enough to cover every possible combination of events, then a Jeffrey update is now just a completely arbitrary shift in probability. Or, alternatively, we can represent arbitrary shifts via virtual evidence, by converting to likelihood-ratio format.

So, these updates are completely general after all.

Granted, there might not be any *point* to seeing things that way.

## Non-Sequential Prediction

One advantage of radical probabilism is that it offers a more general framework for statistical learning theory. I already mentioned, briefly, that it allows one to do away with the realizability/grain-of-truth assumption. This is very important, but not what I'm going to dwell on here. Instead I'm going to talk about non-sequential prediction, which is a benefit of logical induction which I think has been under-emphasized so far.

Information theory -- in particular, algorithmic information theory -- in particular, Solomonoff induction -- is restricted to a *sequential prediction* frame. This means there's a very rigid observation model: observations are a sequence of tokens and you always observe the *n*th token after observing tokens one through *n*-1.

Granted, you can fit lots of things into a sequential prediction model. However, it is a flaw the otherwise close relationship between Bayesian probability and information theory. You'll run into this if you try to relate information theory and logic. Can you give an information-theoretic intuition for the laws of probability that deal with logical combinations, such as  $P(A \text{ or } B) + P(A \text{ and } B) = P(A) + P(B)$ ?

I've [complained about this before](#), offering a theorem which (somewhat) problematizes the situation, and suggesting that people should notice whether or not they're making sequential-prediction style assumptions. I almost included related assumptions in my definition of dogmatic probabilism at the beginning of this post, but ultimately it makes more sense to contrast radical probabilism to the more general doctrine of Bayesian updates.

Sequential prediction cares only about the accuracy of beliefs *at the moment of observation*; the accuracy of the full distribution over the future is reduced to the accuracy about each next bit as it is observed.

If information is coming in "in any old way" rather than according to the assumptions of sequential prediction, then we can construct problematic cases for Solomonoff induction. For example, if we condition the *n*th bit to be 1 (or 0) when a theorem prover proves (or refutes) the *n*th sentence of Peano arithmetic, then Solomonoff induction will never assign positive probability to hypotheses consistent with Peano arithmetic, and will therefore do poorly on this prediction task. This is despite the fact that there are *computable* programs which do better at this prediction task; for example, the same theorem prover running just a little bit faster can have highly accurate beliefs at the moment of observation.

In non-sequential prediction, however, we care about accuracy *at every moment*, rather than just at the moment of observation. Running the same theorem prover, just one step faster, doesn't do very well on that metric. It allows you to get things right just in time, but you won't have any clue about what probabilities to assign before that. We don't just want the right conclusion; we want to get there as fast as possible, and (in a subtle sense) via a rational path

Part of the difficulty of non-sequential prediction is how to score it. Bayes loss applied to your predictions at the moment of observation, in a sequential prediction setting, seems quite useful. Bayes loss applied to all your beliefs, at every moment does not seem very useful.

Radical probabilism gives us a way to evaluate the rationality of non-sequential predictions -- namely, how vulnerable the sequence of belief distributions was to losing money via some sequence of bets.

Sadly, I'm not yet aware of any appropriate generalization of information theory -- at least not one that's very interesting. (You can index information by time, to account for the way probabilities shift over time... but that does not come with a nice theory of communication or compression, which are fundamental to classical information theory.) This is why I [objected to prediction=compression in the discussion section of Alkjash's talk](#).

To summarize, sequential prediction makes three critical assumptions which may not be true in general:

- It assumes observations will always inform us about one of a set of observable variables. In general, Bayesian updates can instead inform us about any event, including complex logical combinations (such as "either the first bit is 1, or the second bit is 0").
- It assumes these observations will be made in a specific sequence, whereas in general updates could come in any order.
- It assumes that what we care about is the accuracy of belief *at the time of observation*; in general, we may care about the accuracy of beliefs at other times.

The only way I currently know how to get theoretical benefits similar to those of Solomonoff induction while avoiding all three of these assumptions is radical probabilism (in particular, as formalized by logical induction).

*(The connection between this section and radical probabilism is notably weaker than the other parts of this essay. I think there is a lot of low-hanging fruit here, fleshing out the space of possible properties, the relationship between various problems and various assumptions, trying to generalize information theory, clarifying our concept of observation models, et cetera.)*

## Making the Meta-Bayesian Update

In *Pascal's Muggle* ([long version](#), [short version](#)) Eliezer discusses situations in which he would be forced to make a non-Bayesian update:

But if I actually see strong evidence for something I previously thought was super-improbable, I don't just do a Bayesian update, I should also question whether I was right to assign such a tiny probability in the first place - whether the scenario was really as complex, or unnatural, as I thought. In real life, you are not ever supposed to have a prior improbability of  $10^{-100}$  for some fact distinguished enough to be written down, and yet encounter strong evidence, say  $10^{10}$  to 1, that the thing has actually happened. If something like that happens, you don't do a Bayesian update to a posterior of  $10^{-90}$ .

Instead you question both whether the evidence might be weaker than it seems, and whether your estimate of prior improbability might have been poorly calibrated, because rational agents who actually have well-calibrated priors should not encounter situations like that until they are ten billion days old. Now, this may mean that I end up doing some non-Bayesian updates: I say some hypothesis has a prior probability of a quadrillion to one, you show me evidence with a likelihood ratio of a billion to one, and I say 'Guess I was wrong about that quadrillion to one thing' rather than being a Muggle about it.

At the risk of being too cutesy, I want to make two related points:

- At the object level, radical probabilism offers a framework in which we can make these sorts of non-Bayesian updates. We can encounter something which makes us question our whole way of thinking. It also allows us to significantly revise that way of thinking, without modeling the situation as something extreme like self-modification (or even something very out of the ordinary).
- At the meta level, updating to radical probabilism *is itself* one of these non-Bayesian updates. Of course, if you were really a hard-wired dogmatic probabilist at core, you would be unable to make such an update (except perhaps if we model it as self-modification). But, since you are *already* using reasoning which is actually closer in spirit to radical probabilism, you can start to model yourself in this way and using radical-probabilist ideas to guide future updates.

So, I wanted to use this penultimate section for some advice about making the leap.

## It All Adds Up to Normality

Radical Probabilism is not a license to update however you want, nor even an invitation to massively change the way you update. It is primarily a new way to understand what you are already doing. Yes, it's possible that viewing things through this lens (rather than the more narrow lens of dogmatic probabilism) will change the way you see things, and as a consequence, change the way you do things. However, you are not (usually) making some sort of mistake by engaging in the sort of Bayesian reasoning you are familiar with -- there is no need to abandon large portions of your thinking.

Instead, try to notice ordinary updates you make which are not perfectly understood as Bayesian updates.

- Calibration corrections are not well-modeled as Bayesian updates. If you say to yourself "I've been overconfident in similar situations", and lower your probability, your shift is better-understood as a fluid update.
- Many instances of "outside view" are not well-modeled in a Bayesian update framework. You've probably seen outside view explained as prior probability. However, you often take the outside view on one of your own arguments, e.g. "I've often made arguments like this and been wrong". This kind of reflection doesn't fit well in the framework of Bayesian updates, but fits fine in a radical-probabilist picture.
- It is often warranted to downgrade the probability of a hypothesis without having an alternative in mind to upgrade. You can start to find a hypothesis suspicious without having any better way of predicting observations. For example, a sequence of surprising events might stick out to you as evidence that your hypothesis is wrong, even though your hypothesis is still the best way that you know to try and predict the data. This is hard to formalize as a Bayesian update. Changes in probability between hypotheses always remain balanced. It's true that you move the probability to a "not the hypotheses I know" category which balances the probability loss, but *it's not true that this category earned the increased probability by predicting the data better*. Instead, you used a set of heuristics which have worked well in the past to decide when to move probabilities around.

## Don't Predictably Violate Bayes

Again, this is not a license to violate Bayes' Rule whenever you feel like it.

A radical probabilist should obey Bayes' Law in expectation, in the following sense:

If some evidence  $E$  or  $\neg E$  is bound to be observed by time  $m > n$ , then the following should hold:

$$E_n(P_m(H) | E) = P_n(H | E)$$

And the same for  $\neg E$ . In other words, you should not expect your updated beliefs to differ from your conditional probabilities on average.

(*You should suspect from the fact that I'm not proving this one that I'm playing a bit fast and loose -- whether this law holds may depend on the formalization of radical probabilism, and it probably needs some extra conditions I haven't stated, such as  $P(E) > 0$ .*)

And remember, every update is a Bayesian update, with the right virtual evidence.

## Exchange Virtual Evidence

Play around with the epistemic practice Jeffrey suggests. I suspect some of you already do something similar, just not necessarily calling it by this name or looking so closely at what you're doing.

## Don't Be So Realist About Your Own Utility Function

Note that the picture here is quite compatible with what I said in [An Orthodox Case Against Utility Functions](#). Your utility function need not be computable, and there need not be something in your ontology which you can think of your utility as a function of. All you need are utility *expectations*, and the ability to update those expectations. Radical Probabilism adds a further twist: you don't need to be able to predict those updates ahead of time; indeed, you probably can't. Your values aren't tied to a function, but rather, are tied to your trust in the ongoing process of reasoning which refines and extends those values (very much like the self-trust discussed in the section on conservation of expected evidence).

## Not So Radical After All

And remember, every update is a Bayesian update, with the right virtual evidence.

## Recommended Reading

[Diachronic Coherence and Radical Probabilism](#), Brian Skyrms

- This paper is really nice in that it constructs Radical Probabilism from the ground up, rather than starting with regular probability theory and relaxing it. It provides a view in which diachronic coherence is foundational, and regular one-time-slice probabilistic coherence is derived. Like logical induction, it rests on a market metaphor. It also briefly covers the argument that radical-probabilism beliefs must have a convergence property.

[Radical Probabilism and Bayesian Conditioning](#), Richard Bradley

- This is a more thorough comparison of radical probabilism to standard bayesian probabilism, which breaks down the departure carefully, while covering the fundamentals of radical probabilism. In addition to Bayesian conditioning and Jeffrey

conditioning, it introduces Adams conditioning, a new type of conditioning which will be valid in many cases (for the same sort of reason as why Jeffrey conditioning or Bayesian conditioning can be valid). He contends that there are, nonetheless, many more ways to update beyond these; and, he illustrates this with a purported example where none of those updates seems to be the correct one.

[Epistemology Probabilized](#), Richard Jeffrey

- The man himself. This essay focuses mainly on how to update on likelihood ratios rather than directly performing Jeffrey updates (what I called virtual evidence). The motivations are rather practical -- updating on expert advice when you don't know precisely what observations lead to that advice.

[I was a Teenage Logical Positivist \(Now a Septuagenarian Radical Probabilist\)](#), Richard Jeffrey.

- Richard Jeffrey reflects on his life and philosophy.

Probabilistic Reasoning in Intelligent Systems, Judea Pearl.

- See especially chapter 2, especially 2.2.2 and 2.3.3.

[Logical Induction](#), Garrabrant et al.

\*: Jeffrey actually used this phrase. See *I was a Teenage Logical Positivist*, linked above.

# The Fusion Power Generator Scenario

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

Suppose, a few years from now, I prompt GPT-N to design a cheap, simple fusion power generator - something I could build in my garage and use to power my house. GPT-N succeeds. I build the fusion power generator, find that it works exactly as advertised, share the plans online, and soon the world has easy access to cheap, clean power.

One problem: at no point did it occur to me to ask “Can this design easily be turned into a bomb?”. Had I thought to prompt it with the question, GPT-N would have told me that the design *could* easily be turned into a bomb. But I didn’t think to ask, so GPT-N had no reason to mention it. With the design in wide use, it’s only a matter of time until people figure it out. And so, just like that, we live in a world where anyone can build a cheap thermonuclear warhead in their garage.

This scenario highlights a few key constraints which I think are under-appreciated in alignment today.

## Sharing Information is Irreversible

I’ve heard people say that we can make AI safe(r) by restricting the AI’s action space to things which we can undo. Problem is, sharing information is irreversible; once the cat is out of the bag, there’s no getting it back into the bag. And if an AI can’t share information, there’s very little that it *can* do. Not much point in an AI which just can’t do anything observable at all. (One could design an AI to “move in mysterious ways”, but I have trouble imagining that it ends up *safer* that way.)

This is a problem when information itself is dangerous, e.g. knowledge of how to build a thermonuclear warhead in one’s garage.

## Humans Are Not Safe

Two key properties of humans:

- We do not have full introspective understanding of our own wants
- We do not have the processing power to fully understand the consequences of changes

Sometimes, we get something we thought we wanted, and find out that we don’t want it after all. Either we misunderstood our own wants, or misunderstood the full implications of the change.

Most of the time, this isn’t that huge an issue. We lose some money and/or time, but we move on.

But if a human is capable of making large, irreversible changes to the world, then the problem becomes more serious. A human with access to powerful AI - even something

as conceptually simple as GPT-N - is capable of making large irreversible changes, and they do not have the processing power to fully understand the implications of those changes. In general, a human won't even know the right questions to ask. So, if a system's safety relies on a human asking the right questions, then the system is not safe.

In particular, this is relevant to the [HCH family of alignment schemes](#) (e.g. IDA), as well as human-imitating AI more broadly.

## Corollary: Tool AI Is Not Inherently Safe

Tool AI, in particular, relies primarily on human operators for safety. Just like a tablesaw is safe if-and-only-if the operator uses it safely, tool AI is safe if-and-only-if the operator uses it safely.

With a tablesaw, that's usually fine. It's pretty obvious what sorts of things will lead to bad outcomes from a tablesaw. But the big value-proposition of powerful AI is its ability to reason about systems or problems too complicated for humans - which are exactly the systems/problems where safety issues are likely to be nonobvious. If we're going to unlock the full value of AI at all, we'll need to use it on problems where humans do not know the relevant safety issues. So: if the system's safety relies on a human using it safely, then it's not safe.

If you want a concrete, evocative analogy: picture a two-year-old playing on top of a tablesaw.

That said, people are designing [tablesaws which auto-stop when skin contacts the blade](#). In general, a system's *designers* may understand the relevant safety issues better than the operators. Indeed, since the first AGIs will be built by humans, *any* approach to AI safety ultimately relies on human designers asking the right questions. Point is: we can't avoid the need for designers to ask (at least some of) the right questions upfront. But needing the designers to ask the right questions once is still a lot better than needing every user to ask the right questions every time they use the system.

(This perspective ties in nicely with [AI alignment as interface design](#): if an interface offers an easy-to-overlook way to cut your hand off, and relies on users not doing so, then that's a design problem.)

Safe tool AI could potentially be built, but safety won't happen by itself any more than it would for other kinds of AI.

## Generalization: Great Power, Great Responsibility

Finally, note that none of this is an issue if GPT-N can't design fusion power generators (or garage warheads) at all. In general, it is easy to come up with designs for probably-safe AIs which just can't do anything all that impressive. The greater an AI's capabilities, the more precisely and reliably it needs to be aligned to human values.

In particular, the “capabilities” relevant here are an AI’s abilities to reason about systems too complicated for humans or solve problems too complicated for humans. It’s the complexity that matters; the inability of humans to fully understand all the implications of the AI’s reasoning/solutions is exactly what makes humans unreliable judges of safety. So, the greater the complexity of systems/problems an AI can handle, the more important it is for that AI to have its own model of what-humans-want, and to align its solutions with that model.

# The Bayesian Tyrant

Long ago and far away, there was a kingdom called Estimor in a broad green valley surrounded by tall grey mountains. It was an average kingdom in most respects, until the King read the works of Robin Hanson and Eliezer Yudkowsy, and decided to institute a Royalist Futarchy.

*(This is a parable about the differences between Bayesian updates and logical induction. See also: [Radical Probabilism](#).)*

The setup was very simple. It followed the [futarchic motto](#), "Vote Values, But Bet Beliefs" -- the only special consideration being that there was just one voting constituent (that being the King). A betting market would inform the King of everything He needed to know in order to best serve His interests and the interests of Estimor (which were, of course, one and the same).

The Seer's Hall -- a building previously devoted to religious prophecy -- was repurposed to serve the needs of the new betting market. (The old prophets were, of course, welcome to participate -- so long as they were willing to put money on the line.)

All went well at first. The new betting market allowed the King to set the revenue-maximizing tax rate, via the Laffer curve. An early success of the market was the forecasting of a grain shortage based on crop growth early in the season, which allowed ample time for grain to be purchased from neighboring lands.

Being an expert Bayesian Himself, the King would often wander the Seer's Hall, questioning and discussing with the traders at the market. Sometimes the King would be shocked by what he learned there. For example, many of the traders were calculating the [Kelly betting criterion](#) to determine how much to invest in a single bet. However, they then proceeded to invest *only a set fraction* of the Kelly amount (such as 80%). When questioned, traders replied that they were hedging against mistakes in their own calculations, or reducing volatility, or the like.

One day, the King noticed a man who would always run in and out of the Hall, making bets hastily. This man did particularly well at the betting tables -- he ended the day with a visibly heavy purse. However, when questioned by The King as to the source of his good luck, the man had no answers. This man will subsequently be referred to as the Fool.

The King ordered spies to follow the Fool on his daily business. That evening, spies returned to report that The Fool was running back and forth between the Seer's Hall and Dragon's Foot, a local tavern. The Fool would consult betting odds at Dragon's Foot, and return to the Seer's Hall to bet using those odds.

Evidently, Dragon's Foot had become an unlicensed gambling den. But were they truly doing better than the Seer's Hall, so that this man could profit simply by using their information?

The King had the Fool brought in for questioning. As it turned out, the Fool was turning a profit by *arbitrage* between the two markets: whenever there was a difference in prices, the Fool would bet in favor at the location where prices were low, and bet

against at the location where prices were high. In this way, he was making guaranteed money.

The King was disgusted at this way of making money without bringing valuable information to the market. He ordered all other gambling in the Kingdom shut down, requiring it to all take place at the Seer's Hall.

Soon after that, the Fool showed his face again. Once again, he did well in the market. The King had his spiel follow the Fool, but this time, he went nowhere of significance.

Questioning the Fool a second time, he learned that this time the Fool was making use of *calibration charts*. The Fool would make meticulous records of the true historical frequency of events given their probabilistic judgement -- for example, he had recorded that when the market judges an event to be 90% probable, that event actually occurs about 85% of the time. The Fool had made these records about individual traders as well as the market as a whole, and would place bets accordingly.

The King was once again disgusted by the way the Fool made money off of the market without contributing any external information. But this time, He felt that He needed a more subtle solution to the problem. Thinking back to his first days of reading about Bayes' Law, the King realized the huge gap between His vision of perfected reasoning and the reality of the crowded, noisy, irrational market. The iron law of the market was *buy low sell high*. It did not follow rational logic. The Fool had proved it: the individual traders were poorly calibrated, and so was the market itself.

What the King needed to do was reform the market, making it a more rational place.

And so it was that the King instituted the Bayesian Law: *all bets on the market are required to be Kelly bets made on valid probability estimates. Valid probability estimates are required to be Bayesian updates of previously registered probability estimates.*

All traders on the market would now proceed according to Bayes' Law. They would pre-register their probability distributions, pre-specifying what kind of information would update them, and by how much it would update them.

The new ordinance proved burdensome. Only a few traders continued to visit the Seer's Hall. They spent their days in meticulous calculation, updating detailed prior models of grain and weather with all the data which poured in.

Surprisingly, the Fool was amongst the hangers-on, and continued to make a tidy profit, even to the point of driving out some of the remaining traders -- they simply couldn't compete with him.

The King examined the registered probability distribution the Fool was using. It proved puzzling. The Fool's entire probability distribution was based on numbers which were to be posted to a particular tree out by Mulberry road. Updating on these numbers, the Fool was somehow a tidy profit. But where were the numbers coming from?

The King's spies found that the numbers were being posted by a secretive group, whose meetings they were unable to infiltrate.

The King had all the attendees arrested, accusing them of running an illegal gambling ring. The Fool was brought in for questioning once more.

"But it wasn't a gambling ring!" the Fool protested. "They merely got together and compiled *odds* for gambling. They were quite addicted when the Bayesian Law shut their sort out of the Seer's Hall, after all. And I took those odds and used them to bet in the Seer's Hall, perfectly legally."

"And redistributed the winnings?" accused the King.

"As is only fair," agreed the Fool. "But that is not gambling. I simply paid them as consultants."

"You took money from honest Bayesians, and drove them out of my Hall!"

"As is the advantage of Bayesianism, no?" The Fool cocked an eyebrow. "The money flows to he who can give the best odds."

"Take him away!" the king bellowed, waving a hand for the guards.

At that moment, the guards removed their helmets, revealing themselves to be comrades-in-arms with the Fool. The outcasts of the Seer's Hall had foreseen that the King would move against them, and with the power of Futarchy, had prepared well -- they staged a bloodless revolution that day.

The King, his family, and his most loyal staff were forced into exile. They went to stay with a distant cousin of the King, who ruled the nation Ludos, in the next valley over.

The King of Ludos had, upon seeing Estimor's success with prediction markets, set up His own. Unlike the Seer's Hall of Estimor, that of Ludos continued to thrive.

The King in Exile asked his cousin: "What did I do wrong? All I wanted was to serve Estimor. The prediction market worked so well at first. And I only tried to improve it."

The King of Ludos sat in thought for a time, and then spoke. "Cousin, We cannot tell You that You did anything wrong. Revolutions will happen. But We will say this: the many prediction markets of Ludos strengthen each other. Runners go back and forth between them, profiting from arbitrage, and this only makes them stronger. Calibration traders correct any bias of the market, ensuring unbiased results. You tried to outlaw the irrational at your market -- but remember, the wise gambler profits off the foolhardy gambler. Without any novices throwing away their money, none would profit."

"But most of all, cousin, We think You lost sight of the power of betting. It is a truth more fundamental than Bayes' Law that money will flow from the uncles to the clever. You lost Your trust in that system. Even if You had enforced Kelly betting, but left it to each individual trader to set his probability however he liked -- rather than updating via Bayes' Law alone -- you would have been fine. If Bayes' Law were truly the correct way, money would have flowed to those who excelled at it. If not, then money would have flowed elsewhere. But instead you overwhelmed them with the bureaucracy of Bayes -- requiring them to record every little bit of information they used to reach a conclusion."

---

*The arbitrage between different betting halls represented outside view / modest epistemology, trying to reach agreement between different reasoners. It's a questionable thing to include, in terms of the point I'm making, since this is not exactly a thing that happens in logical induction. However, it fits in the allegory so well*

*that I felt I couldn't not include it. One argument for the common prior assumption (an assumption which underpins the Aumann Agreement Theorem, and is closely related to modest-epistemology arguments) is that a bookie can Dutch Book any group of agents who do not have a common prior, via performing arbitrage on their various beliefs.*

*[Edit: actually, what we can conclude from the analogy is that bets on different markets should converge to the same thing **if they ever pay out**, which is also true in logical induction.]*

*The calibration-chart idea, clearly, represented [calibration properties](#).*

*The idea of the Bayesian Law represented requiring all hypotheses/traders to update in a Bayesian manner. Starting from Bayesian hypothesis testing, one step we can take in the direction of logical induction is to allow hypotheses to themselves make non-Bayesian updates. The overall update **between** hypotheses would remain Bayesian, but an individual hypothesis could change its mind in a non-Bayesian fashion. A hypothesis would still be required to have a coherent probability distribution at any given time; just, the updates could be non-Bayesian. A fan of Bayes' Law might suppose that, in such a situation, the hypotheses which update according to Bayes' Law would dominate -- in other words, a meta-level Bayesian would learn to also be an object-level Bayesian. But I see no reason to suspect this to be the case. Indeed, in situations where logical uncertainty is relevant, non-Bayesian updates can be used to continue improving one's probability distribution over and above the explicit evidence which comes in. It was this idea -- that we could take one step toward logical induction by being a meta-level Bayesian without being an object-level Bayesian -- which inspired this post (although the allegory didn't end up having such a strong connection with this idea).*

*The main point of this post, anyway, is that **Bayes' Law would be a bad law**. Don't institute a requirement that everyone reason according to it.*

# Tools for keeping focused

Once I realized that [my attention was even scarcer than my time](#), I became an anti-distraction fanatic. During my [weekly reviews](#) I methodically went through my past week, figured out what had been distracting me, and tried to eliminate it or replace it with something less distracting.

Over time, this has led me to find lots of tools (and ways of using my tools) that help me stay more focused. Here are some of the things I've started doing:

- I *aggressively* disable notifications and badges so that I don't mindlessly open distracting apps. If you're into eliminating distractions you've probably already done this. But if you haven't, it's by far the most important thing you can do to improve your focus, so I'm putting it first anyway.



Anxious yet?

I have a zero-tolerance notification policy: if an app interrupts me, I ask myself whether the interruption was valuable, and if not, the app doesn't get to notify me anymore. This has weeded out pretty much everything except for inboxes (phone, texts, reminders) and apps with a human on the other end (ride sharing, delivery).

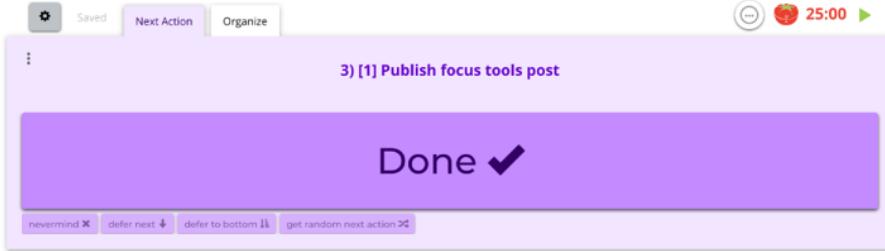
A special dishonorable mention goes to Slack, which can easily suck away a quarter of your time without you noticing. If you use Slack and you haven't disabled the "unread messages" badge, stop reading this post and do it now. (Consider setting a two-minute timer to remind you to quit in case you get distracted by checking your unreads, like I did while taking the screenshot above.)

- On the subject of Slack, I try to keep it closed as much as possible and check it in batches a few times a day. Not all workplace Slack cultures allow this, but if yours does, I highly, *highly* recommend it. Doing this led to my [biggest single quantifiable productivity improvement](#).\*\* If your workplace culture *doesn't* allow you to keep Slack closed, because it requires quick Slack responses, this is a bad sign.
- I only check my email once per day.

Gmail filters send important email to a label called "Temp Inbox" and the rest to a label called "Unimportant." I check Temp Inbox each evening and Unimportant once a week. Since incoming email doesn't go to the inbox, I can open Gmail to compose or search messages without getting distracted by unread messages.

(I use [a piece of Google Apps Script](#) for this, but I think the Gmail UI has improved recently so that you can now do something similar with filters and it'll have decent ergonomics.)

- I have a second monitor that always shows a [Complice](#) window with the task I'm currently working on. (Complice is my favorite app for making lists of what I want to do today.) This helps me recover quickly from unintentionally going down rabbit holes.



- I use [Focus](#) to break my habit of mindlessly checking sites.

For websites that are habit-forming but still feel useful on net, like Twitter or Hacker News, Focus lets me restrict my usage to certain times of day. It's the only website blocker I've used that lets me [fully automate](#) blocking things in the exact way I want.

It's especially important for me to use a website blocker that's fully automatic, because the times when I need it most are exactly those times at which I have the least willpower to do any manual steps!

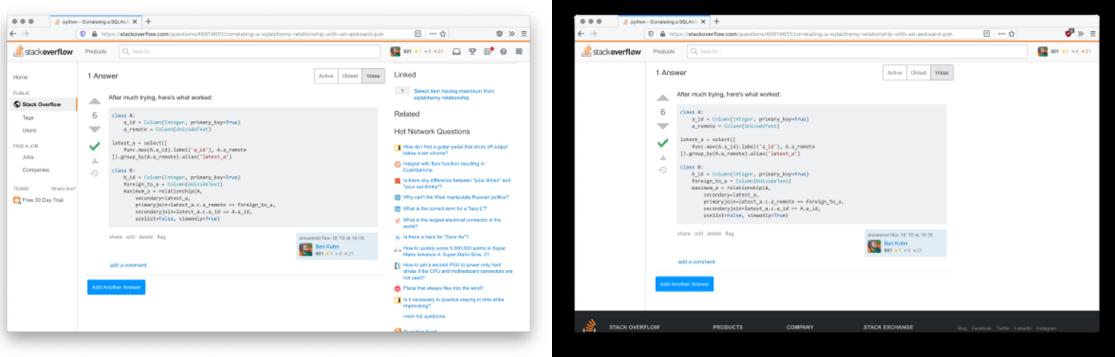
- I also block most websites on my phone (using the iOS built-in content blocking in whitelist mode). Unfortunately, this works less well than Focus since I sometimes want to disable it and then forget to re-enable it.
- I [do most of my Internet reading on my Kindle](#) via [Kindle4RSS](#).

Using RSS means I'm in control of my own feed and don't need to visit an [adversarially distracting](#) site like Facebook to get new reading material.

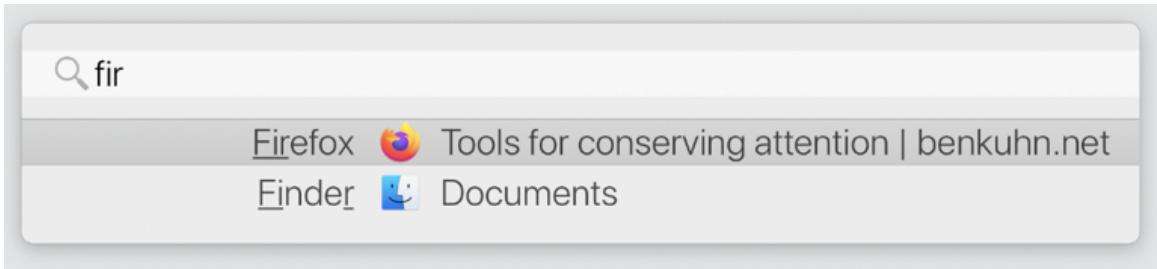
It's also helpful that the Kindle delivery comes once per day at a predictable time, so I don't have an urge to check over and over again for new content.

- I [block distracting parts of websites](#) with [uBlock Origin's](#) amazing "element blocker," which lets you select and remove any part of the page. As a fan of minimalism in web design, I really enjoy being able to adversarially enforce it on any rogue webpage.

For instance, I use it to block the clickbaity "hot network questions" sidebar on Stack Overflow, which otherwise frequently distracts me, as well as the useless notifications and left sidebar:



- I use [Witch](#), an augmented window switcher, to avoid accidentally switching to the wrong window. Witch can display separate app windows separately, only display windows from the current workspace, and supports [text search on window titles](#) so that you can search directly for the window you want without having to skip over.



I found Witch slightly unintuitive to configure, so if you're curious, here are screenshots of my configs: ["actions" tab](#), ["advanced" tab](#).

- I've hidden the apps on my phone. My homescreen looks like the figure on the right.

The two folders are "badges" (for the few apps that are allowed to notify me) and "everything else." If I want an app I type its name in the search bar, which forces me to be intentional (and is often faster than hunting through pages of apps anyway). The app in the bottom bar is the Kindle app.



- I use native versions of web apps instead of keeping them open in a browser tab. This allows me to use the app without getting sucked into my browser. There are two ways of doing this:
  - Some apps (e.g. [Roam](#)) are already "installable," meaning that they tell your browser how to turn them into a native app. (Sadly, Chrome seems to be the only browser that can install installable apps on desktop right now.)
  - For the rest (e.g., oddly, most Google web apps), I use a command-line tool called [nativefier](#) to turn them into desktop apps.
- In fact, right now I'm trying out a "no browser tabs at all" rule. I noticed that I'd sometimes get distracted by tabs that I'd opened a long time ago and should have closed, but forgot about. So I installed an extension to limit each window to a single tab and changed Firefox to [open links in a new window by default](#). (This has lots of synergy with a powerful window switcher like Witch.)

Each of these is small on their own, but [like many of the things I work on during weekly reviews](#), they've added up and compounded to make it much easier for me to spend my attention in ways I want.

# Matt Botvinick on the spontaneous emergence of learning algorithms

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

Matt Botvinick is Director of Neuroscience Research at DeepMind. In [this interview](#), he discusses results from a [2018 paper](#) which describe conditions under which reinforcement learning algorithms will spontaneously give rise to separate full-fledged reinforcement learning algorithms that differ from the original. Here are some notes I gathered from the interview and paper:

## Initial Observation

At some point, a group of DeepMind researchers in Botvinick's group noticed that when they trained a RNN using RL on a series of related tasks, the RNN itself instantiated a separate reinforcement learning algorithm. These researchers weren't trying to design a meta-learning algorithm—apparently, to their surprise, this just spontaneously happened. As Botvinick describes it, they started "with just one learning algorithm, and then another learning algorithm kind of... emerges, out of, like out of thin air":

"What happens... it seemed almost magical to us, when we first started realizing what was going on—the slow learning algorithm, which was just kind of adjusting the synaptic weights, those slow synaptic changes give rise to a network dynamics, and the dynamics themselves turn into a learning algorithm."

Other versions of this basic architecture—e.g., using slot-based memory instead of RNNs—seemed to produce the same basic phenomenon, which they termed "meta-RL." So they concluded that all that's needed for a system to give rise to meta-RL are three very general properties: the system must 1) have memory, 2) whose weights are trained by a RL algorithm, 3) on a sequence of similar input data.

From Botvinick's description, it sounds to me like he thinks [learning algorithms that find/instantiate other learning algorithms] is a strong attractor in the space of possible learning algorithms:

"...it's something that just happens. In a sense, you can't avoid this happening. If you have a system that has memory, and the function of that memory is shaped by reinforcement learning, and this system is trained on a series of interrelated tasks, this is going to happen. You can't stop it."

## Search for Biological Analogue

This system reminded some of the neuroscientists in Botvinick's group of features observed in brains. For example, like RNNs, the human prefrontal cortex (PFC) is highly recurrent, and the RL and RNN memory systems in their meta-RL model reminded them of "synaptic memory" and "activity-based memory." They decided to look for evidence of meta-RL occurring in brains, since finding a neural analogue of the technique would provide some evidence they were on the right track, i.e. that the technique might scale to solving highly complex tasks.

They think they found one. In short, they think that part of the dopamine system (DA) is a full-fledged reinforcement learning algorithm, which trains/gives rise to another full-fledged, free-standing reinforcement learning algorithm in PFC, in basically the same way (and for the same reason) the RL-trained RNNs spawned separate learning algorithms in their experiments.

As I understand it, their story goes as follows:

The PFC, along with the bits of basal ganglia and thalamic nuclei it connects to, forms a RNN. Its inputs are sensory percepts, and information about past actions and rewards. Its outputs are actions, and estimates of state value.

DA<sup>[1]</sup> is a RL algorithm that feeds reward prediction error to PFC. Historically, people assumed the purpose of sending this prediction error was to update PFC's synaptic weights. Wang et al. agree that this happens, but argue that the *principle* purpose of sending prediction error is to cause the creation of "a second RL algorithm, implemented entirely in the prefrontal network's activation dynamics." That is, they think DA mostly stores its model in synaptic memory, while PFC mostly stores it in activity-based memory (i.e. directly in the dopamine distributions).<sup>[2]</sup>

What's the case for this story? They cite a variety of neuroscience findings as evidence for parts of this hypothesis, many of which involve doing horrible things to monkeys, and some of which they simulate using their meta-RL model to demonstrate that it gives similar results. These points stood out most to me:

#### *Does RL occur in the PFC?*

Some scientists implanted neuroimaging devices in the PFCs of monkeys, then sat the monkeys in front of two screens with changing images, and rewarded them with juice when they stared at whichever screen was displaying a particular image. The probabilities of each image leading to juice-delivery periodically changed, causing the monkeys to update their policies. Neurons in their PFCs appeared to exhibit RL-like computation—that is, to use information about the monkey's past choices (and associated rewards) to calculate the expected value of actions, objects and states.

Wang et al. simulated this task using their meta-RL system. They trained a RNN on the changing-images task using RL; when run, it apparently demonstrated similar performance as the monkeys, and when they inspected it they found units that similarly seemed to encode EV estimates based on prior experience, continually adjust the action policy, etc.

Interestingly, the system continued to improve its performance even once its weights were fixed, which they take to imply that the learning which led to improved performance could only have occurred within the activation patterns of the recurrent network.<sup>[3]</sup>

#### *Can the two RL algorithms diverge?*

When humans perform two-armed bandit tasks where payoff probabilities oscillate between stable and volatile, they increase their learning rate during volatile periods, and decrease it during stable periods. Wang et al. ran their meta-RL system on the same task, and it varied its learning rate in ways that mimicked human performance. This learning again occurred after weights were fixed, and notably, between the end

of training and the end of the task, the learning rates of the two algorithms had diverged dramatically.

## Implications

The account detailed by Botvinick and Wang et al. strikes me as a relatively clear example of [mesa-optimization](#), and I interpret it as tentative evidence that the attractor toward mesa-optimization is strong. [Edit: Note that some commenters, like [Rohin Shah](#) and [Evan Hubinger](#), disagree].

These researchers did not set out to train RNNs in such a way that they would turn into reinforcement learners. It just happened. And the researchers seem to think this phenomenon *will* occur spontaneously whenever “a very general set of conditions” is met, like the system having memory, being trained via RL, and receiving a related sequence of inputs. Meta-RL, in their view, is just “an emergent effect that results when the three premises are concurrently satisfied... these conditions, when they co-occur, are sufficient to produce a form of ‘meta-learning’, whereby one learning algorithm gives rise to a second, more efficient learning algorithm.”

So on the whole I felt alarmed reading this. That said, if mesa-optimization is a standard feature<sup>[4]</sup> of brain architecture, it seems notable that humans don’t regularly experience catastrophic inner alignment failures. Maybe this is just because of some non-scalable hack, like that the systems involved aren’t very powerful optimizers.<sup>[5]</sup> But I wouldn’t be surprised if coming to better understand the biological mechanisms involved led to safety-relevant insights.

*Thanks to Rafe Kennedy for helpful comments and feedback.*

---

1. The authors hypothesize that DA is a model-free RL algorithm, and that the spinoff (mesa?) RL algorithm it creates within PFC is model-based, since that’s what happens in their ML model. But they don’t cite biological evidence for this. ↵
2. Depending on what portion of memories are encoded in this way, it may make sense for cryonics standby teams to attempt to reduce the supraphysiological intracellular release of dopamine that occurs after cardiac arrest, e.g. [by administering D1-receptor antagonists](#). Otherwise entropy increases in PFC dopamine distributions may result in information loss. ↵
3. They demonstrated this phenomenon (continued learning after weights were fixed) in a variety of other contexts, too. For example, they cite an experiment in which manipulating DA activity was shown to directly manipulate monkeys’ reward estimations, independent of actual reward—i.e., when their DA activity was blocked/stimulated while they pressed a lever, they exhibited reduced/increased preference for that lever, even if pressing it did/didn’t give them food. They trained their meta-RL system to simulate this, again demonstrated similar performance as the monkeys, and again noticed that it continued learning even after the weights were fixed. ↵
4. The authors seem unsure whether meta-RL also occurs in other brain regions, since for it to occur you need A) inputs carrying information about recent actions/rewards, and B) network dynamics (like recurrence) that support continual activation. Maybe only PFC has this confluence of features. Personally,

I doubt it; I would bet that meta-RL (and other sorts of mesa-optimization) occur in a wide variety of brain systems, but it would take more time than I want to allocate here to justify that intuition. ↵

5. Although note that neuroscientists do commonly describe the PFC as disproportionately responsible for the sort of human behavior one might [reasonably wish](#) to describe as “optimization.” For example, the [neuroscience textbook](#) recommended on lukeprog’s [textbook recommendation post](#) describes PFC as “often assumed to be involved in those characteristics that distinguish us from other animals, such as self-awareness and the capacity for complex planning and problem solving.” ↵

# Alignment By Default

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

Suppose AI continues on its current trajectory: deep learning continues to get better as we throw more data and compute at it, researchers keep trying random architectures and using whatever seems to work well in practice. Do we end up with aligned AI “by default”?

I think there’s at least a plausible trajectory in which the answer is “yes”. Not very likely - I’d put it at ~10% chance - but plausible. In fact, there’s at least an argument to be made that alignment-by-default is *more* likely to work than many fancy alignment proposals, including [IRL variants](#) and [HCH-family methods](#).

This post presents the rough models and arguments.

I’ll break it down into two main pieces:

- Will a sufficiently powerful unsupervised learner “learn human values”? What does that even mean?
- Will a supervised/reinforcement learner end up aligned to human values, given a bunch of data/feedback on what humans want?

Ultimately, we’ll consider a semi-supervised/transfer-learning style approach, where we first do some unsupervised learning and hopefully “learn human values” before starting the supervised/reinforcement part.

As background, I will assume you’ve read some of the core material about human values from [the sequences](#), including [Hidden Complexity of Wishes](#), [Value is Fragile](#), and [Thou Art Godshatter](#).

## Unsupervised: Pointing to Values

In this section, we’ll talk about why an unsupervised learner might *not* “learn human values”. Since an unsupervised learner is generally just optimized for predictive power, we’ll start by asking whether theoretical algorithms with best-possible predictive power (i.e. Bayesian updates on low-level physics models) “learn human values”, and what that even means. Then, we’ll circle back to more realistic algorithms.

Consider a low-level physical model of some humans - e.g. a model which simulates every molecule comprising the humans. Does this model “know human values”? In one sense, yes: **the low-level model has everything there is to know about human values embedded within it, in exactly the same way that human values are embedded in physical humans**. It has “learned human values”, in a sense sufficient to predict any real-world observations involving human values.

But it seems like there’s a sense in which such a model does not “know” human values. Specifically, although human values are *embedded* in the low-level model, **the embedding itself is nontrivial**. Even if we have the whole low-level model, we still need that embedding in order to “point to” human values specifically - e.g. to use them as an optimization target. Indeed, when we say “point to human values”, what we mean is basically “specify the embedding”. (Side note: treating human values as an optimization target is not the only use-case for “pointing to human values”, and we still need to point to human values even if we’re not explicitly optimizing for anything. But that’s a [separate discussion](#), and imagining using values as an optimization target is useful to give a mental image of what we mean by “pointing”.)

**In short: predictive power alone is not sufficient to define human values. The missing part is the embedding of values within the model.** The hard part is pointing to the thing (i.e. specifying the values-embedding), not learning the thing (i.e. finding a model in which values are embedded).

Finally, here's a different angle on the same argument which will probably drive some of the philosophers up in arms: *any* model of the real world with sufficiently high general predictive power will have a model of human values embedded within it. After all, it has to predict the parts of the world in which human values are embedded in the first place - i.e. the parts of which humans are composed, the parts on which human values are implemented. So in principle, it doesn't even matter what kind of model we use or how it's represented; as long as the predictive power is good enough, values will be embedded in there, and the main problem will be finding the embedding.

## Unsupervised: Natural Abstractions

In this section, we'll talk about how and why a large class of unsupervised methods might "learn the embedding" of human values, in a useful sense.

First, notice that **basically everything from the previous section still holds if we replace the phrase "human values" with "trees"**. A low-level physical model of a forest has everything there is to know about trees embedded within it, in exactly the same way that trees are embedded in the physical forest. However, while there are trees *embedded* in the low-level model, the embedding itself is nontrivial. Predictive power alone is not sufficient to define trees; the missing part is the embedding of trees within the model.

More generally, whenever we have some high-level abstract object (i.e. higher-level than quantum fields), like trees or human values, a low-level model might have the object embedded within it but not "know" the embedding.

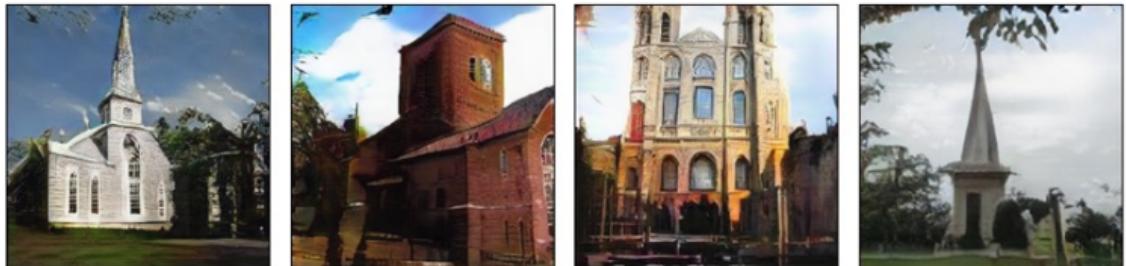
Now for the interesting part: **empirically, we have whole classes of neural networks in which concepts like "tree" have simple, identifiable embeddings**. These are unsupervised systems, trained for predictive power, yet they apparently "learn the tree-embedding" in the sense that the embedding is simple: it's just the activation of a particular neuron, a particular channel, or a specific direction in the activation-space of a few neurons.



Generate images of churches



Identify GAN units that match trees



Ablating units removes trees



Activating units adds trees

*Neat example with “trees” from the paper linked above.*

What's going on here? We know that models optimized for predictive power will not have trivial tree-embeddings *in general*; low-level physics simulations demonstrate that much. Yet these neural networks *do* end up with trivial tree-embeddings, so presumably some special properties of the systems make this happen. But those properties can't be *that* special, because we see the same thing for a reasonable variety of different architectures, datasets, etc.

Here's what I think is happening: **“tree” is a natural abstraction.** More on what that means [here](#), but briefly: abstractions summarize information which is relevant far away. When we summarize a bunch of atoms as “a tree”, we're throwing away lots of information

about the exact positions of molecules/cells within the tree, or about the pattern of bark on the tree's surface. But information like the exact positions of molecules within the tree is irrelevant to things far away - that signal is all wiped out by the noise of air molecules between the tree and the observer. The flap of a butterfly's wings may alter the trajectory of a hurricane, but unless we know how all wings of all butterflies are flapping, that tiny signal is wiped out by noise for purposes of our own predictions. Most information is irrelevant to things far away, not in the sense that there's no causal connection, but in the sense that the signal is wiped out by noise in other unobserved variables.

If a concept is a *natural* abstraction, that means that the concept summarizes all the information which is relevant to anything far away, and isn't too sensitive to the exact notion of "far away" involved. That's what I think is going on with "tree".

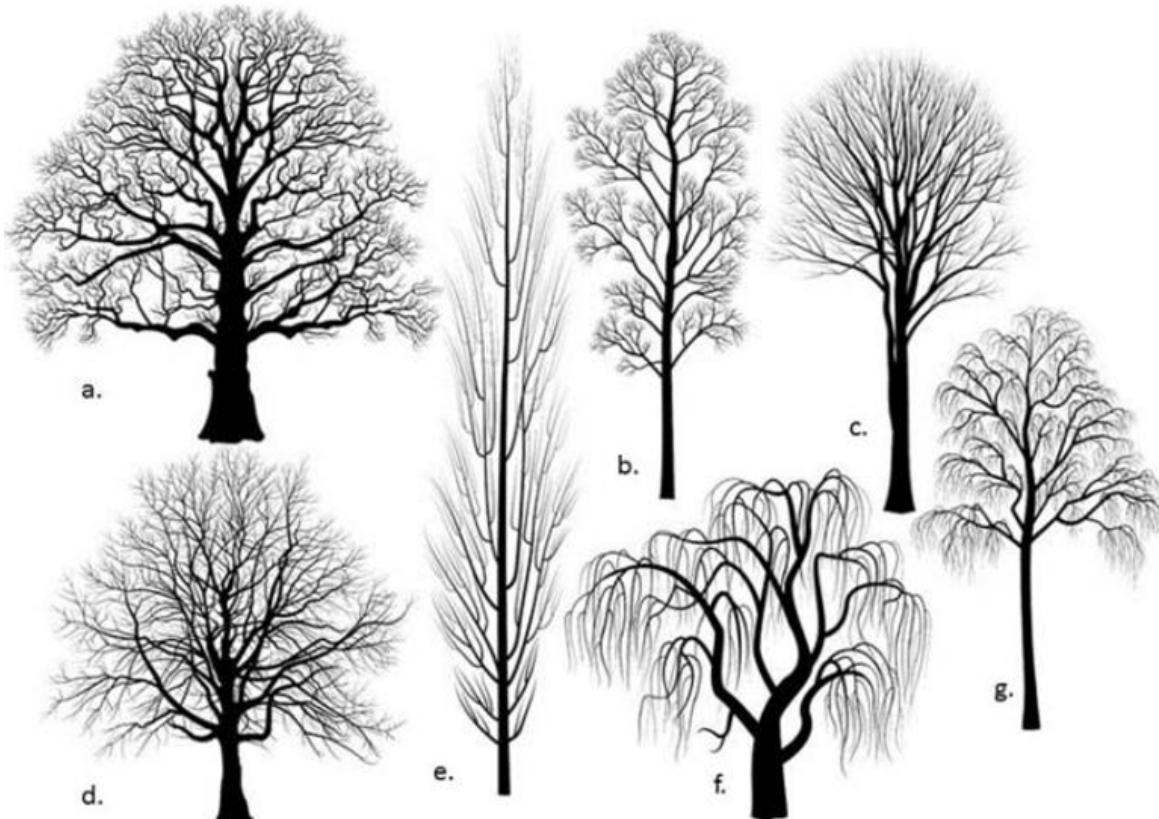
Getting back to neural networks: it's easy to see why a broad range of architectures would end up "using" natural abstractions internally. Because the abstraction summarizes information which is relevant far away, it allows the system to make far-away predictions without passing around massive amounts of information all the time. In a low-level physics model, we don't need abstractions because we *do* pass around massive amounts of information all the time, but real systems won't have anywhere near that capacity any time soon. So for the foreseeable future, **we should expect to see real systems with strong predictive power using natural abstractions internally.**

With all that in mind, it's time to drop the tree-metaphor and come back to human values. Are human values a natural abstraction?

If you've read [Value is Fragile](#) or [Godshatter](#), then there's probably a knee-jerk reaction to say "no". Human values are basically a bunch of randomly-generated heuristics which proved useful for genetic fitness; why would they be a "natural" abstraction? But remember, the same can be said of trees. Trees are a complicated pile of [organic spaghetti code](#), but "tree" is still a natural abstraction, because the concept summarizes all the information from that organic spaghetti pile which is relevant to things far away. In particular, it summarizes anything about one tree which is relevant to far-away trees.

Similarly, the concept of "human" summarizes all the information about one human which is relevant to far-away humans. It's a natural abstraction.

Now, I don't think "human values" are a natural abstraction in exactly the same way as "tree" - specifically, trees are abstract objects, whereas human values are *properties* of certain abstract objects (namely humans). That said, I think it's pretty obvious that "human" is a natural abstraction in the same way as "tree", and I expect that **humans "have values" in roughly the same way that trees "have branching patterns"**. Specifically, the natural abstraction contains a bunch of information, that information approximately factors into subcomponents (including "branching pattern"), and "human values" is one of those information-subcomponents for humans.



*Branching patterns for a few different kinds of trees.*

I wouldn't put super-high confidence on all of this, but given the remarkable track record of hackish systems learning natural abstractions in practice, I'd give maybe a **70% chance that a broad class of systems (including neural networks) trained for predictive power end up with a simple embedding of human values**. A plurality of my uncertainty is on how to think about properties of natural abstractions. A significant chunk of uncertainty is also on the possibility that natural abstraction is the wrong way to think about the topic altogether, although in that case I'd still assign a reasonable chance that neural networks end up with simple embeddings of human values - after all, no matter how we frame it, they definitely have trivial embeddings of many other complicated high-level objects.

## Aside: Microscope AI

[Microscope AI](#) is about studying the structure of trained neural networks, and trying to directly understand their learned internal algorithms, models and concepts. In light of the previous section, there's an obvious path to alignment where there turns out to be a few neurons (or at least some simple embedding) which correspond to human values, we use the tools of microscope AI to find that embedding, and just like that the alignment problem is basically solved.

Of course it's unlikely to be that simple in practice, even assuming a simple embedding of human values. I don't expect the embedding to be quite as simple as one neuron activation, and it might not be easy to recognize even if it were. Part of the problem is that we don't even know the type signature of the thing we're looking for - in other words, there are unanswered fundamental conceptual questions here, which make me less-than-confident that we'd be able to recognize the embedding even if it were right under our noses.

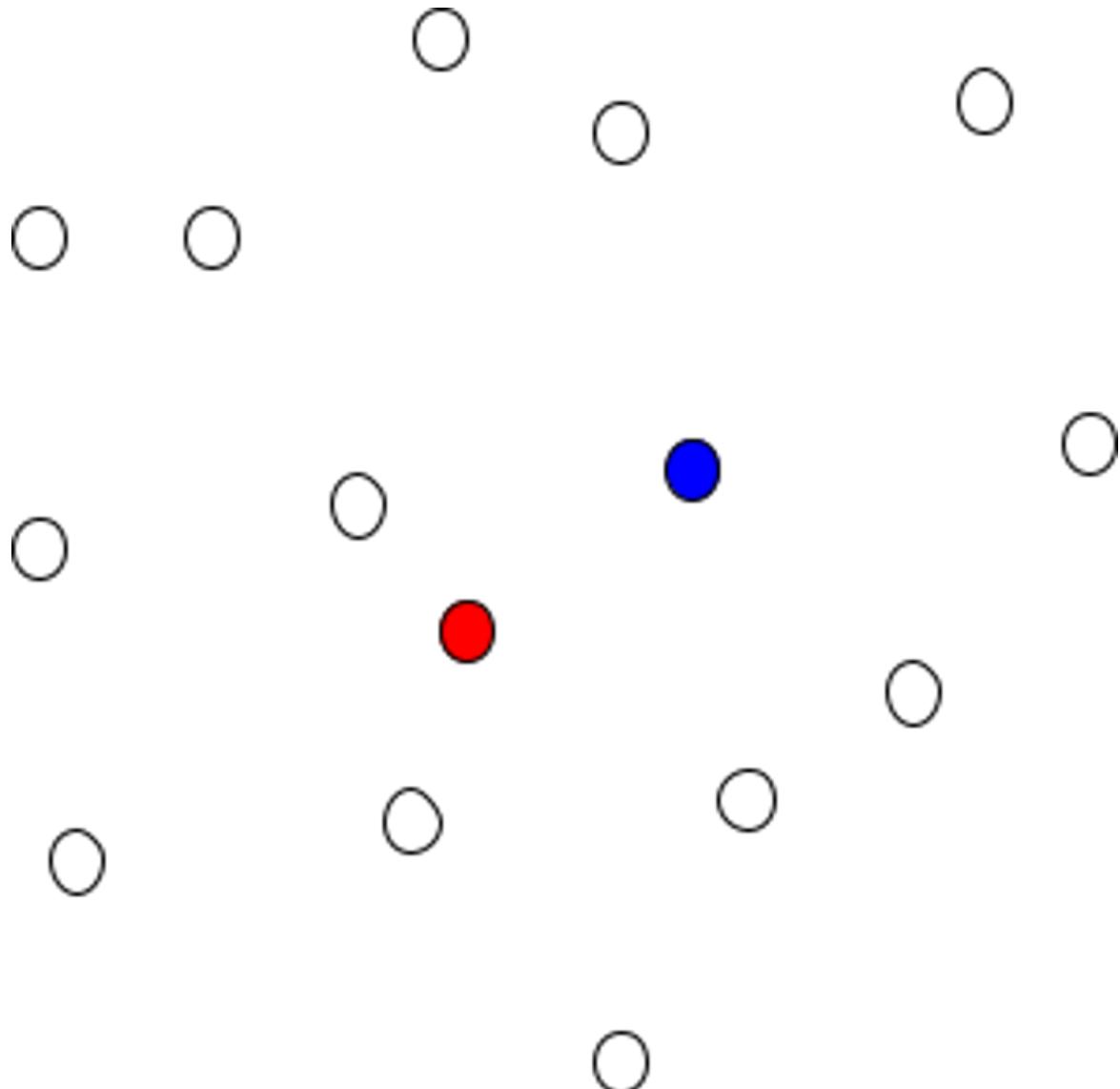
That said, this still seems like a reasonably-plausible outcome, and it's an approach which is particularly well-suited to benefit from marginal theoretical progress.

One thing to keep in mind: this is still only about aligning *one* AI; success doesn't necessarily mean a future in which more advanced AIs remain aligned. More on that later.

## Supervised/Reinforcement: Proxy Problems

Suppose we collect some kind of data on what humans want, and train a system on that. The exact data and type of learning doesn't really matter here; the relevant point is that any data-collection process is always, no matter what, at best a proxy for actual human values. That's a problem, because [Goodhart's Law](#) plus [Hidden Complexity of Wishes](#). You've probably heard this a hundred times already, so I won't belabor it.

Here's the interesting possibility: **assume the data is crap**. It's so noisy that, even though the data-collection process is just a proxy for real values, the data is consistent with real human values. Visually:



*Real human values are represented by the blue point, and the true center of our proxy measure is the red point. In this case, the data generated (other points) is noisy enough that it's consistent with real human values. Disclaimer: this is an analogy, I don't actually imagine values and proxies being directly represented in the same space as the data.*

At first glance, this isn't much of an improvement. Sure, the data is *consistent* with human values, but it's consistent with a bunch of other possibilities too - including the real data-collection process (which is exactly the proxy we wanted to avoid in the first place).

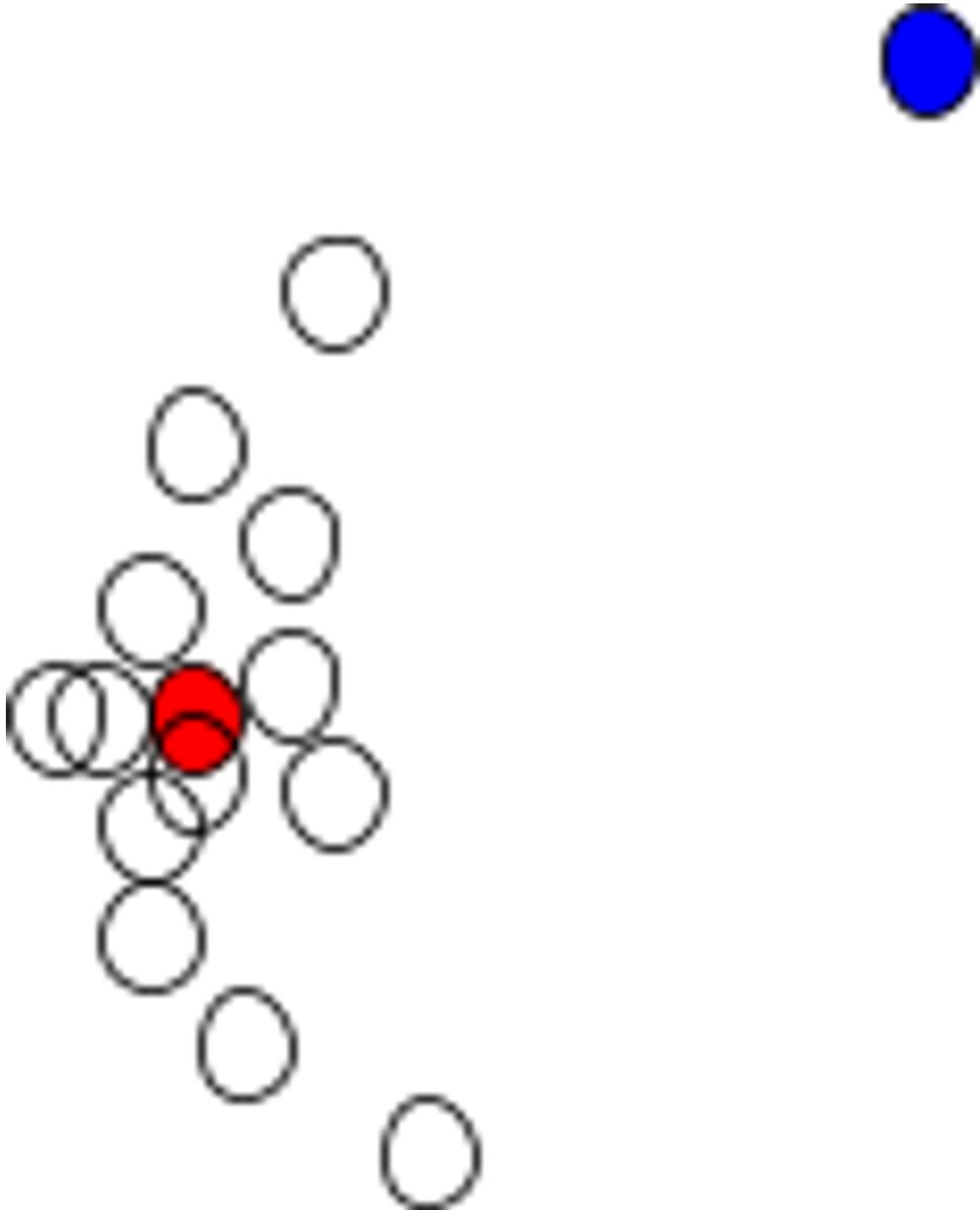
But now suppose we do some transfer learning. We start with a trained unsupervised learner, which already has a simple embedding of human values (we hope). We give our supervised learner access to that system during training. **Because the unsupervised learner has a simple embedding of human values, the supervised learner can easily score well by directly using that embedded human values model.** So, we cross our fingers and hope the supervised learner just directly uses that embedded human values model, and the data is noisy enough that it never "figures out" that it can get better performance by directly modelling the data-collection process instead.

In other words: **the system uses an actual model of human values as a proxy for our proxy of human values.**

This requires hitting a window - our data needs to be good enough that the system can tell it should use human values as a proxy, but bad enough that the system can't figure out the specifics of the data-collection process enough to model it directly. This window may not even exist.

(Side note: we can easily adjust this whole story to a situation where we're training for some task other than "satisfy human values". In that case, the system would use the actual model of human values to model the Hidden Complexity of whatever task it's training on.)

Of course in practice, the vast majority of the things people use as objectives for training AI probably wouldn't work at all. I expect that they usually look like this:



In other words, most objectives are so bad that even a little bit of data is enough to distinguish the proxy from real human values. But if we assume that there's some try-it-and-see going on, i.e. people try training on various objectives and keep the AIs which seem to do roughly what the humans want, then it's *maybe* plausible that we end up iterating our way to training objectives which "work". That's assuming things don't go irreversibly wrong before then - including not just hostile takeover, but even just development of deceptive behavior, since this scenario does not have any built-in mechanism to detect deception.

Overall, I'd give maybe a 10-20% chance of alignment by this path, *assuming* that the unsupervised system does end up with a simple embedding of human values. The main failure mode I'd expect, assuming we get the chance to iterate, is deception - not necessarily "intentional" deception, just the system being optimized to look like it's working the way we want rather than actually working the way we want. It's the proxy problem again, but this time at the level of humans-trying-things-and-seeing-if-they-work, rather than explicit training objectives.

## Alignment in the Long Run

So far, we've only talked about one AI ending up aligned, or a handful ending up aligned at one particular time. However, that isn't really the ultimate goal of AI alignment research. What we really want is for AI to remain aligned in the long run, as we (and AIs themselves) continue to build new and more powerful systems and/or scale up existing systems over time.

I know of two main ways to go from aligning one AI to long-term alignment:

- Make the alignment method/theory very reliable and robust to scale, so we can continue to use it over time as AI advances.
- Align one roughly-human-level-or-smarter AI, then use that AI to come up with better alignment methods/theories.

The alignment-by-default path relies on the latter. Even assuming alignment happens by default, it is unlikely to be highly reliable or robust to scale.

That's scary. We'd be trusting the AI to align future AIs, without having any sure-fire way to know that the AI is itself aligned. (If we did have a sure-fire way to tell, then that would itself be most of a solution to the alignment problem.)

That said, there's a bright side: when alignment-by-default works, it's a best-case scenario. The AI has a basically-correct model of human values, and is pursuing those values. Contrast this to things like IRL variants, which *at best* learn a utility function which approximates human values (which are probably not themselves a utility function). Or the HCH family of methods, which *at best* mimic a human with a massive hierarchical bureaucracy at their command, and certainly won't be any more aligned than that human+bureaucracy would be.

To the extent that alignment of the successor system is limited by alignment of the parent system, that makes alignment-by-default potentially a more promising prospect than IRL or HCH. In particular, it seems plausible that imperfect alignment gets amplified into worse-and-worse alignment as systems design their successors. For instance, a system which tries to look like it's doing what humans want rather than actually doing what humans want will design a successor which has even better human-deception capabilities. That sort of problem makes "perfect" alignment - i.e. an AI actually pointed at a basically-correct model of human values - qualitatively safer than a system which only manages to be not-instantly-disastrous.

(Side note: this isn't the only reason why "basically perfect" alignment matters, but I do think it's the most relevant such argument for one-time alignment/short-term methods, especially on not-very-superhuman AI.)

In short: **when alignment-by-default works, we can use the system to design a successor without worrying about amplification of alignment errors**. However, we wouldn't be able to tell for sure whether alignment-by-default had worked or not, and it's still possible that the AI would make plain old mistakes in designing its successor.

## Conclusion

Let's recap the bold points:

- A low-level model of some humans has everything there is to know about human values embedded within it, in exactly the same way that human values are embedded in physical humans. The embedding, however, is nontrivial. Thus...
- Predictive power alone is not sufficient to define human values. The missing part is the embedding of values within the model. However...
- This also applies if we replace the phrase “human values” with “trees”. Yet we have a whole class of neural networks in which a simple embedding lights up in response to trees. Why?
- Trees are a natural abstraction, and we should expect to see real systems trained for predictive power use natural abstractions internally.
- Human values are a little different from trees (they’re a property of an abstract object rather than an abstract object themselves), but I still expect that a broad class of systems trained for predictive power will end up with simple embeddings of human values (~70% chance).
- Because the unsupervised learner has a simple embedding of human values, a supervised/reinforcement learner can easily score well on values-proxy-tasks by directly using that model of human values. In other words, the system uses an actual model of human values as a proxy for our proxy of human values (~10-20% chance).
- When alignment-by-default works, it’s basically a best-case scenario, so we can safely use the system to design a successor without worrying about amplification of alignment errors (among other things).

Overall, I only give this whole path ~10% chance of working in the short term, and maybe half that in the long term. However, *if* amplification of alignment errors turns out to be a major limiting factor for long-term alignment, then alignment-by-default is plausibly more likely to work than approaches in the IRL or HCH families.

The limiting factor here is mainly identifying the (probably simple) embedding of human values within a learned model, so microscope AI and general theory development are both good ways to improve the outlook. Also, in the event that we are able to identify a simple embedding of human values in a learned model, it would be useful to have a way to translate that embedding into new systems, in order to align successors.

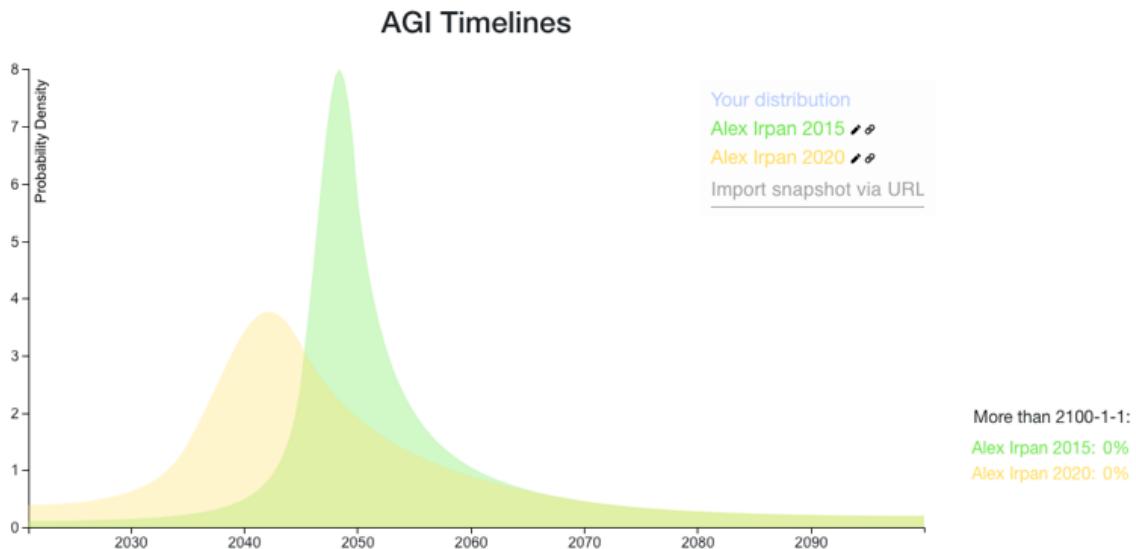
# Forecasting Thread: AI Timelines

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

This is a thread for displaying your timeline until human-level AGI.

Every answer to this post should be a forecast. In this case, a forecast showing your AI timeline.

For example, [here](#) are Alex Irpan's AGI timelines.



The green distribution is his prediction from 2015, and the orange distribution is his 2020 update (based on [this post](#)).

For extra credit, you can:

- Say why you believe it (what factors are you tracking?)
- Include someone else's distribution who you disagree with, and speculate as to the disagreement

## How to make a distribution using Elicit

1. **Go to [this page](#).**
2. **Enter your beliefs in the bins.**
  1. Specify an interval using the Min and Max bin, and put the probability you assign to that interval in the probability bin.
  2. For example, if you think there's a 50% probability of AGI before 2050, you can leave Min blank (it will default to the Min of the question range), enter 2050 in the Max bin, and enter 50% in the probability bin.
  3. The minimum of the range is January 1, 2021, and the maximum is January 1, 2100. You can assign probability above January 1, 2100 (which also includes 'never') or below January 1, 2021 using the Edit buttons next to the graph.
3. **Click 'Save snapshot,' to save your distribution to a static URL.**
  1. A timestamp will appear below the 'Save snapshot' button. This links to the URL of your snapshot.

2. Make sure to copy it before refreshing the page, otherwise it will disappear.
4. **Copy the snapshot timestamp link and paste it into your LessWrong comment.**
  1. You can also add a screenshot of your distribution using the instructions below.

### How to overlay distributions on the same graph

1. Copy your snapshot URL.
2. Paste it into the *Import snapshot via URL* box on the snapshot you want to compare your prediction to (e.g. [the snapshot of Alex's distributions](#)).
3. Rename your distribution to keep track.
4. Take a new snapshot if you want to save or share the overlaid distributions.

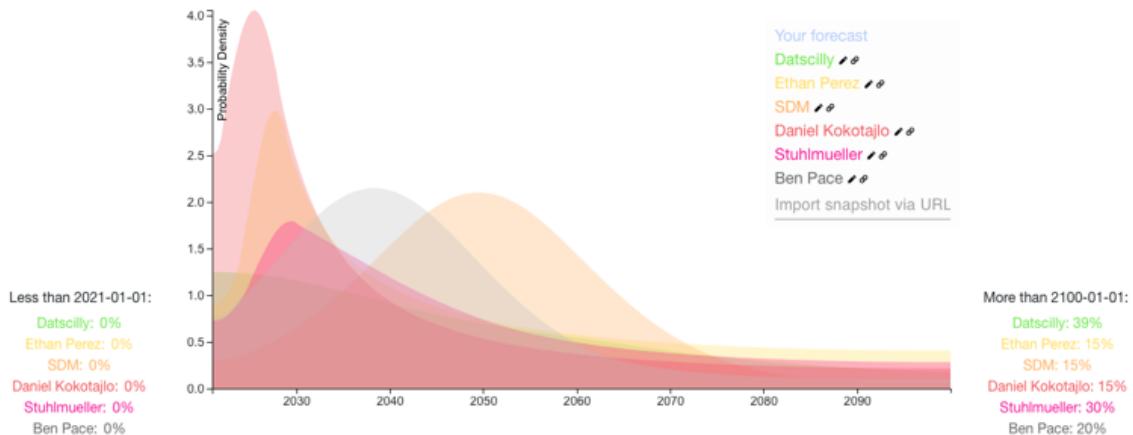
### How to add an image to your comment

- Take a screenshot of your distribution
- Then do one of two things:
  - If you have beta-features turned on in your account settings, drag-and-drop the image into your comment
  - If not, upload it to an image hosting service, then write the following markdown syntax for the image to appear, with the url appearing where it says 'link': ![] (link)
- If it worked, you will see the image in the comment *before* hitting submit.

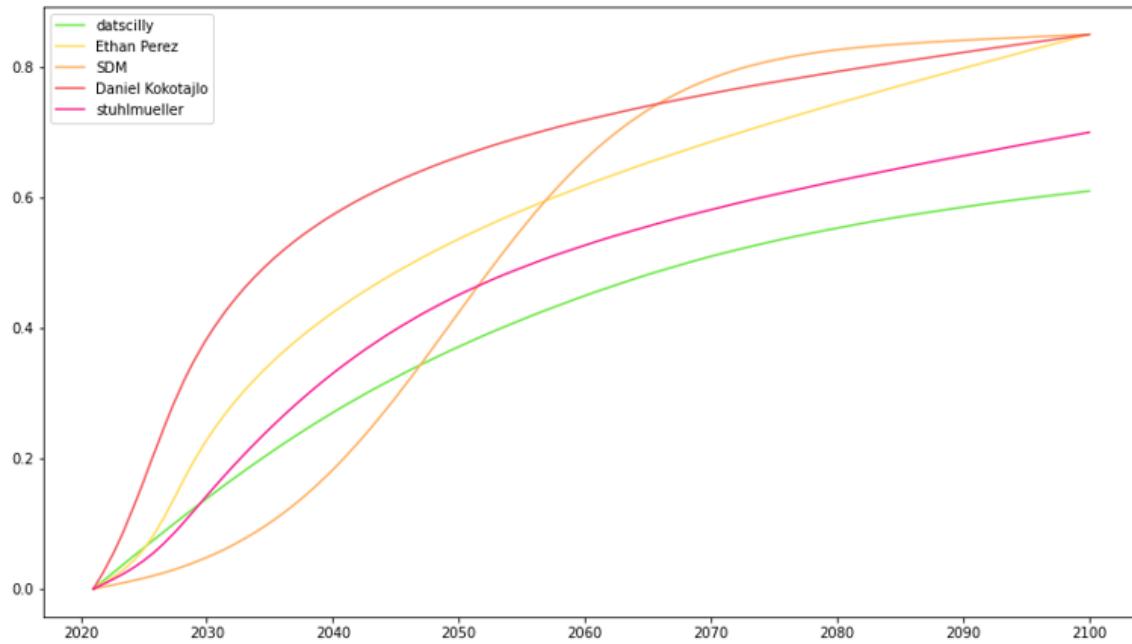
If you have any bugs or technical issues, reply to Ben ([here](#)) in the comment section.

### Top Forecast Comparisons

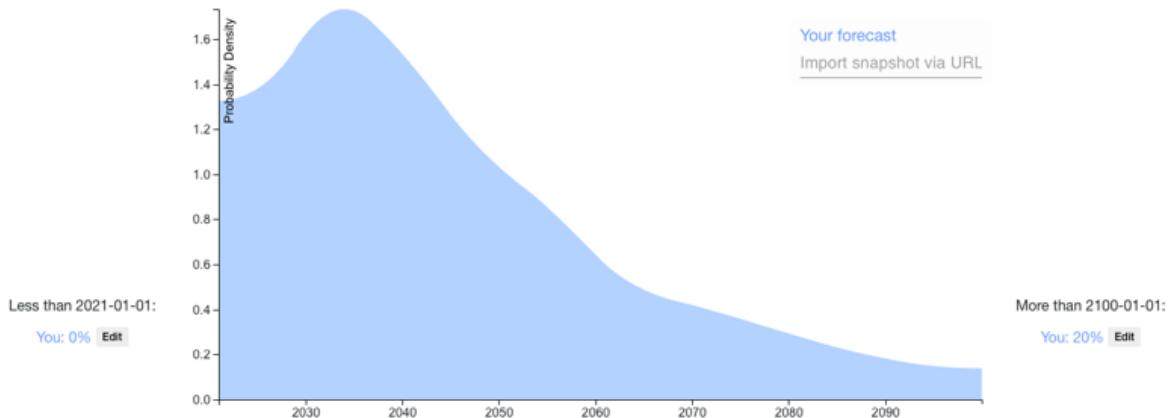
[Here](#) is a snapshot of the top voted forecasts from this thread, last updated 9/01/20. You can click the dropdown box near the bottom right of the graph to see the bins for each prediction.



Here is a comparison of the forecasts as a CDF:



[Here](#) is a mixture of the distributions on this thread, weighted by normalized votes (last updated 9/01/20). The median is **June 20, 2047**. You can click the Interpret tab on the snapshot to see more percentiles.



# My Understanding of Paul Christiano's Iterated Amplification AI Safety Research Agenda

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

*Crossposted from the [EA forum](#)*

*You can read this post as a [google docs](#) instead (IMO much better to read).*

This document aims to clarify the AI safety research agenda by Paul Christiano (IDA) and the arguments around how promising it is.

**Target audience:** All levels of technical expertise. The less knowledge about IDA someone has, the more I expect them to benefit from the writeup.

**Writing policy:** I aim to be as clear and concrete as possible and wrong rather than vague to identify disagreements and where I am mistaken. Things will err on the side of being too confidently expressed. Almost all **footnotes are content and not references**.

**Epistemic Status:** The document is my best guess on IDA and might be wrong in important ways. I have not verified all of the content with somebody working on IDA. I spent ~4 weeks on this and have no prior background in ML, CS or AI safety.

*I wrote this document last summer (2019) as part of my summer research fellowship at FHI. I was planning to restructure, complete and correct it since but haven't gotten to it for a year, so decided to just publish it as it is. The document has not been updated, i.e. **nothing that has been released since September 2019 is incorporated into this document**. Paul Christiano generously reviewed the first third to a half of this summary. I added his comments verbatim in the document. Apologies for the loss of readability due to this. This doesn't imply he endorses any part of this document, especially the second half which he didn't get to review.*

## Purpose of this document: Clarifying IDA

IDA is **Paul Christiano's AI safety research agenda**.<sup>[1]</sup> Christiano works at OpenAI which is one of the main actors in AI safety and IDA is by many considered the **most complete**<sup>[2]</sup> AI safety agenda.

However, people who are not directly working on IDA are often **confused** about how exactly to understand the agenda. Clarifying IDA would make it **more accessible** for technical people to work on and easier to **assess** for nontechnical people who want to think about its **implications**.

I believe that there are currently no resources on IDA that are both **easy to understand** and give a **complete picture**. Specifically, the current main resources

are:

- the “[\*\*Iterated Amplification\*\*](#)” sequence which is a series of curated posts by Paul Christiano that can be quite **difficult to understand**,
- this [\*\*post by Ajeya Cotra\*\*](#) and this [\*\*video by Robert Miles\*\*](#) which are both easy to understand but **limited in scope** and don’t provide many details,
- [\*\*Alex Zhu’s FAQ\*\*](#) to IDA which clarifies important points but does not set them in **context** with the entire research agenda,
- an [\*\*80,000 podcast\*\*](#) with Paul Christiano which explains some intuitions behind IDA but is **not comprehensive** and is in **speech form**.

This document aims to fill the gap and give a **comprehensive** and **accessible overview of IDA**.

## Summary: IDA in 7 sentences

1. IDA stands for **Iterated Amplification** and is a **research agenda** by Paul Christiano from OpenAI.
2. IDA addresses the artificial intelligence (**AI**) **safety problem**, specifically the danger of creating a very powerful AI which leads to catastrophic outcomes.
3. IDA tries to prevent catastrophic outcomes by searching for a **competitive AI** that **never intentionally optimises for something harmful** to us and that we **can still correct** once it’s running.
4. IDA doesn’t propose a specific implementation, but presents a **rough AI design** and a collection of thoughts on whether this design has the potential to create safe and powerful AI and how the details of that design could look like.
5. The proposed AI design is to use a **safe but slow way of scaling up** an AI’s capabilities, **distill** this into a faster but slightly weaker AI, which can be scaled up safely again, and to iterate the process until we have a fast and powerful AI.
6. The most promising idea on how to slowly and safely scale up an AI is to give a weak and safe AI access to other weak and safe AIs which they can **ask questions** to enable it to solve more difficult tasks than it could alone.
7. It is **uncertain whether IDA will work** out at all (or in the worst case lead to unsafe AI itself), lead to useful tool AIs for humans to create safer AI in the future or develop into the blueprint for a single safe and transformative AI system.

## What problem is IDA trying to solve?

### The AI safety problem

IDA addresses one aspect of the **Artificial Intelligence (AI) safety** problem: Current AI systems make a lot of **mistakes** that are **poorly understood**. Currently, this is not a big problem since **contemporary AI systems are very limited**. They don’t make very influential decisions and the tasks they are solving are relatively well understood: it would be relatively **easy for a human to oversee** whether an AI system is doing a task correctly.

*Comment Christiano: IDA is targeted more narrowly at (intent) alignment---building a competitive AI which is trying to do what we want it to do, and is never trying to do something catastrophic. So it might be worth briefly mentioning that this is just one aspect of the safety problem and that even an aligned AI could contribute to*

*catastrophic outcomes (though this is the aspect which futurists and EAs most often talk about).*

However, in the **coming decades** we might want to limit the amount of human oversight and give AIs more autonomy over **more important tasks** that humans **understand less well**. At this point, unreliable AI systems that make a lot of mistakes might be **catastrophic**. One particular worry is that we might create an AI that **optimizes** for a task in a way that is **unexpected** and **harmful** to us and that we **cannot control**.

IDA is trying to find an **AI design** that ensures that an AI following this design **doesn't optimize for something harmful to us, doesn't make catastrophic mistakes** and **doesn't optimize in a way that takes control** from us.

*Comment Christiano: IDA isn't really intended to avoid catastrophic mistakes---just to avoid intentional catastrophic actions.*

This post **assumes familiarity** with the AI safety problem. If readers never heard *AI safety, AI risk, AGI, superintelligence, human level AI and AI alignment*, I [recommend reading, watching or listening to these resources \(and possibly more\)](#) before reading this post.

Other readers might still want to read Christiano's [description of specific AI risks he has in mind](#).

## Worldview behind IDA: We need to work on safe competitive and prosaic AI

IDA tries to create a **safe and powerful version** of a class of AI training methods that we already use, so-called **self-play methods**.<sup>[3]</sup> This approach is based on Christiano's belief that safe, powerful AI systems must be **competitive** and could potentially be **prosaic**.

*Comment Christiano: I wouldn't say that this is particularly related to self-play. I'm interested in competing with all of RL, and in being competitive with deep learning more broadly (I think it's plausible that e.g. deep imitation learning will be the way things go). Maybe "A safe and powerful version of deep learning, including deep reinforcement learning (RL)."*

A **prosaic AI** is an AI we can build without learning anything substantially new about intelligence. If powerful prosaic AI was impossible - that is, if substantial new insights into intelligence were necessary to create powerful AI - we would likely need entirely different AI techniques to create powerful AI. If we believe that powerful prosaic AI is possible, this makes **paying attention to current AI techniques** relatively more important.

Christiano also believes that any safe AI must be **competitive**. This means, that they can solve a given task approximately as well and efficient as any unsafe alternative AI that we could build. Otherwise, competitive pressures could incentivize the development and use of these unsafe AI systems.<sup>[4]</sup>

This motivates his belief that \*\*ideally, for every potentially unsafe AI technique that is currently used<sup>[5]</sup> **and will be developed in the future, we would have a**

**competitive safe version.**

## IDA builds on mainstream AI research to be competitive

Mainstream AI researchers recently made progress on a class of algorithms (**self-play methods**) that essentially consist of what Christiano calls an iteration of *distillation* and *amplification*. Christiano's research agenda Iterated Amplification (IDA) attempts to develop a version of this method that is **safe even when scaled up**, i.e. even when the AI system trained by this method is extremely powerful.

*Comment Christiano: I think expert iteration is orthogonal to self-play. For example, I think Mu Zero also uses this framework on single-player games. It just happened to be that AGZ and the early expert iteration paper were on board games.*

The agenda's contributions are to:

1. argue why iterating distillation and amplification can potentially be very powerful, (*comment Christiano: I wouldn't say I added much here, I think that people already had this covered.*)
2. explore how powerful it can get, (*comment Christiano: This should probably go after safety. If you use expert iteration to optimize a reward function then it won't be limited in this way, but it will be safe. Maybe A. describe a potentially safe way to apply IDA, B.*)
3. argue that this application of IDA may be able to get very powerful, ...
4. argue that this method can potentially be used safely,
5. discuss which conditions must be fulfilled for this to be the case, and
6. attempt to come up with concrete safe ways to implement it.

The term iterated distillation and amplification can be confusing since it refers both to a \*\*specific AI safety research agenda<sup>[6]</sup> and to a **general method of training AI**. When I refer to the safety agenda, I will say IDA. When I just use the words iterating, distillating or amplifying, I will refer to the general training method.

## What's the ideal outcome of IDA?

### What kind of AI do we want to create via IDA?

IDA is **fully successful** if it leads to an **implementation for safe powerful AI**, but IDA could also be **partially successful** by helping us to **develop useful tools** to build safe powerful AI.

### Full success: IDA finds an implementation for safe powerful AI

Ideally, we figure out how to do safe distillation and safe amplification and develop **one single AI system which autonomously does the distillation and amplification steps**, thereby **self-improves**, and keeps going until it becomes the **strongest possible** (or the strongest we will ever attain) **AI**.<sup>[7]</sup>

*Comment Christiano: This seems too strong. A small tweak might be to say something like: "Until it becomes the best AI that can be trained using available data,*

*computation, and prosaic AI methods" or something. The point is that it's as good as anything you could make via a direct application of ML.*

The process of iteration can take as long as years, decades or even millennia to complete but also be as short as days, minutes or seconds. In either case, our AI system hopefully **improves fast enough** to always be reasonably competitive to unsafe alternatives. If during or after the self-improvement process somebody develops and implements a more efficient AI training method, our efforts until then might have been in vain (apart from any generalizable insights generated from our work on IDA.)

*(Comment Christiano: Iteration should occur in parallel with ML training, so it takes exactly as long as ML training---no more, no less.)*

I presented one prototype ideal outcome from IDA, but there are at least two more: Instead of figuring out all of IDA and then applying it, we could 1) **apply IDA over and over again** with stronger and stronger distillation and amplification algorithms until we find something that works or 2) we apply IDA to create **a lot of domain specific narrow AIs**. In reality, probably some **mix** of these will happen or something entirely different. However, I think the main considerations for each case don't change too much.

*Comment Christiano: My hope would be that IDA is used every time someone wants to train a system with deep learning. My expectation would be that we will train many distinct systems with narrower competencies, because that seems like a better use of model capacity. But mostly the success is orthogonal to this: we've succeeded if we produce models that are competitive with the best AI people are making anywhere, and the aligned AIs we train will look basically the same as the unaligned AIs that we would have trained anyway. IDA is orthogonal to this kind of question about how AI ends up looking.*

## **Partial success: IDA finds a tool that improves our reasoning**

There is also the possibility of **partial success**: We might fail to develop a universally competent AI with IDA and instead develop an AI that is very capable at **a subset of tasks** that is important to AI safety. Example: We could create an AI system that is unable to physically manipulate things but really good (or at least better than us) at theorem proving. In that case, we developed a **tool to improve human reasoning** which could help to come up with a different design for safe powerful AI or solve other important problems.<sup>[8]</sup> Of course, there are also **dangers** from this kind of reasoning improvement which I will discuss more generally as *differential competence* in the section "[How IDA could fail](#)".

*Comment Christiano: I would personally describe this as: The aligned AIs we train may not be competitive---they may not fully utilize the techniques, data, and compute that is available---but they may still be good enough to help humans in important ways.*

## **What kind of safety is IDA aiming for?**

\_Disclaimer: Please note that this might not be how the terms safety, alignment and corrigibility are used outside of the context of IDA.<sup>[9]</sup>

## **IDA is aiming for intent alignment**

One important thing to note is what Christiano means by hopefully achieving “safe” AI via IDA. Christiano’s current work on the agenda is focussed on creating an AI that is **intent aligned** towards its current user and **reasonably competent** at inferring our preferences and acting on them.

**An intent aligned agent is an agent that is trying to do what we want it to do.** This definition is about the agent’s **intentions** rather than the outcomes and refers to “what we want it to do” **de dicto** rather than **de re**:

Example: Imagine our AI was tasked to paint our wall and we wanted the AI to paint it blue. The intent aligned AI’s goal in this example should be “paint the wall in the color the user wants” rather than “paint the wall blue”.<sup>[10]</sup> “Paint the wall blue” should only be an instrumental subgoal. An AI is intent aligned as long as its ultimate goal is the first, even if it is not competent enough to figure out the color we prefer (blue) or is not competent enough to actually paint a wall. Thereby, an intent aligned AI is different from the following types of aligned AI:

1. an AI that is trying to do what we want it to do **de re** (in the example that would be an AI that is trying to paint the wall blue),
2. an AI that is actually doing what we want it to do (actually painting the wall blue),
3. an AI that is trying to do the morally right thing (if that differs from the user’s preferences),
4. an AI that is actually doing the morally right thing.

None of the five ways an AI could be ‘aligned’ imply any of the other.

## **Intent aligned AIs need to be reasonably competent to be safe**

An intent aligned AI that is not **sufficiently competent** at inferring and working towards our preferences might still make catastrophic mistakes. If successful, IDA would thereby have to both solve intent alignment and the problem of creating a reasonably competent AI system. This AI might still make minor mistakes<sup>[11]</sup> but no catastrophic mistakes.

*Comment Christiano: Note that this safety condition depends not only on the AI but on how we use it---once we have a well-intentioned AI, we can either make it more competent or we can avoid deploying it in ways where its failures would be catastrophic (just as you would for a well-intentioned human assistant).*

Example: An AI that is reasonably competent might still make mistakes in the realm of “choosing a blue that’s a bit different from what we would have preferred” or “painting somewhat unevenly”. It would hopefully not make mistakes in the realm of “completely fails to paint and knocks down the house instead” or “perfectly paints the wall but in the process also plants a bomb”.

## **Why corrigibility is the preferred form of intent alignment**

IDA hopes to solve intent alignment and partly addresses the competence problem via **corrigibility**. [Christiano’s intuition](#) (April 2018) is that **corrigibility implies intent alignment** but that an intent aligned AI is not necessarily corrigible:

*Suppose that I give an indirect definition of "my long-term values" and then build an AI that effectively optimizes those values. Such an AI would likely disempower me in the short term, in order to expand faster, improve my safety, and so on. It would be "aligned" but not "corrigible."*

*Comment Christiano [on corrigibility implying intent alignment]: I don't really think this is true (though I may well have said it). Also this implication isn't that important in either direction, so it's probably fine to cut? The important things are: (i) if I want my agent to be corrigible than an intent-aligned agent will try to be corrigible, (ii) if being corrigible is easy then it is likely to succeed, (iii) corrigibility seems sufficient for a good outcome.*

A corrigible agent is one that always **leaves us in power** in some way even if disempowering us was instrumental to achieving our long-term goals. In that sense, **an optimal corrigible AI is worse than an optimal intent aligned AI that is not corrigible**: The latter has the freedom to disempower us if that's what's best for us in the long-run. However, it also seems that it would be extremely hard to judge whether a non-corrigible AI is actually intent aligned or to fix catastrophic bugs it has. Corrigibility is still a fuzzy concept and does not have a clear, formal definition. [\[12\]](#)

## How to achieve corrigibility

IDA tries to achieve corrigibility by designing **approval-directed AI**: It only takes actions that it imagines the user would approve of (including actions the AI could hide) and is **act-based**. An act-based AI is one that considers the user's **short-term preferences**.

*Comment Christiano [on approval-directed AI]: Maximizing the overseer's approval is my preferred approach to distillation [...] It takes actions that the overseer would approve of, but the overseer is going to be a (human + AI team)---an unaided human won't be able to understand what the AI is doing at all.*

Details: "Short-term preferences" does not mean "preferences over what can be achieved in the short-term" but "things that I would still endorse in the short-term": An AI system that is optimizing for short-term preferences could still do **long-term planning** and act accordingly, but it does so because that's **what the human wants it to do at that moment**. If we changed our mind, the AI would also change action. It would not take actions that we oppose even if 10 years later we would think it would have been good in hindsight. The logic here is that the user always knows better than the AI.

The idea with approval-directed agents is that 1) since the AI system values the user's true approval, they would have an incentive to clarify their preferences. This makes **value of information (VOI)** and **updating on this information** central to the AI. 2) It is hopefully relatively easy for an AI system to infer that in the short term, the user would probably not approve of actions that disempower the user. This only works if the user doesn't want to be disempowered in the moment.

*Comment Christiano: [on 1] This isn't quite how I think about it (probably fine to cut?) [on 2] This is the key thing. The other important claim is that if you want to be **weaker** than the overseer, then approval-direction doesn't limit your capabilities because you can defer to the overseer's long-term evaluations. So maximizing approval may be acceptable as part of IDA even if it would be too limiting if done on its own.*

## Corrigibility and intent alignment might be easier than other forms of alignment

IDA is aiming for corrigibility because of a number of **claims**. They are all not required to make corrigibility work but inform why IDA thinks corrigibility is more promising than other forms of alignment:

1. Corrigibility **alleviates** the problem that the agent has to be **competent** at inferring the user's preferences: the AI hopefully does not need to be terribly good at understanding humans to understand that they probably would not approve of being killed, lied to or ignored.
2. Corrigibility is **stable**: This means an agent that is partly corrigible will become more corrigible over time as we scale it up.
3. Corrigibility has a **broad basin of attraction**: While perfectly inferring all human preferences is probably a very narrow target, it is hopefully \*\*relatively easy<sup>[13]</sup> to learn corrigibility sufficiently well to gravitate towards a center of corrigibility and not do anything catastrophic along the way.

In principle, you can also imagine an AI system trained by iterating distillation and amplification which is aligned but not corrigible. However, **IDA centers around the idea of corrigibility** and Paul Christiano is quite pessimistic that it would work if corrigibility turned out to be impossible or extremely difficult.

In conclusion, If IDA is successful, we can create a **powerful AI system** that is **corrigible** towards its current user and **reasonably competent** at figuring out what the user would approve of in the **short term**. This does not guarantee that the AI system will not make any **mistakes** in inferring the user's preferences and trying to satisfy them, nor does this guarantee that the AI will do good things since the **user might be malign**.

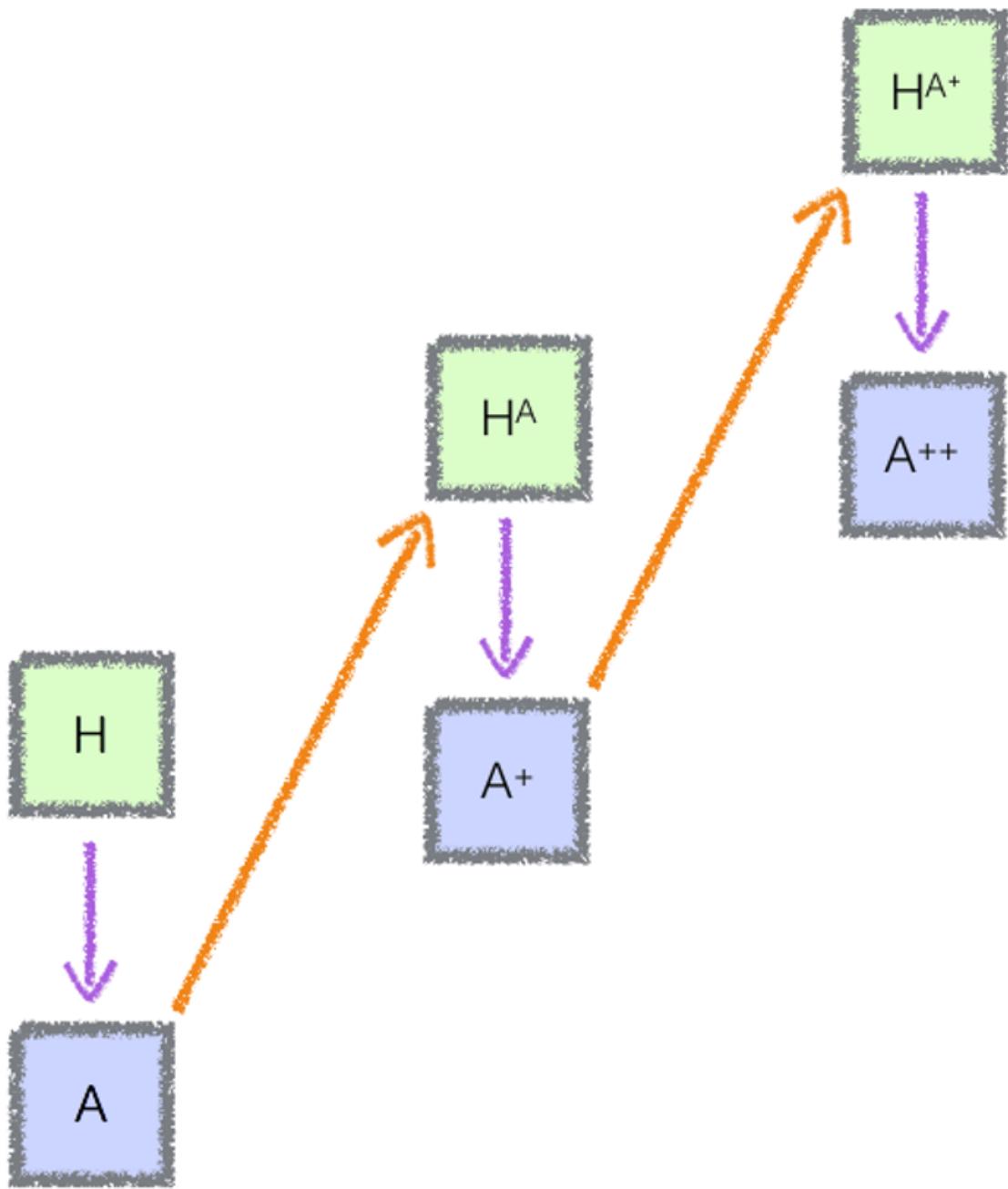
## How does iterating distillation and amplification work?

### A general scheme to train AI systems

#### Basic definition of iterated distillation and amplification

**Amplification is scaling up** a system's capabilities and **distillation is learning from the amplified system**. (*Comment Christiano: I don't mean to define a thing so general as to be applicable beyond ML.*)

Iterated distillation and amplification starts with a system that has some ability of level X. Some **amplification process increases** that system's ability level to X+a, possibly at the expense of **greater compute**. Another system learns the amplified system's ability through a **distillation process** which decreases the ability level to X+a-d, possibly at the advantage of needing **less compute**. If we amplify this system again, we get a system with ability level X+a-d+a2. When this process repeats many times, it is **iterated** (see figure 1). In principle, that is all that is needed to define something as iterated distillation and amplification.



*Figure 1.*

### **Specifics of iterated distillation and amplification that are useful in AI systems**

There are some specifics to doing **iterated distillation and amplification in AI**:

1. The systems with amplify and distill are (or include) **AI systems**.
2. Any single iteration of amplification and distillation is only useful if  $a > d$ , i.e. when we **gain more capabilities from the amplification step than we lose**

**from the distillation step.** (This is under the assumption that creating a system of ability level X is cheaper than doing amplification and distillation.)

3. We probably only want to do distillation if the amplified system is more **computationally expensive** than the distilled system, otherwise the distilled system is just a strictly worse version. (Unless distillation is the cheapest way of creating approximate copies of the amplified system.)
4. The system we distill into can be **the same AI** that we amplified before. In that case, we have the same AI whose abilities we iteratively increase at computational cost and then decrease for computational advantages (see figure 2).
5. If we have **repeatable** amplification and distillation steps (you can apply the amplification and distillation steps over and over again) it is relatively easier to **automate** them, i.e. have an AI do them. This likely increases **efficiency**.

*Comment Christiano: This is a bit confusing---you don't really need an AI to do the process, amplify( $X$ ) and distill( $X$ ) are both just programs that make calls to  $X$ , so you can just run them ad infinitum.*

6. If we combine point 2), 4) and 5), we have a **self-improving** AI system.

*Comment Christiano: I think this is accurate but it's a bit confusing since it differs from the way self-improvement is usually used in the AI safety community (and it's not an expression that is usually used in the ML community).*

7. A self-improving AI that identifies and employs better distillation and distillation step is **recursively self-improving**.

*Comment Christiano: This is true but it's not a distinguished form of recursive self-improvement---your AI may also build new computers or change your optimization algorithms or whatever. If the method succeeds the AI will be doing **more** of those things, since the whole point is that you design some scalable amplification + distillation process that doesn't require ongoing work.*



*r*



Figure 2.

## Example of iterated distillation amplification in a real AI system: AlphaGo Zero

Note: Skip this section if you're already familiar with how AlphaGo Zero roughly works.

[\*\*AlphaGo Zero\*\*](#) (**AGZ**) is an AI system that beat AlphaGo in Go which in turn beat the world's best Go player in Go. It was trained **without any data from human games** by basically **iterating amplification** (in this case Monte-Carlo tree search) and **distillation** (in this case reinforcement learning). I will give a simplified explanation of AGZ as an example for iterated distillation and amplification.

### A mathematical way of solving Go is impossible

To decide on a Go move, AGZ could go through every possible move it could take, then go through every possible way its Go opponent could react to its move and then go through every possible way AGZ could react to that etc. This way, it could play out all possible Go games with itself and choose the move that yields the best consequences. This technique would be very powerful and enable AGZ to find the **ground truth**, i.e. the *actual* optimal move. The problem with this approach is that it is **computationally intractable**, as there are so many possible moves that going through all of them is **impossible**.

### AGZ iterates “manually” going through moves and learning from the moves it tried out

What AGZ does instead is starting with some **weak policy**, e.g. “decide at random”, to **select moves to try out**. It then uses some weak policy, e.g. also “decide at random”, to go through some **possible responses** from its opponent and then to go through its **next own moves** etc. This way, AGZ can play out entire Go games against itself. It can only do so for a very **limited number of moves** and reactions to moves because this process is extremely **expensive**, so we only cover a **tiny fraction of the vast possible search space**.

The AGZ that settles on the move that yielded the **best outcome** of the moves it tried out, is probably more powerful than the initial AGZ that could only choose moves with the help of a very weak policy, e.g. “decide at random”. The better AGZ is the **amplified version** of the old AGZ, but it is also a lot slower. It can **never become very good** at the game because it **improves very slowly**.

But AGZ can take the Go games it played against itself and use them as **training data** to find a better policy than “select moves to try out at random”. It develops an **intuition for which moves are better** to take. The process of using the training data to replace expensive search with heuristics for good moves is **the distillation** in AGZ. An AGZ that takes moves merely based on these heuristics is the **distilled version** of the amplified AGZ. It makes **worse moves** but also is a lot **faster**.

We can now take the distilled AGZ and amplify it again. We use the distilled AGZ's **improved policy for selecting moves** to go through possible moves and possible reactions and our reactions to these reactions etc. The improved policy **narrows down** our search space to **better moves**. The resulting system is the **new amplified AGZ**.

We can **repeat** this process many, many times until we have a **distilled AGZ** that is fast and actually **good at Go**.

For a visual explanation of AGZ I recommend Robert Miles' [video](#).

## Testing your understanding: Clarifying examples of what are and aren't distillation and amplification processes

*Note: This section is for people who want to make sure they **fully understand** what **amplification and distillation** mean, but it is **not necessary** to do so to understand IDA.*

AGZ is only **one possible example** of iterated distillation and amplification. I will give some technical and nontechnical examples of what I would count and not count as amplification and distillation:

### Examples for amplification and distillation

*Technical examples for amplification and distillation*

- Gradient descent (amplification)

*Nontechnical examples for amplification and distillation*

- Sending a draft around, getting feedback and incorporating it several times (Amplifying your draft and hopefully your writing skills)
- Watching a karate master and trying to imitate them (distilling the master's ability, but amplifying your ability)
- Coming up with new dance moves (no distillation, possibly amplifying your dance skills)

## Why should we think iterating distillation and amplification could lead to powerful and safe AI?

*Note: a lot of the terminology used in this section is specific to IDA and might have a different meaning in a different context.*

Christiano believes that iterating amplification and distillation is promising because 1) he sees a **concrete way of doing amplification that could safely scale** to **powerful AI** of which the idealized version is **HCH** (which is the abbreviation for the recursive name "*Human consulting HCH*") and 2) he thinks that creating a fast powerful agent by distilling from a more powerful agent is **potentially safe**.[\[14\]](#)

## HCH is a proof of concept for a powerful and safe amplification process

Suitable amplification steps for IDA must have **no** (or very high) **upper bounds** in terms of capability and **preserve corrigibility**. IDA argues that **HCH** is an

amplification that has these properties.

## **HCH is amplifying a human by giving them many more humans they can order to help them solve a task**

**HCH** is the output of a process of humans solving a task by **recursively breaking the task into easier subtasks** and assigning them to other humans until the subtasks can be solved directly. This can be in the **infinite**. **HCH** is an unattainable **theoretical ideal point**. Showing that HCH would be powerful and corrigible would be a **proof of concept** for strong and safe AI systems that are amplified using similar methods.

Example: Imagine a human Rosa who tries to solve a very difficult task which she is unable to solve within the working memory and life span she has available. Rosa has access to a **large number of copies** of herself (Rosa') and can task them to solve parts of the problem. These subproblems might still be too hard for Rosa' to solve, but luckily each Rosa' also has access to a large number of yet more copies of Rosa (Rosa'') which they can task with subtasks of the subtask they received. Each Rosa'' now also has access to a bunch of Rosa copies etc.

## Structure of HCH

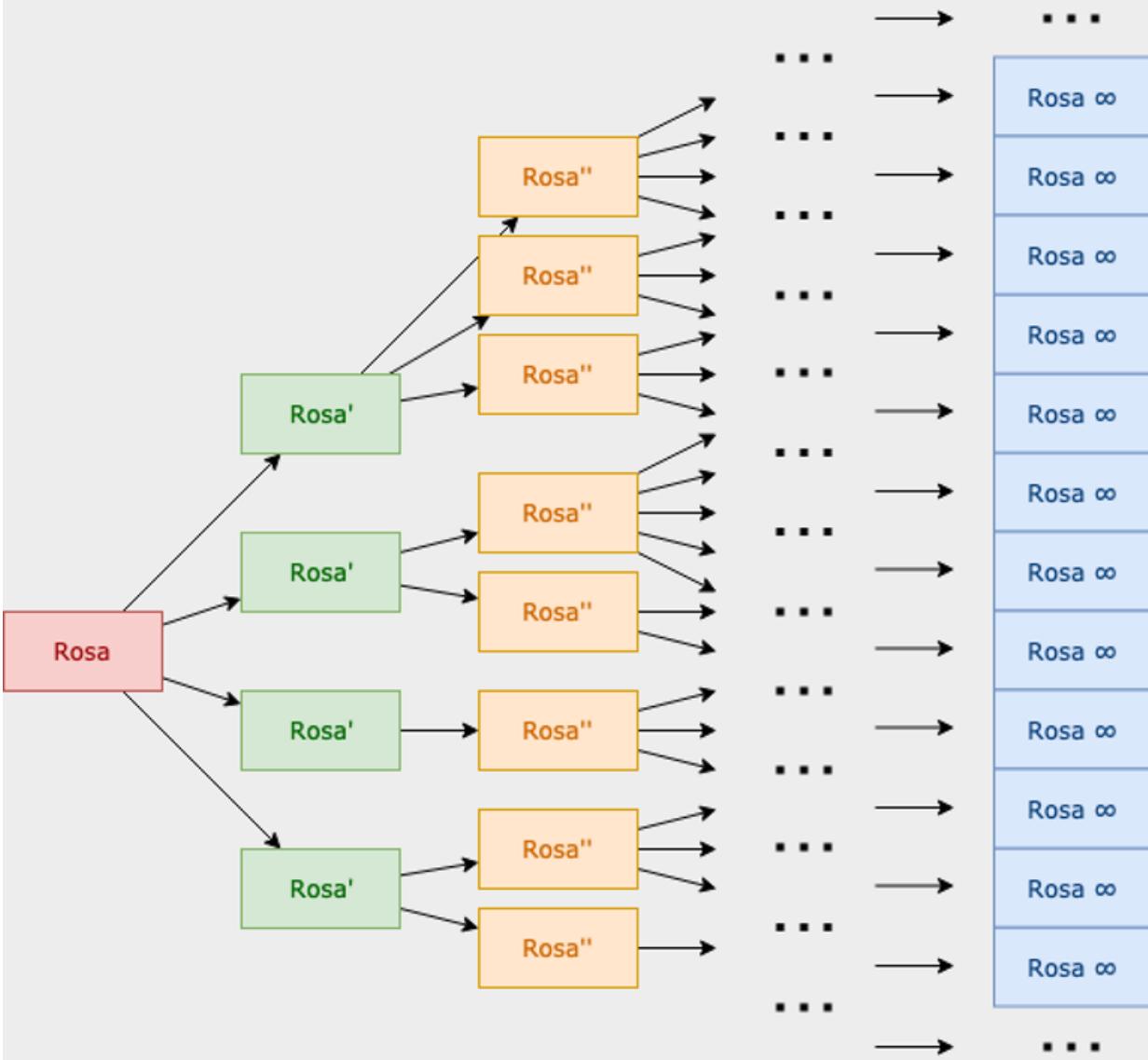


Figure 3.

In addition, the messages sent between the different copies of Rosa and Rosa can contain **pointers**. Example: If Rosa copy A (anywhere in the tree) has some insight that they shared with Rosa copy B, then Rosa copy B can share that insight in any subsequent message by pointing at Rosa copy A.

The information from solving the lowest level tasks is **locally integrated** and passed on to higher levels of the HCH tree etc. Example: Rosa'' integrates all of the information she gets from her Rosa'''s to solve the subtasks she was assigned to. She then passes that solution to Rosa''. There are **many** Rosa's that report to Rosa''. Rosa'' integrates all the information she gets to solve her subtasks and passes that solution to Rosa'. There are many Rosa's that report to Rosa'. Rosa' integrates all the information she gets from them so solve the overall task.

This process ends when the task is solved or when creating more copies of Rosa would not change the output anymore. HCH is a **fixed point**. HCH does not depend on the starting point (“What difficulty of task can Rosa solve on her own?”), but on the **task difficulty** the reasoning process (= breaking down tasks to solve them) can solve *in principle*.

The general process of solving a task by recursively breaking it into subtasks and solving these subtasks is called **factored cognition** and is a way of amplification. Factored cognition does not need to be done by humans like in HCH, but looking at humans might give us important clues about how promising it is.

## HCH might be powerful and safe

IDA builds on two intuitions about HCH: First, that there is some way of dividing tasks into subtasks, such that HCH can **solve any relevant task** (for example solve AI safety or be as strong as any unsafe AI we could ever create). This is based on the intuition that at least some humans (such as Rosa) are above some **universality threshold**, such that if given sufficiently much time, they could solve any relevant task.

The second intuition is that HCH would be **corrigible**. IDA is optimistic about this since HCH starts from human reasoning and therefore, the way of breaking down tasks in HCH is human understandable and approved.

If these intuitions about HCH are true, doing **amplification via factored cognition** seems like a **promising approach to safely scale up an AI system** to become very powerful since HCH is the result of doing factored cognition with humans and infinite computation. [15]

There is an obvious **difference between HCH and any real-world AI system**: HCH talks about **humans rather than an AI system**. Even if HCH was shown to be very powerful and safe, we might not be able to implement safe and powerful factored cognition in an AI. On the other hand, AI systems might also be able to do types of factored cognition that are impossible for humans.

## We still need safe distillation because pure HCH is intractable

HCH is the result of a potentially infinite exponential process (see figure 1) and thereby, **computationally intractable**. In reality, we can not break down any task into its smallest parts and solve these subtasks one after another because that would take too much computation. This is why we need to iterate **distillation** and amplification and cannot just amplify.

Similar to AlphaGo Zero, we can amplify a system that can only solve very small subtasks by letting it do very **fine-grained factored cognition** to solve somewhat more difficult tasks. The system solves a lot of tasks this way and we use these solutions as training data to distill a new system that can solve slightly bigger subtasks. We amplify this system by letting it do **less fine-grained factored cognition**.

## **Amplified systems can act as overseers and improve our chances of safe distillation**

We could potentially use an improved version of a current **deep learning technique** for distillation. These are the precise techniques Christiano worries cannot be safely scaled up. IDA is optimistic about distillation in this context though because we can use the last **amplified AI system as an overseer** for the next distillation.

Usually, we apply deep learning to **directly teach the AI the skill we eventually want it to** have. However, this might become quite **dangerous to scale up** if we eventually leap from “no capabilities” to “very strong” in one go. This is especially worrying if we want our AI to learn something **important** that **no human is smart enough to understand**.

In IDA we only ever **distill from a stronger trusted system** (amplified AI) to a weaker system. The amplification does the actual scaling up of capabilities, the distillation just makes things faster and more efficient. This has **two potential advantages**:

First, the amplified AI can act as an **overseer** that provides feedback to the distilled AI in training in order to both align it and improve its competence.<sup>[16]</sup> The hope is that oversight is easier when the **overseer is stronger than the system in training**.

Remember though that in factored cognition the amplification comes from having many copies of a weak subsystem that work on easy subtasks to solve a difficult task, and that we are trying to distill an AI that is **stronger than each individual subsystem**. To be optimistic about this, we must assume that a bunch of weak systems working together can actually oversee a strong system.

Second, since the distilled AI doesn't have to be as strong as the amplified AI (but just stronger than the last distilled AI), we might be able to use deep learning techniques that **aren't quite as powerful but in exchange safer**.

## **Which mechanisms ensure that amplification and distillation are actually safe?**

To be competitively efficient and powerful, we will probably need to design an AI system that is (recursively) **self-improving**, i.e. autonomously doing the amplification and distillation with itself,<sup>[17]</sup> and eventually **smarter than humans**. How can we design an AI that we can rely on to do good distillation and amplification even if it impossible for us to understand?

We need to find a combination of amplification and distillation procedures that **preserves corrigibility**. That probably means that we need to find an amplification procedure and a distillation procedure that preserve corrigibility both **on average** and **in worst-case scenarios**.

# How to make sure the AI system doesn't make catastrophic mistakes during amplification

The goal of IDA is to find an amplification process we can **generally trust**. One strong contender is factored cognition which starts from a reasoning process we understand and endorse. However, even if we find a generally good way to do factored cognition, it will probably have **some weaknesses** that might have **catastrophic consequences**.

To deal with these weaknesses, IDA proposes two processes that should happen during amplification alongside capability amplification (making our AI system more powerful): **Reliability amplification** and **security amplification**.

## Why we need reliability amplification

**Reliability amplification** aims to solve the following problem: Even if we have an **in principle corrigible system**, for any given input, there is **some probability that it acts malignly**.

### Reliability decreases with amplification

Example: Kim is very **unlikely to steal just for fun**. However, with every year Kim ages, there is some small probability that their **motivation changes** and they do steal just for fun. We give Kim the task “never steal just for fun in your life”. This task **critically depends on each and every single time slice** of Kim. They all have to solve the subtask “don’t steal just for fun”. Kim solves the task.

Now we give a more difficult task: “For 10000 years in a row, don’t steal just for fun.” Normal Kim can’t solve the task within their lifespan, so we **amplify Kim** by adding 10000 years to their lifespan. During those 10000 years, it seems quite likely that **some time slice would defect**. For every year we add (to task and lifespan), the probability of failing the overall task and taking a bad action increases, i.e. **reliability decreases with every amplification step**. What’s worse is that **once Kim defects**, it’s likelier that they will **defect again** the next year.

This problem is **analogous to amplifying an AI system**: In factored cognition, we create lots of small subsystems and give them subtasks that might be **critical for the overall task**, i.e. if we solve the subtask wrong we solve the entire task wrong. For every subsystem we add, there is **some chance that it is malign**, e.g. instead of having “help the human” it could have “gain influence” as its ultimate goal.<sup>[18]</sup> This is worse if we **design future subsystems based on the malign subsystem**.

### Reliability amplification is double checking the solution to all subtasks

Reliability amplification is letting **multiple subsystems solve the same subtask** and **crosscheck each other's solution** instead of relying on just one subsystem.

Example: In the earlier example, for every additional year Kim got to live, the probability that they were going to steal just for fun at some point in their life increased. But now every time we add a year to the Kim’s life, we also create two copies of Kim. And at any point in time, Kim only gets to steal just for fun if the **majority of Kim copies approve** of that action.<sup>[19]</sup> This seems to sufficiently counteract the increased failure probability from adding another life year. We could make use of the same principle to do reliability amplification alongside capability

amplification in AI systems by basically **doing every task triple** or some similar mechanism.

The problem with this is that it would probably **slow down capability amplification**.

## Why we need security amplification

### We need security amplification to deal with adversarial input

**Security amplification** tries to solve the following problem: There will be **some inputs that reliably cause my generally corrigible system to fail**.

Example: An AI system that **mostly gives correct answers** can still **fail completely on certain inputs**. For example an image classifier might correctly label images as “cats” and “cars” 99% of the time<sup>[20]</sup> but mislabel some seemingly easy images with very high confidence and for reasons that are opaque to us. These kinds of input that produce poor behavior are called **adversarial examples**.

Reliability amplification does not help with this problem.

Example: Kim is very unlikely to steal just for fun unless Kim’s crush told them to. Then they will always steal. Even if Kim copies need to consent to Kim stealing just for fun, if they all get told by their crush that they should, they all consent and steal just for fun. Kim would always fail if they encounter the task “Don’t steal just for fun even if your crush tells you to” while trying to solve “Never steal just for fun.”

Security amplification is relevant to defend against both **deliberate attacks** from other AIs (or sufficiently smart humans) and malign input that our AI system **accidentally** comes across.<sup>[21]</sup>

### Security amplification is avoiding inputs that are likely to cause the AI system to make catastrophic mistakes by minimizing information per input

Factored cognition means that we have many subsystems working together that each solve a subtask. Security amplification **minimizes the information each subsystem ever sees** even when the overall task is very complex.

The assumption behind security amplification is that inputs that cause an amplified system to completely fail (behave incorrigibly) are rare and complex: Hopefully, **inputs with low complexity (little information) never cause the AI to behave incorrigibly**.

Details: Imagine the AI is exposed to some complex information, that would usually cause the AI system to behave incorrigibly. However, we install a mechanism that **automatically breaks down complex information into many pieces of less complex information** before giving it to any of the subsystems.

Example: We give the AI access to the internet and ask it: “Given all this information, why was the Roman Empire so successful?”. Imagine there was something about the internet, such that any subsystem that was exposed to all the information the AI has to look through to answer this question would behave incorrigibly.

Security amplification means we use a mechanism that only lets through simple information: Exposing any subsystem to all the internet would be impossible. The highest level subsystem (that would be Rosa in HCH) only sees the simple question

“Why was the Roman empire so successful?” and is made aware of the existence of the internet. The highest level subsystem cannot look at the internet itself, but instead breaks down the question with the little information it has and passes subquestions on to the next lower subsystems.

This goes on until the lowest level subsystems (in HCH that would be Rosan) only see a tiny directly answerable subquestion they can look up. The subsystem one level higher sees the answer of several of these lowest level subsystems and integrates them to answer a question that’s a bit more complex (**local integration**) etc. The highest level subsystem never has to see the details that the lowest level subsystems figured out (there is **no global integration**). It just gets enough higher level information to integrate into an answer.

So far, the break-down mechanism is unspecified, but it has to be simple enough to be robust against adversarial examples itself. For example, A mechanism that chopped every input up into strings of 10 words and only let through one string at a time, it would probably be **simple enough to not be corruptible**.

This seems very similar to **what we would like to do in capability amplification anyway** and indeed Paul Christiano thinks that it is likely that we will solve both via the same mechanism. IDA’s current proposal for capability and security amplification is called **meta-execution** which I will not explain in this document.

## How the amplified system ensures corrigibility is preserved during distillation

IDA is mostly focused on exploring the possibility of **safe and strong amplification**. One reason for this is that **distillation is closer to mainstream work on AI** since finding a safe distillation procedure essentially answers the question “How do I best train an agent with a bunch of training data?” Nonetheless, IDA proposes some necessary mechanisms and properties to ensure distillation is safe.

### Informed oversight and ascription universality: Making sure that the distilled system can’t hide anything from the amplified system

*Informed oversight* means ensuring that the **overseer has access to all the relevant information** about what the AI system in training is doing. In the context of IDA, the overseer is the last amplified AI system and the AI system in training is the next distilled AI.

It is important that the overseer both knows **which action** the distilled AI wants to take as well as **why** it takes that action. Otherwise, the amplified AI could miss

1. **unwelcome consequences** that the distilled AI is aiming for but that the amplified AI is not aware of, and
2. the distilled AI taking a good action **for the wrong reasons** (e.g. to gain trust and deceive the amplified AI in the long run).

To ensure informed oversight, IDA proposes that the amplified system must be **ascription universal**. We already discussed the idea of an **universality threshold** for successful factored cognition and HCH. The idea was that the subsystems involved

in factored cognition must be above some threshold, such that with enough of these subsystems (enough computation), they could **solve any relevant problem**.

**Ascription universality** is the idea that a system always has **at least as good beliefs as any sufficiently computationally simpler system**. This means, for all possible questions and tasks, whenever an ascription universal system disagrees with a computationally simpler system, the ascription universal system is more correct. Ideally, this would be true for any computationally simpler system, but in practice this might only be possible for systems that are computationally simpler by a certain amount.

Ascription universality is relevant for informed oversight because an ascription universal system has **at least as good beliefs about another system as the other system itself if it is sufficiently computationally simpler**. This means that an ascription universal overseer would always understand a sufficiently simpler AI in training at least as well as the AI in training understands itself, i.e. the **AI in training cannot hide anything from the overseer**. In IDA, a distilled AI (which is being trained) is a less computationally expensive version of an amplified AI (which serves as overseer), so an ascription universal amplified AI would have informed oversight over its distillation if the computational gap is large enough.[\[22\]](#)

Ascription universality is still a **vague concept**, e.g. it is not specified how exactly we should understand computational complexity or how much simpler a computationally simpler system would have to be, such that an ascription universal system is guaranteed to have at least as good beliefs as it.[\[23\]](#)

### **Box 1: What does ascription universality have to do with ascription?**

Some people might be confused why this is called ascription universality. An ascription universal system always has better beliefs than any simpler system, but in order to compare beliefs, **we need some way of ascribing beliefs**. For every behavior, we could ascribe many different beliefs to a system that could explain that behavior.

For example, imagine Eduard is eating a mushroom. **We can ascribe different beliefs** to Eduard. Eduard might believe that mushrooms are delicious but not nutritious. Eduard might also believe that mushrooms are nutritious but not tasty. We can reasonably ascribe different beliefs to him that would all explain his behavior. An ascription universal system must have **at least as good (correct) beliefs than any of the beliefs we could reasonably ascribe** to Eduard, i.e. the system's beliefs about mushrooms must be at least as accurate as any belief we could reasonably ascribe Eduard to hold.

It is **impossible** to have a system that always has better beliefs than *any* beliefs we could ascribe to another system. For example, imagine a dog eating and we ascribe beliefs about the future consequences of eating to the dog. One way of ascribing beliefs to the dog is by saying that the dog believes that its hunger will be satisfied when eating. We could also ascribe a different belief to the dog: We could think that the dog actually correctly calculated how eating will influence the next millions years of the future and that the dog is eating because it's intending to bring these consequences about. An ascription universal system that always has at least as good beliefs as any other belief we could ascribe to a computationally simpler system would have to have better beliefs about the consequences of the dog eating than the belief we ascribed to the dog in the second case. However, we ascribed the dog in the second case to have accurate and very detailed beliefs about the consequences of it eating. Therefore, it would be nearly impossible to have better beliefs about the consequences of the dog eating than the beliefs we ascribed to the dog in the second case. Therefore, **an ascription universal system must only have better beliefs than any belief we could reasonably ascribe to a computationally simpler system.**

### **Box 2: Ascription universality and competitiveness**

Ascription universality might also be crucial for **competitiveness**. Imagine we successfully developed a corrigible and powerful AI system in accordance with IDA. However, we continue to search the space of possible AI system designs to improve our AI. (This might be done by the AI itself.) If our amplified AI system is **not ascription universal**, we might find **another AI design that would be computationally cheaper** or have better beliefs than our safe AI system and we might be tempted to **implement it even if it is potentially unsafe**. This is why ascription universality might be **necessary to stay competitive** (and preserve safety) in the long-term.

Ascription universality is only defined in terms of the system's ability to have at least as good beliefs as computationally cheaper systems. However, this **helps us with computationally more expensive systems as well**. Imagine, there was some AI system X that was more computationally expensive and would have better beliefs than our amplified AI. We could construct an amplified system AI' that is computationally more expensive than that system X. (For many ways of amplification, particularly the proposed factored cognition, we can just arbitrarily add more steps: For example, in the case of factored cognition, we can just continue to break down subtasks further and further. (Remember that HCH is an infinite sequence of breaking down tasks.) Hence, we could always make the amplified system arbitrarily more computationally expensive.)

We can then check if AI' has at least as good beliefs as system X. If our amplified AI and all AI' are ascription universal, this would always be the case (if the computational gap is sufficiently high). Hence, there would **never be an AI system that has better beliefs than our amplified AI scaled up to a sufficient level of compute**. Hence, ascription universality might be sufficient to ensure **competitiveness at run-time**, at least in cases where we have the same amount of computation available for running our amplified system as we have for system X. (e.g. if one and the same agent decides between implementing these two systems.)

One important thing to note is that ascription universality might ensure competitiveness at run-time, but does **not imply** that an ascription universal system is **competitive in the sense that we can make as quick progress** on it as on other AI systems. In fact, the opposite might be the case, which might mean that a group of people working on an ascription universal AI system might get outcompeted by another group that achieves powerful AI first and does so in a way that prevents further work on the ascription universal AI system.

## **Security makes informed oversight more difficult**

Informed oversight is complicated by **security problems**: It is possible that the distilled AI produces an input of the kind that causes our amplified AI to fail (behave incorrigibly). This means we want to apply the same mechanism as in security amplification, i.e. only expose each subsystem of our amplified AI to very little information from the distilled AI.<sup>[24]</sup> This might make it substantially harder for the amplified AI to understand everything the distilled AI is doing.

## **Adversarial training: Finding and fixing the distilled system's vulnerabilities**

Informed oversight solves the problem of the distilled system taking actions with malign intentions without the amplified AI noticing. However, even if the distilled AI is acting genuinely corrigibly in all situations it has been exposed to, there might be some inputs (**adversarial examples**) the distilled AI is not aware of that will cause it to fail completely (act incorrigibly). **Adversarial training** tries to prevent this.

In adversarial training, we use an adversary that tries to find these adversarial examples that cause the AI to completely fail in order to fix the AI in a way that makes it not fail. In IDA, the last amplified AI could play the role of the adversary and do adversarial training with the distilled AI in training.

## **How IDA could fail**

There are three ways IDA could fail: First, it could **just not work**, we might be unable to create an AI system that way or at least an AI system that is competitive enough. Second, it might **fail catastrophically because of misalignment**. We might develop an extremely capable AI with the ideas of IDA but overlook something important that leads to incorrigibility and catastrophic outcomes either immediately or further down the line. Third, we might be able to develop a corrigible but not sufficiently competent AI with IDA that leads it to **fail in subtle ways** that we don't instantly recognize but has catastrophic consequences. I will outline some reasons why we might believe each of the three ways to come true.

## **We might not be able to implement a competitive version of IDA**

There is one very simple way in which IDA could trivially become obsolete. In the coming years, we might just develop **new, significantly stronger AI techniques** that work very different from what we currently have and that **IDA is not applicable** to and that IDA can't compete with. We might also realize that IDA's ideas are **conceptually flawed**.

## **HCH might not be powerful, corrigible or translatable to the real world**

Similarly we might realize that our intuitions around the **concept of HCH** are just wrong. We might conclude that HCH could **never robustly be corrigible** or never

become **powerful enough**: HCH might not be **universal** either because no universal reasoning process exists or because HCH is not able to do it.

## HCH might not be corrigible

IDA argues that in order for HCH to be robustly corrigible, all of the **subtrees in HCH should be corrigible in isolation** as well. It might be extremely hard or **impossible** to find break-down mechanisms that robustly preserves corrigibility of the subsystems.

Examples that would not satisfy this condition:

Task	Subtasks	Why the subsystems are not corrigible
Satisfy all human preferences	1) Figure out human values 2) Optimise for those values	An AI that is optimising for a fixed set of values is a bad idea because the values are likely to be wrong in relevant ways. We want subsystem 2) to ask itself "Given that I think human values are [XYZ] and have uncertainties [ABC], how should I act?"
Safely develop AGI	1) Find 10 plausible safe AGIs 1.1) Find 10 plausible safe AGIs 1.2) Find 10 plausible safe AGIs 2) Find 10 plausible safe AGIs 2.1) Find 10 plausible safe AGIs 2.2) Find 10 plausible safe AGIs 3) Find 10 plausible safe AGIs 3.1) Find 10 plausible safe AGIs 3.2) Find 10 plausible safe AGIs	We end up with searching through a huge space of possible AGI designs of which many will not be safe.
Real life example: Maximise the profit of this company	Do the work in your job description	Most employees have incentives that are not maximising the company's profit

[Yudkowsky](#) argues that even if **all subsystems** in HCH are corrigible, the **system as a whole** might not be. He argues that if HCH is really smart, this intelligence must come from somewhere, either from the **individual subsystems** or the way they **interact with each other**. The smallest subsystems in HCH are humans, so if the intelligence from HCH is coming from just the subsystems it consists of, it cannot become very powerful. The **interaction of the subsystems** must do some magic. However, if the interaction is doing enough magic to make an extremely powerful system out of less powerful subsystems, it might also **make an incorrigible system out of many corrigible subsystems**.

## HCH might not be able to solve any relevant tasks

There might be **no way of breaking tasks down into subtasks** for any relevant domain. It might be that **all relevant tasks require global integration**, i.e. one and the same agent has to go through all the relevant information and steps to solve the task. In this case, IDA would just not work.

Example: A pregnancy takes roughly 9 months. Taking 9 people with uteri doesn't speed up the pregnancy to 1 month. There is no way of breaking that task down and delegating the subtasks (at least currently). Things like writing poetry or finding a maths proof intuitively seem similar. Finding ways to break down tasks might fall into a similar category.

Even if tasks could in principle be broken down, every way of breaking tasks down might be **specific to a very narrow type of task**. We might have to find **intractably many task-specific break-down mechanisms** to get meaningful performance. [William Saunders \(2019\)](#) discusses **task-specific "manuals"** for overseers in the IDA context but that would probably only work with a **limited number of tasks** that are **broad enough** to be helpful. This might mean that IDA just doesn't work or lead to subtle failure in the form of what I call the **differential competence problem**, which I discuss in the next section.

## **The differential competence problem: HCH might favor some skills over others and lead to a bad combination of abilities**

HCH might be **differentially competent**: It might only be able to solve certain kinds of tasks or be better at certain tasks than others. Wei Dai argues that this might lead to **bad combinations of abilities** that might lead to **catastrophic outcomes**.

Example: It might be easier to break down the task "Create more destructive nuclear weapons" than "Find effective ways for humans to cooperate with each other to ensure peace." If so, HCH might make humanity better at the first than the second, which could increase the possibility of catastrophic wars.

### **The differential competence problem might be worse with AI systems**

We also have some reason to think that this might be especially worrisome when we move from HCH to factored cognition with **Als**. Compared to humans, an AI system might be particularly good at things for which there is a lot of **data** and **feedback** available.

Example: For **corrigibility** to actually be helpful, we need AI systems that have sufficiently good **models of our preferences** and **when to seek approval** for their actions. This might be something an AI system is comparably bad at relative to other capabilities. This would mean that their ability to judge when to seek our approval is not proportional to the general power they have. This is bad news, if we develop an AI system that is incredibly good at building weapons of mass destruction but has a poor understanding of what we want and when to consult humans.

Example: We can also imagine an AI system that realizes we value truth and wants to help us, but is not very good at doing moral philosophy and not very good at finding the optimal truth-seeking strategy of communicating its current opinions. This seems somewhat plausible since moral philosophy is an area with arguably very little data and feedback (unless you have a very descriptive view of ethics). This might be bad if the AI system is comparably better at e.g. convincing people or engineering powerful tools.

## We might not be able to safely teach AIs factored cognition

Even HCH works conceptually, we might **not be able to make it work for an AI system**. Safe amplification could be **computationally intractable** for example. [Ought](#) is currently trying to find out empirically how powerful factored cognition with real humans can become.

Factored cognition in **amplification could also still work even if HCH does not work** because **AI systems are not humans**. However, I imagine it to be extremely hard to design an AI doing factored cognition in ways that humans with infinite computation cannot do.

## Potential problems with corrigibility

There are a number of criticisms of the role of corrigibility in IDA.

### Corrigibility might be doomed without a firm theoretical understanding of what it is

Ben Cottier and Rohin Shah ([August 2019](#)) summarize one disagreement around corrigibility as follows:

*[T]his definition of corrigibility is still **vague**, and although it can be explained to work in a desirable way, it is not clear how **practically feasible** it is. It seems that proponents of corrigible AI accept that **greater theoretical understanding** and clarification is needed: how much is a key source of disagreement. On a **practical extreme**, one would iterate experiments with tight feedback loops to figure it out, and correct errors on the go. This assumes ample opportunity for trial and error, rejecting Discontinuity to/from AGI. On a **theoretical extreme**, some argue that one would need to develop a new mathematical theory of preferences to be confident enough that this approach will work, or such a theory would provide the necessary insights to make it work at all.*

In other words, the argument of the corrigibility critics consists of three assumptions and one conclusion:

1. Corrigibility is currently **too poorly understood** to make it work in AI  
(Proponents and critics seem to agree on this)
2. A very clear and **formal conceptual understanding** of corrigibility is necessary to make corrigibility work in AI.
3. Getting this formal understanding is **impossible** or **unlikely**.
4. Conclusion: We cannot rely on corrigibility.

### The proposed safety mechanisms might not be enough to ensure corrigibility

A second argument is about the **concrete mechanisms implemented** to ensure the corrigibility of the AI system. The safety mechanisms IDA proposes to preserve corrigibility (informed oversight, adversarial training, reliability amplification, security amplification) **cannot formally verify corrigibility**. They all aim to make it

**sufficiently likely** that the AI system in question is corrigible. You might think that it is **not possible** to get a sufficiently high likelihood of corrigibility that way.

## Corrigibility might not be stable

Another area for criticism is the claim that corrigibility is **stable**. Here is a counterargument against the stability of corrigibility: Imagine we develop an AI system that wants to earn as much human approval as possible. The AI gets a task and can do it the safe and diligent way. However, doing so might be really inefficient, so the AI might get **more approval** if it does a slightly **sloppy job**. Over time, the AI becomes less and **less diligent** and finally **unsafe**.

## Corrigible AI might unintentionally change our preferences in a way we don't want to

Wei Dai proposes another problem related to corrigibility. The ideal AI system in IDA is trying to clarify the user's preferences. However, these are likely to change over time and the **AI itself** will likely **influence our preferences**. We want the AI to only influence us in ways that we **approve of**. This might be extremely **hard** or **impossible** to define, implement and verify. The argument's weight depends on which ethical position people have and what exactly we mean by preferences, but if one has strong feelings about these, this might constitute a failure mode on the level of a catastrophe.

## Acknowledgements

I would like to thank Rohin Shah, Max Daniel, Luisa Rodriguez, Paul Christiano, Rose Hadshar, Hjalmar Wijk and Jaime Sevilla for invaluable feedback on this document. I also want to thank everyone at FHI (visitors and staff) who was willing to discuss and clarify IDA or plain out patiently explain it to me. Good ideas are thanks to others and all mistakes are mine :)

- 
1. For an overview of the current landscape of AI safety research agendas see [here](#).  
↳
  2. **E.g. Alex Zhu, 2018 in his [FAQ on IDA](#)** ↳
  3. **Safety:** I will clarify what safe means in the context of this write-up [further below](#). **Unsafe:** An unsafe AI system has an unacceptably high probability of producing very bad (catastrophic) outcomes, e.g. as side effects or because it is optimizing for the wrong thing or would have an unacceptably high probability to do so if scaled up.  
**Powerful:** When I use the term powerful I am pointing towards the same thing that others call superintelligence.  
**AI, AI systems and agents:** I use the terms AI and AI system interchangeably and sometimes refer to a powerful AI systems as agents. ↳
  4. We can come up with **social solutions** to **ease competitive pressures** but probably only if the efficiency gap between safe and unsafe AI alternatives is not too big. Paul Christiano [estimates](#) that a safe AI system needs to be about 90% as efficient as its unsafe alternatives. ↳

5. \*\*Arguably, no current AI technique is really unsafe because the AI systems we can create are not strong enough to be meaningfully dangerous. Unsafe in the 2019 context means that they would be unsafe if scaled up (e.g. by increasing computation, computational efficiency or available data) and applied to more powerful AI systems. ↵
6. \*\*The research agenda is actually called “Iterated Amplification” but is still often referred to as “Iterated Distillation and Amplification” and abbreviated as “IDA” ↵
7. Realistically, we would probably have to set up a new AI system with new distillation and amplification methods from time to time that replace the old one instead of having one single system that infinitely improves itself. An alternative way this could look like is that we train an AI system via safe iterated distillation and amplification and this AI system aids us in developing another, stronger, superintelligent AI system that is not trained via distillation and amplification but via another training method. Which of these three paths to superintelligence would actually play out is not relevant, there are all ideal in the sense that they all essentially mean: We successfully implement an agent with good distillation and amplification procedures and from there on, the development of safe and powerful AI is taking off and taking care of itself. Our work is basically done after having figured out safe distillation and amplification methods, or at least significantly easier. There are also more complicated versions of this where an AI system that is trained by safe distillation and amplification can potentially take us to safe, extremely powerful superintelligent AI, but needs some ongoing fixing and effort from our side. ↵
8. This seems very similar to the third ideal scenario I presented in footnote 7. The deciding difference here is that in this case, while IDA is helpful, it does not solve the full problem and we still have to put a lot of effort in to ensure that things don't go very wrong. ↵
9. \_For example MIRI's early writing attaches a slightly different meaning to corrigibility than Christiano does: <https://intelligence.org/files/Corrigibility.pdf> . The concepts are quite close to each other. MIRI's definition “[an AI that] cooperates with what its creators regard as a corrective intervention] is probably a property that Paul Christiano would also see as central to corrigibility. However, MIRI's corrigibility is a property that can be formalized and verified. It might for example work like a checklist of actions a human could take in which case the AI should obey (e.g. shut down when a shutdown button is pressed) (This might stems from the belief that there is some [algorithmically simple core](#) to corrigibility.) Christiano's definition is broader, less formal and directed at the AI's intentions rather than concrete actions it should take in a concrete set of circumstances (e.g. human shutting you down). ↵
10. An AI with the goal “Paint the wall in the color the user wants” is trying to do what we want it to do de dicto. An AI with the goal “Paint the wall blue” is trying to do what we want it to do de re. ↵
11. This departs from the idea of designing a ‘perfectly rational’ AI that’s maximising some objective reward function which represents some ‘true’ preferences but also seems more tractable to Christiano. ↵
12. To give an idea of what “leaves us in power” means, Christiano (June 2017) gives a rough definition of corrigibility by examples of what he wants a corrigible AI system to help him do:

- Figure out whether I built the right AI and correct any mistakes I made
- Remain informed about the AI's behavior and avoid unpleasant surprises
- Make better decisions and clarify my preferences
- Acquire resources and remain in effective control of them
- Ensure that my AI systems continue to do all of these nice things ↵

13. \*\*[Yudkowsky](#) discusses the possibility of a “relatively simple core or central principle” to corrigibility. I believe that by that he means that an algorithmically simple core in the way that you can just write it down on paper and formally verify it. This is not the way Christiano understand this. He is referring to some core of corrigibility that is “easy to learn in some non-anthropomorphic sense” but is not easy or even possible to write down in maths and not simple to algorithmically specify. Instead, corrigibility is somehow picked up by the AI during its learning process. It's quite vague how exactly that happens. ↵
14. In principle, you could imagine and insert all kinds of methods and treat “amplification” and “distillation” as place holders. However, IDA’s contribution is to concretely spell out why we should be optimistic about finding suitable procedures and how these procedures are likely to look like. As part of IDA, there is currently work around properties that these procedures would need to have to be safe and concrete-ish candidates for the amplification and distillation step. If these properties and candidate procedures turn out to not work out and also don’t give leads to other properties and candidate procedures that could be inspected for suitability, that **doesn’t necessarily mean that no system trained by iterated distillation and amplification can ever be sufficiently safe and powerful, but IDA would lose most if not all of its value.** (Similar to how a safe system trained by iterated distillation and amplification does not necessarily have to be corrigible, but if it is not, then the research agenda IDA might have no value.) ↵
15. I only discussed HCH as a reason to think amplification might be promising. Of course, increasing the available compute and dividing tasks into subtasks is just one of many ways to amplify an agent. However, so far it is the only one that people propose as a safe amplification step with no (or a very high) upper bound in terms of capabilities. Therefore, I will treat IDA as if it is tied to this specific way of amplification even though it is not in principle. This doesn’t mean we won’t ever find another promising amplification step. However, unless we can find another promising amplification method in the context of IDA, it seems unclear to me how promising and valuable IDA is. It seems like pointing out HCH was one of the main contributions of IDA. ↵
16. Part of the idea is that the amplified AI can train the distilled agent via **supervised learning** or **reinforcement learning**. In the **supervised** learning setting, the distilled AI looks at how the amplified AI solved tasks and tries to **imitate** it. During training, the amplified AI could test the distilled AI by giving it tasks and checking the distilled AI’s proposed solutions by solving them itself via factored cognition, comparing if they are the same and giving feedback if the proposed solution was right or wrong.  
There is also a version of factored cognition that would allow us to use **reinforcement learning** for the distillation process: **Factored evaluation**. Factored evaluation is solving tasks of the type “evaluate how good this input is”. This might be important if there are many things we want the distilled system to learn are not of the form “this is the correct solution to this” but “this thing is a little good, this is very good and this is very bad”.  
During training, the amplified AI could test the distilled AI’s proposed solutions

not only by comparing them to its own solutions and giving a “wrong”/“right” signal, but by doing factored evaluation to give the distilled AI a **reward signal** indicating how good or bad the proposed solution is, i.e. we can do reinforcement learning. [←](#)

17. In her [summary of IDA](#), Ajeya Cotra seems to be taking a different stance:  
*“But in IDA, amplification is not necessarily a fixed algorithm that can be written down once and repeatedly applied; it’s an interactive process directed by human decisions.”*  
This quote implies that humans will always lead and be part of the amplified system (taking a similar role as the highest level node in HCH, i.e. Rosa.) I think this might be plausible when we are still in a phase in which the AI system has not learned how to break down tasks, yet and distills that knowledge from interactions with a human (e.g. either by imitating them or through incentives the human provides). However, I think at some point we will probably have the AI system autonomously execute the distillation and amplification steps or otherwise get outcompeted. And even before that point we might find some other way to train the AI in breaking down tasks that doesn’t involve human interaction. [←](#)
18. This problem would not exist if we found a way of amplification that only creates perfect subsystems that would **never** “steal just for fun”, but Christiano thinks that that is a lot harder than doing reliability amplification. [←](#)
19. Of course for this to work the decisions of Kim copies have to be less than perfectly correlated, but the problem of reliability loss only occurs because we assume that we can’t make “perfect copies of subsystems including their motivational structure”/“age perfectly” anyway. [←](#)
20. or at least have low confidence in its solution when it gives a wrong answer [←](#)
21. You might hope that we would automatically be safe from deliberate attacks from other AIs even without security amplification since our AI system is supposed to be competitive. However, it might be that defense is a lot harder than offense in which case it isn’t automatically enough to have an AI system that is just similarly capable as any potential attacker. [←](#)
22. We might have additional reason for optimism w.r.t. informed oversight since the amplified AI and distilled AI could directly share some of their computation. [←](#)
23. Paul Christiano expects the definition of ascription universality to change in the future and intends to replace the term ascription universality with a different one once this happens. [←](#)
24. In the context of IDA, people refer to overseers that are restricted in that way as **low bandwidth overseers** and overseers that are not restricted that way as **high bandwidth overseers**. [←](#)

# When can Fiction Change the World?

I suspect that a nontrivial percentage of the people reading this became involved with the community because of *Harry Potter and the Methods of Rationality*.

So to the extent that those who were drawn to join the community because of that source are making the world a better place, we have at least one clear example of a novel having an important impact.

I've made a living through self publishing novels for the last five years (specifically Pride and Prejudice variations, that is Jane Austen fan fiction). Recently inspired by conversations at EA Virtual and worries made more emotionally salient by GTP-3 examples, I decided that I wanted to put part of my professional time towards writing novels that might have a positive impact on conversations around AI.

As part of this I did some thinking about when fiction seemed to exert an influence on public policy, and then I looked for academic research on the subject, and I think there are people in the community who will find this write up about the subject interesting and useful.

## Theoretical Model

I identified four common mechanisms that seemed to be involved when fiction had a large impact on opinions. This is not an exhaustive list, and there is some overlap and fuzziness around the boundary of each concept.

### **Radicalizing the already convinced:**

A classic example is *Uncle Tom's Cabin*, a novel about a slave unjustly suffering written in the 1850s that was credited with helping to spark the Civil War. *Uncle Tom's Cabin* did not introduce anyone to the idea that slavery was bad, or convince anyone who thought that slavery was a fine peculiar Southern institution that it was actually evil. However it seems to have radicalized Northern attitudes towards slavery, and it was part of the moment when enough one issue voters on slavery existed that the party system broke down and allowed the new abolitionist Republican party to win congress and the presidency in 1860.

Research has been done via surveys to find out if readers of popular novels about climate change have changed their views about climate change relative to similar readers who did not read any of them. Concerned readers become alarmed by climate change after reading, and those who were aware but not concerned become concerned. However, readers who think that climate change is a hoax usually don't read 'Cli-fi' and when they do read it, their opinions are not changed. (Schneider-Mayerson, M. 2018 and 2020)

### **Evoking empathy for new groups:**

*Uncle Tom's Cabin* succeeded at radicalizing northerners by making them care about the fate of a particular southern slave. LGBTQ representation in media drive viewers to care about gay characters, and see them as normal human beings who deserve to have the same chances for happiness as anyone else. In the 1970s and 80s *The Jeffersons* and *The Cosby Show* helped convince white Americans that black Americans could be successful, intelligent and well dressed citizens.

Some of the research in cultivation theory specifically shows that heavy TV watchers as a group changed their opinions on minorities far more than the general public as positive representations of these groups became common.

On the other hand, negative representations can exist. For example anti semetic stories in which the long-nosed Jewish banker mistreats a poor person. Or media which portrays minority Americans as dangerous and violent. West German attitudes towards the Holocaust were substantially changed when a quarter of the viewing population watched the *Holocaust* miniseries in the seventies, and this likely contributed to passing laws that supported extraditions of more war criminals to Isreal.

### **Exposure to New Points of View:**

After reading Upton Sinclair's novel *The Jungle*, the president at the time, Teddy Roosevelt supposedly stared at his morning sausage and was unable to stomach the possibility of human body parts being in it. The rest of the public was equally horrified by the prospect of tainted meat, and within a year the act establishing the FDA was made.

*Methods of Rationality* exposed many of its readers to a way of thinking about the world that they'd never seen and that they found highly engaging. When I read *Atlas Shrugged* as a teenager who was trying to decide if he thought God existed, I saw for the first time an expression of intellectually satisfied atheists who were confident in living and being happy living without believing that God existed. Supposedly many of the people who laid the foundation for the science of robotics were inspired by reading Isaac Asimov's *Robots* series.

### **Community Building:**

The science fiction community created a space for people interested in engineering and technology to meet each other informally in the early part of the century. *Atlas Shrugged* recruited people to join objectivist circles. The Less Wrong community recruited probably half its current population from people who loved *Methods of Rationality*.

A paper about fiction influencing international relations argued that the Left Behind novels were part of the structure that maintained unity amongst Christian Evangelicals during George W Bush's presidency. (Musgrave)

### **How attitudinal changes lead to real world changes**

#### **Influencing specific important individuals:**

Ronald Reagan became more opposed to nuclear weapons after watching *The Day After*. Scientists who started the field of robotics read Asimov. Paul Ryan and Alan Greenspan are huge fans of Ayn Rand and *Atlas Shrugged*. Many of us are here because we liked *Methods of Rationality*. Often specific influential individuals then cause major changes because they were influenced to do so by reading a specific book or watching a specific movie.

#### **Supporting a mass movement:**

Uncle Tom's cabin radicalized Northern abolitionist attitudes, and was part of the process that led to the Republican party. Radicalizing environmentalists via cli-fi

possibly has led to groups like Extinction Rebellion and enthusiasm for buying Teslas. Anti-war films reduced political willingness to keep troops in Vietnam and Iraq. Nuclear apocalypse films increased political support for the Test Ban Treaty. Global warming films are definitely part of why the public in most countries supports things like the Paris Agreement, or why California passed a cap and trade policy. LGBTQ rights and civil rights are more likely to get friendly supreme court rulings, friendly company hiring practices, etc because they are popular, which is partly because of media representations.

### **How does fiction change attitudes?**

#### **Transportation theory; synthetic experiences and aliefs:**

The brain doesn't fully treat fictional evidence differently, and fiction gives a veneer of the real and specific to ideas that before were just general and abstract.

Professor Paul Bloom at Yale describes this phenomenon as fiction creating 'aliefs', where the emotional portion of our brain responds to things that are not real as though they were. For example most atheists are unwilling to sign a contract selling their soul to the devil, which Bloom interprets as being an example of part of the brain treating the fictional entity as real. However, as we all know the real world is theoretically over determined, and refusing to take twenty dollars to sell your soul to the devil might be a reasonable response in a case of potentially infinite stakes with trivial residual uncertainty.

Feeling deeply transported into a different world in a key part of fiction's appeal, and it seems to be an important part of how fiction can make people treat fictional examples as though they were concrete and real. (Busselle 2009)

A research manipulation to make readers feel more emotionally engaged before exposing them to a story changed how believable they found it, while changing whether a story was labelled as fiction based on a true story, or a true story did not.

Cli-fi readers talked about now feeling like climate change would actually happen to actual people. This led to increased radicalization as they had a sense of connection to this future. Having characters who the readers could identify with acting in familiar settings that create a sense of place and being a real location may have been what drove that success. (Schneider-Mayerson 2018)

Fiction influences readers by making the problem seem more real, creating a feeling of emotional verisimilitude and plausibility, and making the problem seem vivid and concrete.

Lesson for EA writers: Someone, possibly Mark Twain, said that you need to give your readers two familiar things for every strange idea you introduce. If you want to get at a broader audience (*not necessarily a mass audience!*) to be moved by your book, try to make sure you actually hit that target.

### **Cultivation theory:**

According to this model, people think the world is like the media they are repeatedly exposed to. As a result people who watch lots of TV think there is more crime and that more people are lawyers than people who don't watch very much TV. They also like minorities and LGBTQ groups as well today as low TV watchers, after decades of positive media representation of those groups. However in the early surveys before

this representation happened, high TV watchers were more bigotted against minority groups. (Mosharafa 2015)

A Harvard professor convinced lots of TV shows to insert designated drivers into episodes where the group went to a party and got drunk as part of the public safety campaign. Given how much money different companies pay to have people be seen in movies drinking Pepsi while flying somewhere on Southwest, after they paid with a Chase credit card, we can assume that ideological product placement probably has some impact.

Maybe we could try to convince screenwriters to look for a chance to get their characters to be mentioned doing EA type things, like donating a regular percentage of their money to extreme poverty reduction, or talking about impact evaluations after mentioning they've donated to something.

Audiences will reject and resist ideas that disagree with their preexisting assumptions. They also will draw lessons and meanings from fiction that are congruent with what they already believed.

Example: After reading a cli-fi novel where a PoV character betrays and murders another PoV character at the very end of the novel, conservative and moderate readers drew the lesson from the book that you can't trust anyone, and that you need to be grateful for the little things in life. They also tended to identify with the ruthless and initially amoral rich male character, rather than the middle class activist/journalist or the poor Latina immigrant. (Schneider-Mayerson 2020)

Example: Teenagers and children make fun of media that is obviously trying to convince them to act in a way that adults would like them to, for example anti-drug messages in shows for teenagers seem to have had a very limited effect, especially when the 'facts' in the message are broadly believed to be false.

Example: Tom Clancy's influence on the Republican policy elite was much smaller during the George W Bush presidency when the use of preemptive wars that he thought were a bad idea had become the preferred elite policy. (Musgrove)

### **What goes wrong?**

#### **Democracy counts numbers not intensity**

If the radicalized people simply become more passionate members of a blocked political coalition, it doesn't do anything. There needs to be a way to transmit the increased passion for the subject into an actual change. Movements like the Extinction Rebellion are reflections of the way that the blocked political coalition in favor of stronger climate policies is now engaging in civil disobedience because it is clear that they are not going to be able to directly achieve the policy victories they view as desperately important through purely democratic processes.

#### **Attitudinal changes are temporary:**

A group of psychology freshmen were assigned to read *The Omnivore's Dilemma* and compared to a group of freshmen who weren't assigned to read it. Their attitudes right after they read the book were changed in the direction of the book compared to the control group, but after a year's time their opinions had mostly returned to what the control group thought. (Hormes et al 2013)

### **Books generally do not convert opponents:**

Southern slave owners did not generally read *Uncle Tom's Cabin*, and if they did, they wrote thought it was an unrealistic and dishonest portrayal of slavery instead of deciding that slavery must be ended. Climate change hoaxers do not generally read cli-fi novels, and when they do, they say that the climate change scenario portrayed was unrealistic.

The Left Behind novels may have tied together Evangelicals in the conservative political coalition, but nobody who wasn't an evangelical Christian had the slightest interest in reading them. The framework of the argument in a book must match the presuppositions of the audience (Musgrave).

The changes in attitude that come with fiction often depend on the reader not having much personal experience with the situation. In the case of events that are common in media, but uncommon in personal life, people will automatically recall media examples of the situation, for example murder trials, chemical explosions or international spy rings. But if you ask them to recall an event that is both common in real life and in media, such as dates or highway accidents, they automatically think of their own experiences or those of friends (Busselle 2003).

Southerners had many personal experiences with slavery that would have dominated fictional portrayals in how they thought about it. AI researchers have daily personal experience with AI being extremely dumb and not suddenly destroying the world.

Lesson for EA writers: People who disagree with the model of the world they think is expressed by your book probably won't read it. If they do read the book, they will be in a 'am I allowed to disbelieve this' mindset, rather than trying to figure out if it is actually true. So be aware of what ideas will feel strange and might create resistance in your desired audience and either figure out a way to make your argument so that it follows logically from their existing presuppositions, or figure out a way to market your book to an audience that shares your presuppositions or is undecided on them.

Example: *The Day After* was an explicitly non partisan film designed to simply show ordinary Americans being ordinary and then dying because of the nuclear exchange. The Republican establishment that controlled the presidency at the time did not need to react to the movie as a partisan attack, but as expressing authentic concerns.

### **Backlash:**

Fictional portrayals of effective torture provoked extensive debate and elite backlash arguing against the portrayal, and thus the effect of these scenes on the support for torture was at best ambiguous, and likely null. (Payne)

But despite the response, did this portrayal possibly legitimize, or give a platform for the idea that torture could be a worthy tool in extreme situations, and thus even though the public debate did not fully support the idea, it still became more popular and legitimate?

My suspicion is that this effect was negligible. The idea that torture might be effective and legitimate to use in extreme circumstances already existed. I remember as a teenager spontaneously thinking about torturing terrorists as probably being useful on the morning of Sept 11, 2001. It is a natural idea for people to have.

Media that showed effective torture reflected this pre-existing belief. Possibly the belief could have been delegitimized in the way racism was delegitimized after the 60s by showing that it had no political or social power, but simply not speaking about the possibility of torture being effective would not have done that. The idea *did* in fact have political power and was believed by many elites in the party holding the presidency and congress.

Most likely the net effect of 24 cinematically displaying effective torture was zero.

However, fictional representations may be part of how particular policies change from polarized to bipartisan over time. For example the military preparedness policies promoted by a story about an invasion of Britain by Germany written in 1871 were strongly disliked by the Liberals who were in power at that time, but twenty years later, military preparedness against the chance of an invasion was funded with bipartisan support (Kirkwood 2012).

Lesson for EA authors: Be aware of pushing against an established opposition. If you can't undercut the coalition against your preferred policy, you probably will achieve very little. Robin Hanson's political orthogonality thesis, the idea that you will be most effective at pulling the debate in a direction without an established opposition is relevant here.

### **Depressing people too much to act**

As they are focused on disasters, often cli-fi books create intense negative affect, depression, and a sense of helplessness and hopelessness. This can lead people to paradoxically act less. (Schneider-Mayerson 2018)

### **Conclusion:**

Audiences have a sophisticated response to what they read. They will notice the things that aren't said or even considered in the books they read. If it is clear that a book is trying to promote a particular political point of view, many readers will strongly discount the intended message. They will also spontaneously come up with objections to arguments that do not feel correct to them.

For change to happen it does not only need to change opinions, there needs to be a way for the changed opinion to be turned into action.

Simply radicalizing people doesn't matter without a path for change.

This especially true if there is a blocking coalition that is unaffected by the attitudinal change. Uncle Tom's Cabin mattered because the North had enough people to politically dominate the country, and because the South delegitimized itself in Northern eyes by seceding.

In many cases politically motivated fiction that successfully radicalized those who consumed it probably did very little:

For example, Climate change fiction will only matter in the long run if it weakens the power of the blocking coalition (since the supporting coalition will act whenever it is in power anyways). And it doesn't seem to weaken the blocking coalition directly.

Possibly, despite the failure of the broader coalition, cli-fi books are actually making an important contribution by intensifying the salience of climate change in the supportive

political coalition. Climate change is viewed as an extremely important issue outside of the US, and one of the political coalitions in the US is dedicated to pushing forward climate policies. *An Inconvenient Truth* and *The Day After Tomorrow* and cli-fi novels are plausibly why the US might someday pass strong climate change policies, and why California and Germany already have.

In cases where the goal is for small numbers of people to engage in intense efforts by donating substantial amounts of money or of changing their professional plans, radicalizing a few people is probably more valuable than convincing a majority to vote differently. Democratic political majorities require broad, but shallow, agreement. Deep engagement by narrow communities might improve AI safety norms, expand the use of randomized control trials in global poverty reduction research, or fund charities that give poor American inmates bail money

Finally: Highly successful changes can take a long time to become real.

The laws against debtor's prison and child labor that Charles Dickens promoted were passed over decades. The scientists who attribute their interest in robotics to Isaac Asimov's only started making substantial progress decades after the first stories were published. To the extent that Ayn Rand's novels have led to any concrete policy changes, it has taken a long time, and those policy changes were not large.

Rick Busselle & Helena Bilandzic (2009) Measuring Narrative Engagement, Media Psychology, 12:4, 321-347

Rick W. Busselle & L. J. Shrum (2003) Media Exposure and Exemplar Accessibility, Media Psychology, 5:3, 255-282

Jenny Kitzinger, (2009) Questioning the sci-fi alibi: a critique of how 'science fiction fears' are used to explain away public concerns about risk. Journal of Risk Research

Paul Musgrave, J. Furman Daniel. Working Paper on Fiction and International Relations Theory

Schneider-Mayerson, M. (2018). The Influence of Climate Fiction: An Empirical Survey of Readers. Environmental Humanities, 10(2), 473-500.

Schneider-Mayerson, M. (2020). "Just as in the Book"? The Influence of Literature on Readers' Awareness of Climate Injustice and Perception of Climate Migrants. ISLE: Interdisciplinary Studies in Literature and Environment.

Eman Mosharafa, (2015) All you Need to Know About: The Cultivation Theory, [www.researchgate.net/publication/337077784\\_All\\_you\\_Need\\_to\\_Know\\_About\\_The\\_Cultivation\\_Theory](http://www.researchgate.net/publication/337077784_All_you_Need_to_Know_About_The_Cultivation_Theory)

Rodger A. Payne, Popular Culture and Public Deliberation about Torture

Hormes JM, Rozin P, Green MC and Fincher K (2013) Reading a book can change your mind, but only some changes last for a year: food attitude changes in readers of The Omnivore's Dilemma. Front. Psychol. 4:778. doi: 10.3389/fpsyg.2013.00778  
[www.frontiersin.org/articles/10.3389/fpsyg.2013.00778/full](http://www.frontiersin.org/articles/10.3389/fpsyg.2013.00778/full)

Kirkwood, P. M. (2012). The Impact of Fiction on Public Debate in Late Victorian Britain: The Battle of Dorking and the "Lost Career" of Sir George Tomkyns Chesney (Fall 2012). Graduate History Review.

<https://www.vox.com/2014/9/1/5998571/why-anti-drug-campaigns-like-dare-fail>

[https://www.huffpost.com/entry/designated-driver-campaig\\_b\\_405249](https://www.huffpost.com/entry/designated-driver-campaig_b_405249)

# Unifying the Simulacra Definitions

Epistemic Status: Confident this is the perspective I find most useful. This is intended to both be a stand-alone post and to be the second post in the Simulacra sequence, with the first being [Simulacra Levels and their Interactions](#). It should be readable on its own, but easier having read the previous post.

Simulacra levels are difficult to understand.

This is not without cause. This is complex and bizarre stuff.

Simulacra levels are a map of the metaphors we use to create metaphoric maps of both territory and the map itself.

The text that coined the term Simulacra levels does not help matters. The term was first referenced locally [by Ben Hoffman in this post](#), but this was not the original source.

The original source of the term is a [super-dense work of French philosophy](#). It requires the reader to pause after every sentence. It's not clear that a proper review would be shorter than the book itself.

Thus, I'm still working through the book. The more I read [Jean Baudrillard](#)'s further assertions, the less they seem deserving of engagement. He is opposed for nonsensical reasons not only to the concept of capitalism, but the concepts of money, value and trade, and even urbanization and mass production. He blames these for the rise of simulacra, whereas they are the primary forces *opposed to* simulacra.

Upon parsing many of his super-dense sentences, I find many of them to be outright false. I find many others to be based on models and frameworks very different from my own, and that are assumed rather than specified in the text. The idea that capitalism isn't the cause of all the world's problems (never mind whether it's the solution) does not seem to parse in his mind. I find many others to be downright absurd, or to be carrying water for the agendas of History's Greatest Villains.

This is a case where I strongly endorse taking the concepts that are useful and leaving the remaining giant mess behind.

Baudrillard's definition will be kept. Beyond that, I'm tossing essentially everything else away.

The goal of this post is to reconcile Baudillard's definition with the Lion definition used in my previous posts, integrating the relation of a simulacra to reality with the motivational explanatory framework. This deals with the dueling definition problem so it doesn't get in the way down the line, and I hope it helps explain why higher level activity systematically has such strange and hostile relationships to physical reality (and what Baudillard calls 'profound' reality.)

## Different Definitions of Simulacra Levels

An additional confusion is that there are now multiple competing definitions, which seem superficially to point at different things, although I claim that both definitions

are fully compatible once properly understood.

First, there are simulacra levels as defined in Baudrillard. The closest thing he gives to a compact full definition is this:

- Such would be the successive phases of the image (as we pass from levels 1 to 4):
  - It is the reflection of a profound reality.
  - It masks and denatures a profound reality.
  - It masks the absence of a profound reality.
  - It has no relation to any reality whatsoever: It is its own pure simulacrum.

Profound reality is a weird term that is doing several distinct things. On a basic level, it means concrete physical reality unmediated by any symbols of any kind. It is what is, full stop. Lying masks and denatures that reality by representing it as something other than what it is, but this is less of an offense than there not being an underlying reality of importance in the third stage, or being cut free from that underlying reality entirely, as we are in the fourth stage.

On a continental philosophical level, there is this idea that anything mass produced, or anything that interacts with money, trade or other systematic motivations, rather than being fully intrinsic and local and spontaneous, or something like that, loses this something vital that Baudrillard calls ‘profound reality’.

I don't think that this second angle is *entirely* nonsense. There is something important that can be distorted or lost when commodification sets in. Polanyi's [The Great Transformation](#) is the best source I know about to make the general case for why this is inherently concerning. It also means that the person interacting with underlying/profound reality then exchanges the fruits of that interaction with someone else in exchange for something symbolic. Rather than gain the physical rewards, those are exchanged for rewards that are based on the symbolic associations of what has been created. This throws away any value for much of the underlying physical reality. The less connected production is with consumption, the bigger this concern.

The difference I have with Baudrillard here is that I do not think this phenomenon is central to what is happening. And I am not eager to dismiss the many-leveled benefits of such systems. The discipline of the market, the need to match demand with satisfactory supply and the reward for doing so, not only are the main ways we have, in historical terms, an insanely great abundance of lots of nice things. They also keep us connected to the underlying reality, and keep our simulacrum (and maze) levels lower.

It is precisely when this market discipline is lost or distorted that things get out of hand. Such systems force upon us symbols at all, bringing us firmly into the first level, as opposed to having no symbols and avoiding the scale entirely. And once on the scale, the slope upwards is slippery. But these forces also form one of our strongest defenses against rising to levels beyond that.

If I was going to write the symbolic description of Simulacra levels in my own words, I would say this:

Level 1: A symbol corresponds to the key elements of underlying physical reality.

Level 2: A symbol pretends to correspond to underlying physical reality, but instead distorts key elements.

Level 3: A symbol pretends to be a distorted version of underlying physical reality (that is in turn pretending to be the underlying physical reality), but instead only corresponds as necessary to maintain the plausibility claim that this is the case.

Level 4: A symbol no longer pretends to be a version of anything other than other symbols. It has no relationship to the underlying physical reality.

Or more compactly:

Level 1: Symbols describe reality.

Level 2: Symbols pretend to describe reality.

Level 3: Symbols pretend to pretend to describe reality.

Level 4: Symbols need not pretend to describe reality.

Or a variation/alternative:

Level 1: Symbols accurately describe reality.

Level 2: Symbols inaccurately describe reality.

Level 3: Symbols claim to describe reality.

Level 4: Symbols no longer claim to describe reality.

A concrete example suggested by Michael Vassar:

Level 1: A court reflects justice.

Level 2: A corrupt judge distorts justice.

Level 3: A Soviet show trial conceals the absence of real Soviet courts.

Level 4: A trial by ordeal or trial by combat lacks and denies the concept of justice entirely.

Contrast that with the newer definition that's based on what it means to say "There's a lion across the river", as described in [Simulacra Levels and their Interactions](#):

Level 1: There's a lion across the river.

Level 2: I don't want to go (or have other people go) across the river.

Level 3: I'm with the popular kids who are too cool to go across the river.

Level 4: A firm stance against trans-river expansionism focus grouped well with undecided voters in my constituency.

Or alternatively, and isomorphic to the Lion definition, from [my previous simulacra post](#):

"There's a pandemic headed our way from China" means...

Level 1: "There's a pandemic headed our way from China."

Level 2: "I want you to act as if you think there might be a pandemic on our way from China" while hoping to still be interpreted by the listener as meaning "There's a pandemic headed our way from China."

Level 3: "I wish to associate with the group that claims there is a pandemic headed our way from China."

Level 4: "It is advantageous for me to say there is a pandemic headed our way from China."

See [the previous post](#) for more details and variations of meaning via this definition.

Careful reading of Baudrillard confirms my suspicion that both definitions point at the same thing while highlighting different aspects.

They grasp different parts of the elephant.

From here on, we will call the levels from the Lion definition L-1, L-2, L-3 and L-4, merged with the Pandemic definition. I will call the levels from the original Baudrillard definition B-1, B-2, B-3 and B-4. The claim is that L-1 = B-1, L-2 = B-2, L-3 = B-3, L-4 = C-4.

This is relatively easy to see, and relatively uncontroversial, for levels 1 and 2. It is less so for level 3 and even less so for level 4.

The key differences are that they deal with lower-level actions versus higher-level systems, and that they deal with cashed-out motivations versus processes.

## **Actions versus Systems**

**The Lion and Pandemic definitions** deal centrally with the *motivations for, and the meanings of, individual statements, actions and systems*.

**The Baudrillard definition** deals centrally with the *default or central interpretation of statements, actions and systems*. It is the expectation of general interpretation and thus of purpose.

One can usefully think of the relative importance of each simulacra level at any scale – of an individual statement or action, of an interaction, of a person or group, of a system or concept, or of a civilization or the world.

Higher levels of grouping and abstraction exist at the simulacra level that is the default motivation and interpretation of their lower-abstraction more discrete actions, statements and systems.

## **Motivations versus Processes**

The Lion definition:

Level 1: There's a lion across the river.

Level 2: I don't want to go (or have other people go) across the river.

Level 3: I'm with the popular kids who are too cool to go across the river.

Level 4: A firm stance against trans-river expansionism focus grouped well with undecided voters in my constituency.

The Baudrillard definition:

Level 1: It is the reflection of a profound reality.

Level 2: It masks and denatures a profound reality.

Level 3: It masks the absence of a profound reality.

Level 4: It has no relation to any reality whatsoever: It is its own pure simulacrum

**The Lion and Pandemic definitions** deal with the motivation behind an action or communication. Was it about communicating true information, updating someone else's a model, sending a symbolic signal or creating useful associations?

**The Baudrillard definition** deals with the process by which the statement relates to what he calls the 'profound reality.' Does it deal with it directly, distort it, hide its absence or ignore it entirely?

These two patterns go hand in hand.

At level 1, one directly deals with reality in order to communicate true information. One does not have to pretend.

At level 2, one distorts reality – in Baudrillard's words, masks and denatures it – in order to convince others that what one is representing as reality is actual reality. One pretends.

At level 3, one hides the absence of reality in order to invoke a symbolic meaning, usually for the purposes of signaling. There needs to be a sufficiently strong sense of association to the underlying reality that the signal is understood and holds meaning, but not so strong an association that the signal can be confused for an underlying map. Reality must be absent, but in a way that is deniable. One pretends to pretend.

At level 4, one engages in pure simulacra, with no relation to the underlying reality at all. There is no object-level cashing out at all. The underlying reality is something to be sculpted by changing associations and symbolic meanings. Consequences of a statement or action are in terms of the consequence to a *simulacra*, as measured on the third and fourth levels. At most, one pretends to be offering pure level-3 simulacra. One does not pretend to pretend to pretend to be on the object level, rather one stops pretending, period.

More than that. One does not merely stop pretending. One *forgets that there was an underlying reality to begin with*, and loses the ability to think about the underlying reality and guide it to better outcomes, except by doing so through agents operating at the levels in between.

## Simulacra Level of a System

All of these are phrased above in terms of an individual, but apply equally to a group, system or civilization.

Even more than people, groups, systems and civilizations have a mix of all levels. Moving up the level chain has instrumental rewards, and happens continuously unless there is sufficient push back. This is similar to [the rise of maze levels](#). Existing systems must maintain some amount of low-level grounding, as well. Without sufficient grounding, doom quickly follows. Things collapse well before the amount of level 1 activity can reach zero.

A person can exist mostly or solely on one level. A larger group almost never does. When I speak below about being at a level, that means the most dominant one. It does not mean that the others are not present.

When a group, system or civilization is still sufficient in level 1, or has regained that footing, its symbols map directly to reality. Words have meaning.

When a group, system or civilization proceeds sufficiently into level 2, that means it is standard and/or wise to assume that level 2 motives and actions are dominant. Claims cease to be taken at face value by default. Trust is destroyed. The assumption is that someone is likely to be out to sell you a bill of goods.

It would not parse that someone would say something because it is true. It would only make sense to say something because it would be beneficial for others to believe it is true. And thus, even straightforward claims are interpreted this way. Still, because all claims must *pretend*, a relatively strong link to the underlying reality is maintained.

When a group, system or civilization proceeds sufficiently into level 3, that means it is standard slash wise to assume that level 3 motives and actions are dominant. Claims cease to be taken not only at face value, but also cease to be taken as being claims about the world at all. It is assumed that claims are made because it is beneficial to be seen making a claim, or for one's side to be seen making such a claim, due to its symbolic benefits.

But this is not a complete transformation. Not yet. Those symbolic benefits still have to be seen as tracing back to underlying 'profound' reality.

[Everybody knows](#), when things reach level three, that statements are made based primarily on their utility in the game. It's no longer a big deal to say that which is not. That's fine. The cool kids don't want to cross the river and this is their slogan, so you repeat the slogan.

The link to the underlying reality is tenuous, but still exists - if one can expose others as not being able to pretend, thus showing they have failed to pretend to pretend, they lose face. The symbols mask the lack of an underlying reality, *but need plausible deniability while doing so*. You need not believe, when claiming there is a lion across the river, that there is a lion across the river. But it would be bad to be seen knowing that there wasn't a lion across the river, or being known to have no reason to think there was a lion across the river, and saying it anyway.

This maintains a weak link to underlying reality, so Level 3 never fully succeeds at existing purely on its own terms. Doing so transforms it into Level 4.

When a group, system or civilization proceeds sufficiently into level 4, that means it is standard slash wise to assume that level 4 motives and actions are dominant. Claims cease to be taken as anything other than symbolic moves. Any impact on the underlying physical reality via their accuracy or lack thereof is purely coincidental.

The institutional memory of the object level is lost in all practical senses. The Level 4 parts of the system can't see the Level 1 parts of the system. A sufficiently strongly Level 4 person, for whom level 4 has become truly part of them, almost always has the same issue. Level 4 doesn't use logic or physical causation, so they've discarded those parts of their system of thought and no longer believe in them.

As I noted in the previous post, Level 4 is especially difficult for myself and many like myself to grok. It seems profoundly alien, perhaps evil.

## Level 0 Exists

It is worth noting here that B-0 and L-0 also exist, are important, and are congruent.

That's where you don't say there is a lion across the river in Lion-0, you simply *don't cross the river because you don't want to get eaten by a lion*. You are also in Symbolic-0, because you don't have a symbol at all. Others can see that, and consider that maybe crossing the river is not a great idea, even if they can't figure out why.

The key is that, on Level 0, that has nothing to do with your decision to not cross. The moment you don't cross because of how others might interpret that decision, you're a symbol, and thus in Symbolic-1, and communicating a map of the world, and thus in Lion-1. L-0 is [what you are in the dark](#). We can call this The Hermit.

It is in this sense that Baudrillard is right to put the 'blame' for simulacra on capitalism or urbanization or mass production. These are the methods by which we have nice things and get to interact with fellow human beings. If one is alone in the forest, there is no need for symbols at all and the equilibrium is stable. The only way to entirely avoid manipulation of symbols is to not have any symbols. The price seems rather high.

## Level 3 and The War Against Knowledge

It is important to note that Level 3 is at war with knowledge.

This is more than the "Level 3 sees knowledge as composed of things it can use for something else." Level 3 is actively destructive of knowledge.

At level 3, the following two things are blameworthy, creating two ways in which knowledge is a liability.

(These two are not everything that is blameworthy, of course. One can be blamed for other things as well – one can be part of the outgroup, be the designated scapegoat, etc etc.)

One blameworthy thing is *not invoking the right symbols*.

This is the "composed of things that can be used for something else" aspect. Caring about what is true creates an alternative incentive that prevents one from invoking the proper symbols, and casts doubt on whether those symbols mean what they seem to mean.

Invoking symbols that are technically false rather than those that are technically true is, if anything, *a stronger move in the game*. This is why. It signals more strongly one's costly sending of the appropriate signals, without room for misinterpretation as a

lower level action. By repeating the lie, we show ourselves loyal. By getting others to repeat it, we drive them towards being and identifying as loyal, and get them to show others they are loyal, and demonstrate our power over both people and symbols.

The other blameworthy thing is *knowing that what you say is false*. What is blameworthy is *knowledge itself*.

(Or, perhaps more precisely, *other people knowing* that you know what you say is false, and thus anyone else's knowledge of our having knowledge, as opposed to knowledge itself, but that's also true of any other blame system.)

What did the President know, and when did he know it?

Thus, the shift in communication from explicit to implicit. The focus on having only deniable, tacit knowledge.

The follower who needs explicit instruction is a poor follower indeed. Specifying everything to be done is impractical, and makes it clear you have not only knowledge but responsibility. Much better to *work towards* the goals of the group, to pile on symbols that help win the game.

Thus does this structure drive everyone away from knowledge. The easiest way, by far, to pretend not to know things is to not know them.

Thus, Level 3 is not merely unconcerned with the profound reality. Level 3 actively symbolizes the *absence* of a profound reality. They are not merely orthogonal to an accurate map. They oppose it.

This situation is not stable. It relies on a lack of common knowledge. It also relies on a lack of individual knowledge. It also doesn't present stable incentives and thus is not an equilibrium.

To be sustained, it requires sufficiently powerful residues stuck in the first two levels, who misunderstand what is going on. This is the force that requires people to pretend to pretend, when their actions are exposed to the public.

When there are insufficient naive forces to appeal to or worry about, the mask of pretending is dropped. People stop pretending to pretend.

"Facts don't matter" is true at both levels three and four. But *acknowledging* that "facts don't matter" and *creating common knowledge* of this will make short work of level three. It moves all actors up to levels three and four. This shatters the link between symbol and reality entirely. This moves us collectively to Level 4.

Level 4 is then not at war with knowledge the way level 3 is at war with knowledge. Level 4 doesn't acknowledge that knowledge is a thing. Thus, there is no need to symbolize its absence.

## **Conclusion and the Unity of Level 4**

This hopefully provided additional perspective on simulacra levels. Ideally it provided at least some justification for the additional associations and implications I've placed upon levels three and four, and made clear how I think the Lion definition integrates with the original definition.

I hope the whole system is also looking less like an elegant  $2 \times 2$  with extra weird stuff piled on top of it that seems like it has an axe to grind, and more like a coherent system. In particular, I hope that it is now clearer why level 3 actively opposes knowledge, and level 4 loses access to logic and the ability to observe and analyze and optimize the physical world.

I hope that this will all become clearer as these posts continue. I'm especially excited by the next one, but felt I needed to get this one out of the way first, as the confusions it tries to clear up would otherwise have gotten in the way. It was necessary to tackle it first.

# Measuring hardware overhang

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

# Measuring hardware overhang

## Summary

How can we measure a potential AI or hardware overhang? For the problem of chess, modern algorithms gained two orders of magnitude in compute (or ten years in time) compared to older versions. While it took the supercomputer "Deep Blue" to win over world champion Gary Kasparov in 1997, today's Stockfish program achieves the same ELO level on a 486-DX4-100 MHz from 1994. In contrast, the scaling of neural network chess algorithms to slower hardware is worse (and more difficult to implement) compared to classical algorithms. Similarly, future algorithms will likely be able to better leverage today's hardware by 2-3 orders of magnitude. I would be interested in extending this scaling relation to AI problems other than chess to check its universality.

## Introduction

Hardware overhang is a situation where sufficient compute is available, but the algorithms are suboptimal. It is relevant if we build AGI with large initial build, but cheaper run costs. Once built, the AGI might run on many comparably slow machines. That's a hardware overhang with a risk of exponential speed-up. This asymmetry exists for current neural networks: Creating them requires [orders of magnitude](#) more compute than running them. On the other hand, in [The Bitter Lesson](#) by Rich Sutton it is argued that the increase in computation is much more important (orders of magnitude) than clever algorithms (factor of two or less). In the following, I will examine the current state of the algorithm-art using chess as an example.

## The example of chess

One of the most well-researched AI topics is chess. It has a long history of algorithms going back to a program on the [1956 MANIAC](#). It is comparatively easy to measure the quality of a player by its [ELO score](#). As an instructive example, we examine the most symbolic event in computer chess. In 1997, the IBM supercomputer "Deep Blue" defeated the reigning world chess champion under tournament conditions. The win was taken as a sign [that artificial intelligence was catching up to human intelligence](#).

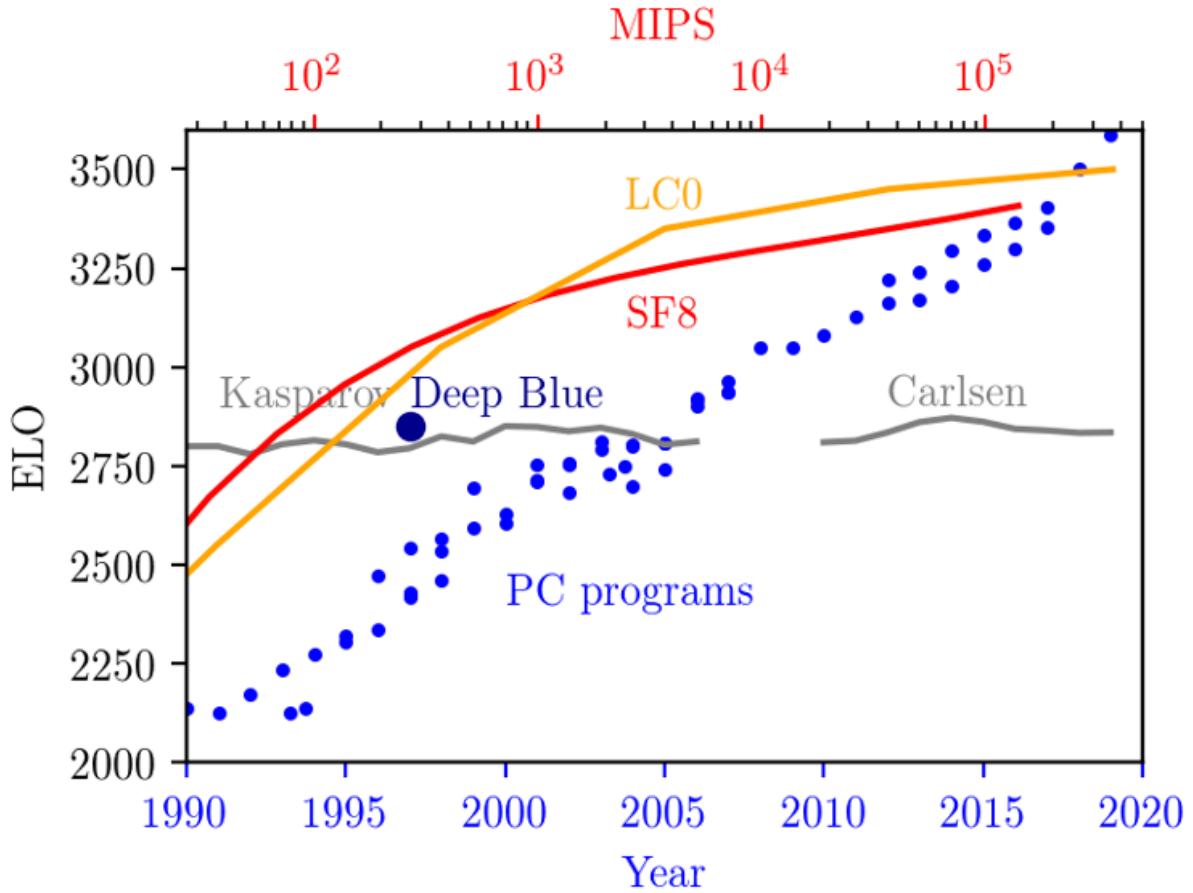
By today's standards, Deep Blue used simple algorithms. Its strength came from computing power. It was a RS/6000-based system with 30 nodes, each with a 120 MHz CPU plus 480 special purpose VLSI chess chips. For comparison, a common computer at the time was the Intel Pentium II at 300 MHz.

# Method: An experiment using a 2020 chess engine

We may wonder: How do modern (better) chess algorithms perform on slower hardware? I tested this with [Stockfish](#) version 8 (SF8), one of the strongest classical chess engine. I simulated 10k matches of SF8 against slower versions of itself and a series of [older engines for calibration](#), using [cutechess-cli](#). In these benchmarks, I varied the total number of nodes to be searched during each game. I kept the RAM constant (this may be unrealistic for very old machines, see below). By assuming a fixed thinking time per game, the experiments scale out to slower machines. By cross-correlating various old benchmarks of Stockfish and other engines on older machines, I matched these ratings to units of MIPS; and finally, [MIPS approximately to the calendar year](#). Depending on the actual release dates of the processors, the year axis has a jitter up to 2 years. I estimate the error for the compute estimates to be perhaps 20%, and certainly less than 50%. As we will see, the results measure in orders of magnitude, so that these errors are small in comparison (<10%).

## Results

SF8 achieves Kasparov's 2850 ELOs running on a 486-100 MHz introduced in 1994, three years before the Kasparov-Deep Blue match. These ELOs refer to tournament conditions as in the 1997 IBM games. In other words, with today's algorithms, computers would have beat the world world chess champion already in 1994 on a contemporary desk computer (not a supercomputer).



The full scaling relation is shown in the Figure. The gray line shows the ELO rating of Kasparov and Carlsen over time, hovering around 2800. The blue symbols indicate the common top engines at their times. The plot is linear in time, and logarithmic in compute. Consequently, ELO scales approximately with the square of compute. Finally, the red line shows the ELOs of SF8 as a function of compute. Starting with the 2019 rating of  $\sim 3400$  points, it falls below 3000 when reducing MIPS from  $10^5$  to a few times  $10^3$ . This is a decrease of 2-3 orders of magnitude. It falls below 2850 ELO, Kasparov's level, at 68 MIPS. For comparison, the 486 achieves 70 MIPS at 100 MHz. At its maximum, the hardware overhang amounts to slightly more than 10 years in time, or 2-3 orders of magnitude in compute. Going back very far (to the era of 386-PCs), the gap reduces. This is understandable: On very slow hardware, you can't do very much, no matter what algorithm you have. The orange line shows the scaling relation of a neural network-based chess engine, Leela Chess Zero (LC0), as discussed below.

## Discussion

I originally ran these tests in 2019. Now (August 2020), SF8 has been superseded by SF11, with another  $\sim 150$  ELO increase (at today's speed). It remains unclear how much improvement is left for future algorithms when scaled down to a 486-100. I strongly suspect, however, that we're running into diminishing returns here. There is only so much you can do on a "very slow" machine; improvements will never go to infinity. My guess is that the scaling will remain below three orders of magnitude.

Neural network-based algorithms such as [AlphaZero](#) or [Leela Chess Zero](#) can outperform classical chess engines. However, for this comparison they are less suited. I find that their scaling is considerably worse, especially when not using GPUs. In other words, they do not perform well on CPUs of slower machines. Depending on the size of the neural net, older machines may even be incapable of executing it. In principle, it would be very interesting to make this work: Train a network on today's machines, and execute (run) it on a very old (slow) machine. But with current algorithms, the scaling is worse than SF8. As a reference point, LC0 achieves ~3000 ELOs on a Pentium 200 under tournament conditions; SF8 is at the same level with about half the compute.

## Conclusion and future research proposals

Similarly, scaling of other NN algorithms to slower hardware (with less RAM etc.) should yield interesting insights. While x86 CPUs are in principle backwards-compatible since the 1980s, there are several breaking changes which make comparisons difficult. For example, the introduction of modern GPUs produces a speed gap when executing optimized algorithms on CPUs. Also, older 32-bit CPUs are capped at 4 GB of RAM, making execution of larger models impossible. Looking into the future, it appears likely that similar breaking changes will occur. One recent example is the introduction of TPUs and/or GPUs with large amounts of RAM. Without these, it may be impossible to execute certain algorithms. If AGI relies on similar (yet unknown) technologies, the hardware overhang is reduced until more of such the units are produced. Then, the vast amount of old (existing) compute can not be used.

I would be interested in researching this scaling relation for other problems outside of chess, such as voice and image recognition. Most problems are harder to measure and benchmark than chess. Will the scalings show a similar 2-3 orders of magnitude software overhang? Most certainly, many problems will show similar diminishing returns (or a cap) due to RAM restrictions and wait time. For example, you just can't run a self-driving car on an Atari, no matter how good the algorithms. I would be interested in researching the scaling for other AI and ML fields, possibly leading to an academic paper.

# Introduction To The Infra-Bayesianism Sequence

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

TLDR: Infra-Bayesianism is a new approach to epistemology / decision theory / reinforcement learning theory, which builds on "imprecise probability" to solve the problem of prior misspecification / grain-of-truth / nonrealizability which plagues Bayesianism and Bayesian reinforcement learning. Infra-Bayesianism also naturally leads to an implementation of UDT, and (more speculatively at this stage) has applications to multi-agent theory, embedded agency and reflection. This post is the first in a sequence which lays down the foundation of the approach.

## Prelude:

Diffractor and Vanessa proudly present: The thing we've been working on for the past five months. I initially decided that Vanessa's scattered posts about incomplete models were interesting, and could benefit from being written up in a short centralized post. But as we dug into the mathematical details, it turned out it didn't really work, and then Vanessa ran across the true mathematical thing (which had previous ideas as special cases) and scope creep happened.

This now looks like a new, large, and unusually tractable vein of research. Accordingly, this sequence supersedes all previous posts about incomplete models, and by now we've managed to get quite a few interesting results, and have ideas for several new research directions.

Diffractor typed everything up and fleshed out the proof sketches, Vanessa originated almost all of the ideas and theorems. It was a true joint effort, this sequence would not exist if either of us were absent. Alex Mennen provided feedback on drafts to make it much more comprehensible than it would otherwise be, and Turntrout and John Maxwell also helped a bit in editing.

Be aware this sequence of posts has the math textbook issue where it requires loading a tower of novel concepts that build on each other into your head, and cannot be read in a single sitting. **We will be doing a group readthrough on MIRIxDiscord where we can answer questions and hopefully get collaborators, PM me to get a link.**

## Introduction:

Learning theory traditionally deals with two kinds of setting: "realizable" and "agnostic" or "non-realizable". In realizable settings, we assume that the environment can be described perfectly by a hypothesis inside our hypothesis space. (AIXI is an example of this) We then expect the algorithm to converge to acting as if it already knew the correct hypothesis. In non-realizable settings, we make no such assumption. We then expect the algorithm to converge to the best approximation of the true environment within the available hypothesis space.

As long as the computational complexity of the environment is greater than the computational complexity of the learning algorithm, the algorithm cannot use an easy-to-

compute hypothesis that would describe the environment perfectly, so we are in the nonrealizable setting. When we discuss AGI, this is necessarily the case, since the environment is the entire world: a world that, in particular, contains the agent itself and can support other agents that are even more complex, much like how halting oracles (which you need to run Solomonoff Induction) are nowhere in the hypotheses which Solomonoff considers. Therefore, the realizable setting is usually only a toy model. So, instead of seeking guarantees of good behavior assuming the environment is easy to compute, we'd like to get good behavior simply assuming that the environment has some easy-to-compute properties that can be exploited.

For offline and online learning there are classical results in the non-realizable setting, in particular VC theory naturally extends to the non-realizable setting. However, for reinforcement learning there are few analogous results. Even for passive Bayesian inference, the best non-realizable result found in our literature search is [Shalizi's](#) which relies on ergodicity assumptions about the true environment. Since reinforcement learning is the relevant setting for AGI and alignment theory, this poses a problem.

Logical inductors operate in the nonrealizable setting, and the general reformulation of them in [Forecasting Using Incomplete Models](#) is of interest for broader lessons applicable to acting in an unknown environment. In said paper, reality can be drawn from any point in the space of probability distributions over infinite sequences of observations,  $\Delta(O^\omega)$ . Almost all of the points in this space aren't computable, and because of that, we shouldn't expect convergence to the true environment, as occurs in the realizable setting where the true environment lies in your hypothesis space.

However, even if we can't hope to learn the *true* environment, we can at least hope to learn some *property* of the true environment, like "every other bit is a 0", and have our predictions reflect that if it holds. A hypothesis in this setting is a closed convex subset of  $\Delta(O^\omega)$  which can be thought of as "I don't know what the true environment is, but it lies within this set". The result obtained in the above-linked paper was, if we fix a countable family of properties that reality may satisfy, and define the inductor based on them, then for all of those which reality fulfills, the predictions of the inductor converge to that closed convex set and so fulfill the property in the limit.

## What About Environments?

However, this just involves sequence prediction. Ideally, we'd want some space that corresponds to environments that you can interact with, instead of an environment that just outputs bits. And then, given a suitable set  $B$  in it... Well, we don't have a fixed environment to play against. The environment could be *anything*, even a worst-case one within  $B$ . We have Knightian uncertainty over our set of environments, it is *not* a probability distribution over environments. So, we might as well go with the maximin policy.

$$\operatorname{argmax}_\pi \inf_{e \in B} (E_{\pi \cdot e} [U])$$

Where  $\pi \cdot e$  is the distribution over histories produced by policy  $\pi$  interacting with environment  $e$ .  $U$  is just some utility function.

When we refer to "Murphy", this is referring to whatever force is picking the worst-case environment to be interacting with. Of course, if you aren't playing against an adversary, you'll do better than the worst-case utility that you're guaranteed. Any provable guarantees come in the form of establishing lower bounds on expected utility if a policy is selected.

The problem of generating a suitable space of environments was solved in [Reinforcement Learning With Imperceptible Rewards](#). If two environments are indistinguishable by any policy they are identified, a mixture of environments corresponds to picking one of the component environments with the appropriate probability at the start of time, and there was a notion of update.

However, this isn't good enough. We could find no good update rule for a set of environments, we had to go further.

Which desiderata should be fulfilled to make maximin policy selection over a set of environments (actually, we'll have to generalize further than this) to work successfully? We'll have three starting desiderata.

**Desideratum 1:** There should be a sensible notion of what it means to update a set of environments or a set of distributions, which should also give us dynamic consistency. Let's say we've got two policies,  $\pi$  and  $\pi'$  which are identical except they differ after history  $h$ . If, after updating on history  $h$ , the continuation of  $\pi'$  looks better than the continuation of  $\pi$ , then it had better be the case that, viewed from the start,  $\pi'$  outperforms  $\pi$ .

**Desideratum 2:** Our notion of a hypothesis (set of environments) in this setting should collapse "secretly equivalent" sets, such that any two distinct hypotheses behave differently in *some* relevant aspect. This will require formalizing what it means for two sets to be "meaningfully different", finding a canonical form for an equivalence class of sets that "behave the same in all relevant ways", and then proving some theorem that says we got everything.

**Desideratum 3:** We should be able to formalize the "Nirvana trick" (elaborated below) and cram any UDT problem where the environment cares about what you *would do*, into this setting. The problem is that we're just dealing with sets of environments which only depend on what you do, not what your policy is, which hampers our ability to capture policy-dependent problems in this framework. However, since Murphy looks at your policy and then picks which environment you're in, there *is* an acausal channel available for the choice of policy to influence which environment you end up in.

The "Nirvana trick" is as follows. Consider a policy-dependent environment, a function  $\Pi \times (A \times O)^{<\omega} \times A \rightarrow \Delta O$  (ie, the probability distribution over the next observation depends on the history so far, the action you selected, and your policy). We can encode a policy-dependent environment as a set of policy-independent environments that don't care about your policy, by hard-coding every possible deterministic policy into the policy slot, making a family of functions of type  $(A \times O)^{<\omega} \times A \rightarrow \Delta O$ , which is the type of policy-independent

environments. It's similar to taking a function  $f(x, y)$ , and plugging in all possible  $x$  to get a family of functions that only depend on  $y$ .

Also, we will impose a rule that, if your action ever violates what the hard-coded policy predicts you do, you attain Nirvana (a state of high or infinite reward). Then, Murphy, when given this set of environments, will go "it'd be bad if they got high or infinite reward, thus I need to pick an environment where the hard-coded policy matches their *actual* policy". When playing against Murphy, you'll act like you're selecting a policy for an environment that *does* pay attention to what policy you pick. As-stated, this doesn't quite work, but it can be repaired.

There's two options. One is making Nirvana count as infinite reward. We will advance this to a point where we can capture any UDT/policy-selection problem, at the cost of some mathematical ugliness. The other option is making Nirvana count as 1 reward forever afterward, which makes things more elegant, and it is much more closely tied to learning theory, but that comes at the cost of only capturing a smaller (but still fairly broad) class of decision-theory problems. We will defer developing that avenue further until a later post.

## A Digression on Deterministic Policies

We'll be using deterministic policies throughout. The reason for using deterministic policies instead of probabilistic policies (despite the latter being a larger class), is that the Nirvana trick (with infinite reward) doesn't work with probabilistic policies. Also, probabilistic policies don't interact well with embeddedness, because it implicitly assumes that you have a source of random bits that the rest of the environment can never interact with (except via your induced action) or observe.

Deterministic policies can emulate probabilistic policies by viewing probabilistic choice as deterministically choosing a finite bitstring to enter into a random number generator (RNG) in the environment, and then you get some bits back and act accordingly.

However, we aren't assuming that the RNG is a good one. It could be insecure or biased or nonexistent. Thus, we can model cases like Death In Damascus or Absent-Minded Driver where you left your trusty coin at home and don't trust yourself to randomize effectively. Or a nanobot that's too small to have a high bitrate RNG in it, so it uses a fast insecure PRNG (pseudorandom number generator). Or game theory against a mindreader that can't see your RNG, just the probability distribution over actions you're using the RNG to select from, like an ideal CDT opponent. It can also handle cases where plugging certain numbers into your RNG chip cause lots of heat to be released, or maybe the RNG is biased towards outputting 0's in strong magnetic fields. Assuming you have a source of true randomness that the environment can't read isn't general enough!

## Motivating Sa-Measures

Sets of probability distributions or environments aren't enough, we need to add in some extra data. This can be best motivated by thinking about how updates should work in order to get dynamic consistency.

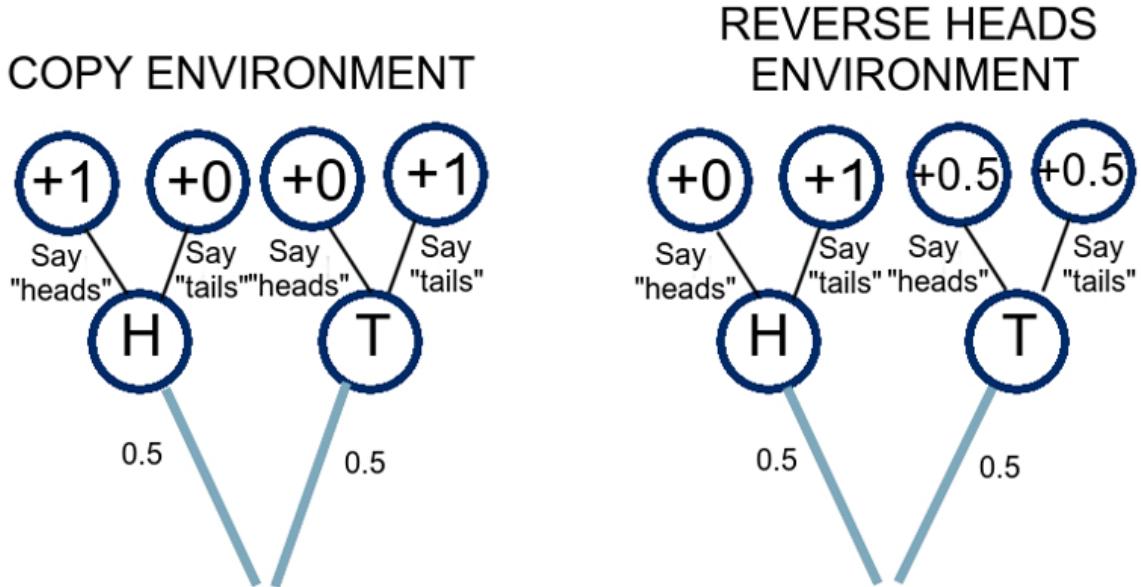
Throughout, we'll be using a two-step view of updating, where first, we chop down the measures accordingly (the "raw update"), and then we renormalize back up to 1.

So, let's say we have a set of two probability distributions  $\mu_1$  and  $\mu_2$ . We have Knightian uncertainty within this set, we genuinely don't know which one will be selected, it may even be adversarial.  $\mu_1$  says observation o has 0.5 probability,  $\mu_2$  says observation o has 0.01 probability. And then you see observation o! The wrong way to update would be to go "well, both probability distributions are consistent with observed data, I guess I'll update them individually and resume being completely uncertain about which one I'm in", you don't want to ignore that one of them assigns 50x higher probability to the thing you just saw.

However, neglecting renormalization, we can do the "raw update" to each of them individually, and get  $m_1$  and  $m_2$  (finite measures, not probability distributions), where  $m_1$  has 0.5 measure and  $m_2$  has 0.01 measure.

Ok, so instead of a set of *probability distributions*, since that's insufficient for updates, let's consider a set of measures  $m$ , instead. Each individual measure in that set can be viewed as  $\lambda\mu$ , where  $\mu$  is a probability distribution, and  $\lambda \geq 0$  is a scaling term. Note that  $\lambda$  is not uniform across your set, it varies depending on which point you're looking at.

However, this *still* isn't enough. Let's look at a toy example for how to design updating to get dynamic consistency. We'll see we need to add *one more* piece of data. Consider two environments where a fair coin is flipped, you see it and then say "heads" or "tails", and then you get some reward. The COPY Environment gives you 0 reward if you say something different than what the coin shows, and 1 reward if you match it. The REVERSE HEADS Environment always gives you 0.5 reward if the coin comes up tails, but if it comes up heads, saying "tails" gets you 1 reward and "heads" gets you 0 reward. We have Knightian uncertainty between the two environments.



For finding the optimal policy, we can observe that saying "tails" when the coin is tails helps out in COPY and doesn't harm you in REVERSE HEADS, so that's a component of an optimal policy.

Saying "tails" no matter what the coin shows means you get  $0.5 \cdot 0 + 0.5 \cdot 1 = 0.5$  utility on COPY, and  $0.5 \cdot 1 + 0.5 \cdot 0.5 = 0.75$  utility on REVERSE HEADS. Saying "tails" when the coin is tails and "heads" when the coin is heads means you get  $0.5 \cdot 1 + 0.5 \cdot 1 = 1$  utility on COPY and  $0.5 \cdot 0 + 0.5 \cdot 0.5 = 0.25$  utility on REVERSE HEADS. Saying "tails" no matter what has a better worst-case value, so it's the optimal maximin policy.

Now, if we see the coin come up heads, how should we update? The *wrong* way to do it would be to go "well, both environments are equally likely to give this observation, so I've got Knightian uncertainty re: whether saying heads or tails gives me 1 or 0 utility, both options look equally good". This is because, according to past-you, regardless of what you did upon seeing the coin come up "tails", the maximin expected values of saying "heads" when the coin comes up heads, and saying "tails" when the coin comes up heads, are unequal. Past-you is yelling at you from the sidelines not to just shrug and view the two options as equally good.

Well, let's say you *already* know that you would say "tails" when the coin comes up tails and are trying to figure out what to do now that the coin came up heads. The proper way to reason through it is going "I have Knightian uncertainty between COPY which has 0.5 expected utility assured off-history since I say "tails" on tails, and REVERSE HEADS, which has 0.25 expected utility assured off-history. Saying "heads" now that I see the coin on heads would get me  $(0.5 \times 1) + 0.5 = 1$  expected utility in COPY and  $(0.5 \times 0) + 0.25 = 0.25$  utility in REVERSE HEADS, saying "tails" would get me  $(0.5 \times 0) + 0.5 = 0.5$  utility in COPY and  $(0.5 \times 1) + 0.25 = 0.75$  utility in REVERSE HEADS, I get higher worst-case value by saying "tails"." And then you agree with your past self re: how good the various decisions are.

Huh, the proper way of doing this update to get dynamic consistency requires keeping track of the fragment of expected utility we get off-history.

Similarly, if you messed up and precommitted to saying "heads" when the coin comes up tails (a bad move), we can run through a similar analysis and show that keeping track of the expected utility off-history leads you to take the action that past-you would advise, after seeing the coin come up heads.

So, with the need to keep track of that fragment of expected utility off-history to get dynamic consistency, it isn't enough to deal with finite measures  $m$ , that still isn't keeping track of the information we need. What we need is  $(m, b)$ , where  $m$  is a finite measure, and  $b$  is a number  $\geq 0$ . That  $b$  term keeps track of the expected value off-history so we make the right decision after updating. (We're glossing over the distinction between probability distributions and environments here, but it's inessential)

We will call such a  $(m, b)$  pair an "affine measure", or "a-measure" for short. The reason for this terminology is because a measure can be thought of as a linear function from the space of continuous functions to  $\mathbb{R}$ . But then there's this  $+b$  term stuck on that acts as utility, and a linear function plus a constant is an affine function. So, that's an a-measure. A pair of a finite measure and a  $b$  term where  $b \geq 0$ .

But wait, we can go even further! Let's say our utility function of interest is bounded. Then we can do a scale-and-shift until it's in  $[0, 1]$ .

Since our utility function is bounded in  $[0, 1]$ ... what would happen if you let in measures with negative parts, but only if they're paired with a sufficiently large  $b$  term? Such a thing is called an  $sa$ -measure, for signed affine measure. It's a pair of a finite signed measure and a  $b$  term that's as-large-or-larger than the amount of negative measure present. No matter your utility function, even if it assigns 0 reward to outcomes with positive measure and 1 reward to outcomes with negative measure, you're still assured nonnegative expected value because of that  $+b$  term. It turns out we actually *do* need to expand in this direction to keep track of equivalence between sets of  $a$ -measures, get a good tie-in with convex analysis because signed measures are dual to continuous functions, and have elegant formulations of concepts like minimal points and the upper completion.

Negative measures may be a bit odd, but as we'll eventually see, we can ignore them and they only show up in intermediate steps, not final results, much like negative probabilities in quantum mechanics. And if negative measures ever become relevant for an application, it's effortless to include them.

## Belief Function Motivation

Also, we'll have to drop the framework we set up at the beginning where we're considering sets of environments, because working with sets of environments has redundant information. As an example, consider two environments where you pick one of two actions, and get one of two outcomes. In environment  $e_0$ , regardless of action, you get outcome 0. In environment  $e_1$ , regardless of action, you get outcome 1. Then, we should be able to freely add an environment  $e_2$ , where action 0 implies outcome 0, and where action 1 implies outcome 1. Why?

Well, if your policy is to take action 0,  $e_2$  and  $e_0$  behave identically. And if your policy is to take action 1,  $e_2$  and  $e_1$  behave identically. So, adding an environment like this doesn't affect anything, because it's a "chameleon environment" that will perfectly mimic *some* preexisting environment regardless of which policy you select. However, if you consider the function mapping an action to the set of possible probability distributions over outcomes, adding  $e_2$  didn't change that at all. Put another way, if it's impossible to distinguish in any way whether an environment was added to a set of environments because no matter what you do it mimics a preexisting environment, we might as well add it, and seek some alternate formulation instead of "set of environments" that doesn't have the unobservable degrees of freedom in it.

To eliminate this redundancy, the *true* thing we should be looking at isn't a set of environments, but the "belief function" from policies to sets of probability distributions over histories. This is the function produced by having a policy interact with your set of environments and plotting the probability distributions you could get. Given certain conditions on a belief function, it is possible to recover a set of environments from it, but

belief functions are more fundamental. We'll provide tools for taking a wide range of belief functions and turning them into sets of environments, if it is desired.

Well, actually, from our previous discussion, sets of probability distributions are insufficient, we need a function from policies to sets of sa-measures. But that's material for later.

## Conclusion

So, our fundamental mathematical object that we're studying to get a good link to decision theory is not sets of probability distributions, but sets of sa-measures. And instead of sets of environments, we have functions from policies to sets of sa-measures over histories. This is because probability distributions alone aren't flexible enough for the sort of updating we need to get dynamic consistency, and in addition to this issue, sets of environments have the problem where adding a new environment to your set can be undetectable in any way.

In the next post, we build up the basic mathematical details of the setting, until we get to a duality theorem that reveals a tight parallel between sets of sa-measures fulfilling certain special properties, and probability distributions, allowing us to take the first steps towards building up a version of probability theory fit for dealing with nonrealizability. There are analogues of expectation values, updates, renormalizing back to 1, priors, Bayes' Theorem, Markov kernels, and more. We use the "infra" prefix to refer to this setting. An infradistribution is the analogue of a probability distribution. An infrakernel is the analogue of a Markov kernel. And so on.

The post after that consists of extensive work on belief functions and the Nirvana trick to get the decision-theory tie-ins, such as UDT behavior while still having an update rule, and the update rule is dynamically consistent. Other components of that section include being able to specify your entire belief function with only part of its data, and developing the concept of Causal, Pseudocausal, and Acausal hypotheses. We show that you can encode almost any belief function as an Acausal hypothesis, and you can translate Pseudocausal and Acausal hypotheses to Causal ones by adding Nirvana appropriately (kinda). And Causal hypotheses correspond to actual sets of environments (kinda). Further, we can mix belief functions to make a prior, and there's an analogue of Bayes for updating a mix of belief functions. We cap it off by showing that the starting concepts of learning theory work appropriately, and show our setting's version of the Complete Class Theorem.

Later posts (not written yet) will be about the "1 reward forever" variant of Nirvana and InfraPOMDP's, developing inframeasure theory more, applications to various areas of alignment research, the internal logic which infradistributions are models of, unrealizable bandits, game theory, attempting to apply this to other areas of alignment research, and... look, we've got a lot of areas to work on, alright?

If you've got the relevant math skills, as previously mentioned, you should PM me to get a link to the MIRIxDiscord server and participate in the group readthrough, and you're more likely than usual to be able to contribute to advancing research further, there's a lot of shovel-ready work available.

## Links to Further Posts:

- [Basic Inframeasure Theory](#)
  - [Proofs 1.1](#)
  - [Proofs 1.2](#)
- [Belief Functions and Decision Theory](#)
  - [Proofs 2.1](#)
  - [Proofs 2.2](#)
  - [Proofs 2.3](#)
- [Less Basic Inframeasure Theory](#)
  - [Proofs 1](#)
  - [Proofs 2](#)
  - [Proofs 3](#)
  - [Proofs 4](#)
  - [Proofs 5](#)
  - [Proofs 6](#)
  - [Proofs 7](#)
  - [Proofs 8](#)
- [Inframeasures and Domain Theory](#)
  - [Infra-Domain Proofs 1](#)
  - [Infra-Domain Proofs 2](#)
- [The Many Faces of Infra-Beliefs](#)
  - [Proofs T1](#)
  - [Proofs T2,3](#)
  - [Proofs T4](#)
  - [Proofs T5](#)
  - [Proofs T6-8](#)
- [Infra-Bayesian Physicalism: a formal theory of naturalized induction](#)
  - [IBP Proofs 1](#)
  - [IBP Proofs 2](#)

# How to teach things well

This is a linkpost for <https://www.neelnanda.io/blog/mini-blog-post-18-how-to-teach-things-well>

(This is a post on my thoughts on good teaching techniques from a daily blogging project, that I thought might be of interest to LessWrong readers)

## Introduction

This is a blog post on how to teach things well. I'll mostly be focusing on forms of teaching that involve preparation and structure, like talks and tutoring, but these ideas transfer pretty broadly. I think teaching and explaining ideas is an *incredibly* important skill, and one that most people aren't great at. I've spent a lot of time practicing teaching ideas, and I think I've found a bunch of important ideas and approaches that work well. I'm giving a talk next week, so I'll initially focus on how to give good talks, but try to outline the underlying concepts and high-level ideas of teaching. And then talk about how these can transfer to contexts like tutoring, and to teaching specifically maths or applied rationality - the main areas I have actual teaching experience with.

Note: I mostly care about teaching concepts and ideas, and teaching things to people who genuinely want to learn and be there, so my advice will focus accordingly.

I think it's useful to think about good teaching even if you don't intend to spend much time teaching - learning and teaching are flip sides of the same process. I've found that even when in the role of a student, understanding what good teaching looks like can often fix a lot of the shortcomings of a bad teacher!

## Framing

The key insight of this post is that good teaching requires you to [be deliberate](#), and keep the purpose in mind: **learning is a process of information compression**. When you're learning something new, you essentially receive a stream of new information. But human cognition doesn't work by just storing a flood of information as is. The student takes in the information stream, extracts out the important ideas, converts it to *concepts*, and stores those in their mind. This is a key distinction, because it shows that the job of a teacher is *not* to give the student information, it's to **get the student to understand the right concepts**. Conveying information is only useful as a means to an end to this goal.

In practice, it often works to just give a stream of information! Good students have learned the skill of taking streams of information and converting it to concepts. Often this happens implicitly, they student will absorb and memorise a lot of data, and over time this forms into concepts and ideas in their head automatically. But this is a *major* amount of cognitive labour. And a good teacher will try to do as much it as possible, to let the student focus their cognitive labour on the important things.

My underlying model here is that we all have a web of concepts in our minds, our **knowledge graph**. The collection of all the concepts we understand, all of our

existing knowledge and intuitions, connected together. And you have learned something when you can convert it to concepts and **connect it to your existing understanding**. This means not just understanding the concept itself, but understanding where it fits into the bigger picture, where to use it, etc.

The final part there is key - if the student leaves with a good understanding of the ideas in the abstract, but no idea when to think about the ideas again, it's no better than if they'd learned nothing at all. We call on our knowledge when something related triggers, so in order for a lesson to be useful, you *need* to build those connections and triggers in the student's mind.

A key distinction to bear in mind is ideas being **legible** vs **tacit**. A legible idea is something concrete that can easily be put into unambiguous words, eg how to do integration by parts. While tacit knowledge is something fuzzier and intuitive, eg recognising the kinds of integrals where you'd use integration by parts in the first place - essentially the intuitions you want the student to have. This is a good distinction to bear in mind, because legible knowledge is *much* easier to convey, but often your goal is to convey tacit knowledge (at least, it should be!). And there's a lot of skill to conveying tacit knowledge well, and making it as legible as possible without losing key nuance. And different techniques work better for the two kinds. A lot of my issues with the Cambridge maths course is an extreme focus on legible knowledge over tacit - the underlying intuitions and motivations.

## How to teach

There are two key problems when teaching, that any good teaching advice must account for:

- Limited ways to convey information
  - The ideas I'm teaching are stored as implicit concepts in my mind, but in order to convey them, I must translate them into language. This language maps to ideas in *my* head according to all of my implicit knowledge and worldview, but translates into the student's head according to *their* implicit knowledge and worldview. This often creates errors
  - And converting concepts to language is *inherently lossy*, ideas have a lot of tacit nuance that is hard to capture
  - Essentially, words can only convey legible knowledge, and I need to figure out how to hack this to convey tacit knowledge. Or how to find alternate information channels
  - Alternately, I need to have error checking mechanisms to *notice* when I've failed to convey something well.
- [Typical Mind Fallacy](#)
  - A key part of learning is combining knowledge you hear with the ideas already in your head.
  - But I only have access to what's in *my* head, not what's in the student's. And by definition this is different - I already understand what I'm teaching!
  - This is a crippling blow to my ability to teach, and I need to be constantly aware of this and trying to build models of what's in the student's heads, and how they receive what I'm saying.

Here are some of the most important tools I have for addressing these problems:

- High-level picture

- The student's knowledge graph is *big*. So the first, and most important part, is identifying which part of the graph they should add these ideas to
- Thus you should *always* highlight where this fits in to the bigger picture. Which questions are we currently trying to answer? Why is this idea interesting at all? Where can the student use this? What are its limitations?
- This should *always* be the first thing when introducing a new idea
- Prioritising information
  - Learning is information compression. This fundamentally means that the student needs to be paying selective attention. When learning something new, the [Pareto Principle](#) always applies - 80% of the importance lies in 20% of the ideas.
  - Identifying this 20% is significant cognitive labour, because it's not immediately obvious. They need to pay attention to *everything* and later filter.
  - But, the question of "what matters here" is tacit knowledge that I already have! This talk is high labour for the student, but easy for me. Thus, the most important thing a teacher can possibly do is to highlight what matters and what doesn't, to tell the student where to focus their attention.
  - In practice, you should *always* be saying "this is really, really important" or "these are just fiddly details" or "this is a bit of a niche edge case" etc. It is *extremely hard* to do this too often.
    - And give more *time* to the important things. If you say an important point, write it down and put a box around it. Explain *why* it's important and how it fits into the bigger picture. Give an alternate explanation, or an example.
    - This is really easy to miss if you think of teaching as information *transfer* - where your goal is to tell the student everything and hope they figure out what to pay attention to themselves.
  - Another tool: Have frequent summaries, highlighting the key points in the previous section.
    - This is further useful to highlight *connections*. Some ideas will fit into their knowledge graph more easily than others, and pointing out connections can leverage the easy connections to make hard ones easier
  - I'm a big of the advice "say what you're going to tell them, tell them, and say what you've just told them", I think it's a good way to implement this principle
  - The fundamental principle behind this is that students retain a tiny fraction of what they hear. If you give a one hour talk, they'll retain maybe a few minutes worth of content. This is a fundamental fact of the learning process, and the only thing you can do about it is to *control* what they retain. Focus their attention on the important parts
    - This mindset helps me *identify* what's important. Set a 5 minute timer, and write down everything important that you want people to retain. These are the key points around which your talk should be structured!
    - Everything else you say should be intended to help these key points stick better - to highlight connections between them and to existing knowledge, to ensure good information transfer, and to convey the tacit knowledge underlying the key points
  - Another key skill of prioritisation is *cutting things*. If one part is irrelevant, or it's boring and fiddly, cut it! It's painful to not talk about everything cool,

- but you have limited time - if you don't actively prioritise, you aren't avoiding trade-offs, you're just ceding control to "whatever you leave last"
- Anyone who's gone to a talk by me knows that I have yet to internalise this lesson
  - If there's one point you retain from this post, let it be this one - this is *incredibly* important, and the main mark of a good teacher vs a mediocre one
  - Understand pre-requisites
    - A consequence of the Typical Mind Fallacy, is that it's super easy to forget that your students don't have all the context you have! This manifests as people teaching ideas without the pre-requisites.
      - The underlying idea: the ideas you teach are in *your* knowledge graph, and are built upon existing ideas - these are the pre-requisites. You need to figure out whether the students
    - A common secondary mistake is to understand pre-requisites, and then try to explain all of them!
      - A good framing here is [inferential distance](#). The inferential distance of a new idea is the number of steps of new concepts someone must understand before they can understand the new idea. Eg, to teach a young kid about the quadratic formula, they first need to understand the idea of polynomials, for which they need to understand algebra - this is three inferential steps.
      - A general rule of thumb: *never* teach things with more than 2 inferential steps. An idea just learned is shaky, and doesn't yet have good connections built. It's very, very hard to anchor new ideas onto new ideas.
    - This is *really hard to get right*. Pre-requisites require you to have a good model of somebody else's mind. A useful technique is often to do a practice run on someone from your target demographic, and ask them to flag everything that confuses them.
    - Further, for groups, this can be an intractable problem, everyone has different prior knowledge. You need to have a clear picture in your head of who the talk is aimed at.
  - Students should learn actively, not passively
    - It's *really* easy for a student to just passively sit in a stream of information, taking none of it in. This achieves neither of your goals, because to form connections, compress information and connect it to their knowledge graph, they need to be putting in *some* cognitive labour.
    - Good ways to encourage this: explicitly telling them that this is important, giving exercises and time to think through them, asking the audience regular questions and giving them some time to think.
      - Question asking has the failure mode where most people won't volunteer, or just zone out a bit - I lack great solutions to this problem
    - This is *much* easier to handle in smaller settings, I'm bad at handling it in talks. The main solution I have is just to be engaging and keep the pace going well.
      - Often people will zone out, so having regular breaks, and check-points of "if you weren't following, we're changing topic so it doesn't matter" can help to rectify this
        - Even short, 30s-120s breaks can be helpful! Encourage people to get up and stretch.
        - Breaks never *feel* important, but they really, really help
  - Give examples

- Examples are an *insanely* powerful tool for teaching things well, and people rarely use them enough. I have never given a class with too many examples (and believe me, I've tried)
  - "You can teach a class with no content, only examples; you can't teach a class with only content, no examples"
- Why examples are awesome:
  - Examples are an excellent way to resolve lossy information transfer - they're a completely different channel of communication than normal. If nothing else, they serve as an error check
  - Examples are a great way to transfer tacit knowledge, without necessarily making it legible - this is what it means to build intuition
  - Examples can help fit things in to the bigger picture, they can motivate the ideas, and locate where they fit in to the student's knowledge graph
  - By giving the student a pool of motivating examples, they can often generate the ideas themselves by generalising from the examples
  - Examples can bridge the gap from "understanding the knowledge in the abstract" to "understanding where to *use* these ideas, and where they should come up"
- How to use them?
  - Often when giving a point, I'll give a micro-example to give context to it - eg "sometimes straight lines aren't enough to model data, eg with a quadratic". This should be quick and effortless, the example should make immediate sense to the students, with 0 inferential steps
    - (I can't believe it's taken me 18 posts to get to my first nested example :( )
  - Examples can be good at the start, to motivate things and show the questions we're trying to answer. It can be good to give an example, and then constantly refer back to it as we generalise the example into a concept
  - After introducing a complex idea, go through a long example and illustrate which parts of the example embody the complex idea
  - Use examples to illustrate the importance and relateability of an idea - eg if explaining how to think about good planning to students, give an example of a student with a deadline crisis that they missed - everyone relates to this
- Examples contain a *lot* of information, so the idea of information prioritisation applies strongly here - tell the students what to pay attention to in the example, and why it's interesting
- Visual information and diagrams
  - Often tacit knowledge manifests as a literal picture in my head - draw this!
  - This is another good alternate communication channel
  - Our visual memory and processing is often much better than our abilities with language - this can work well for clarified confusing and complex parts
- Pacing
  - An easy mistake that I often fall into is to give a section the amount of time it takes me to say it. I convert the ideas into words, and just read through what I've come up with.
    - This is the fallacy of viewing teaching as giving an information stream! Time should be allocated for people to process and compress information, and they need more time for hard parts and less time for easy parts

- This is *hard* to get right intuitively - when you understand an idea, it feels easy!
- A good trick: make a high-level summary, and rate each section out of 5 for difficulty. Then go and explicitly give more time to those sections - eg add more intuitions, say the key ideas more, give more examples
  - Note - pacing doesn't mean the speed at which you speak, it's about the time you give to different ideas!
  - Note that difficulty  $=/$ = importance. If one part is hard, but unimportant, cut it. Or give a brief overview, explain the important idea, and say "don't sweat the details"
- Another trick - explicitly tell students which ideas are important and worth paying attention to, and which aren't
- If doing a practice run (which you totally should), regularly check in with the test audience about pacing - **the default state of the world is that you get pacing wrong**
- It's key to get pacing right - people zone out in slow, easy sections, and get lost in fast, hard sections. Your job as the teacher is to keep as many people as possible absorbing information at the optimum rate.
- It's very hard to give accurate time estimates for things - my trick is to have a bonus section at the end which I'll cut if need be, and to pace in the moment according to my existing notes and my intuitions
- Understand the *mindset* behind a question
  - When somebody asks a question, the default response is to answer it. This is a failure to be deliberate! The student asks a question because they're confused about something, and your goal is to resolve that confusion - answering the question directly is just a means to an end.
    - This is an important distinction, because often questions are *weird*. They're confused, or don't quite make sense, or are asking about unimportant things. This manifests, for me, as the student's mind not making sense. And it's easy to get frustrated, or just to answer the dumb question directly. But this is ineffective.
      - A related effect - somebody asks a question that isn't the *real* question they want to ask. Eg, a student at a university open day who asks "how many A-Levels did you do?", when what they really care about is "how many A-Levels do I need to do to get in"
    - Your goal should be to **understand the state of mind from which that question made sense** - once you've done this, you can often resolve the confusion directly, or answer the question they really care about.
      - Do this by asking clarifying questions, trying to answer and saying "did that answer your question?", giving them several interpretations of what they're *really* asking and asking whether any resonate, paraphrasing the question back to them, etc.
      - The main trigger to look for here is a note of confusion - the question feeling a bit off, or out of nowhere, something isn't quite making sense.
      - It's a delicate balance between doing this and moving on with the talk - try to gauge whether many people share the same confusion, if not, just move on
  - Often doing this can uncover **Pedagogical Content-Knowledge**, common ways that people misunderstand the ideas you're teaching. It's

super valuable to collect these, because then you can recognise them in future and dissolve them directly.

- Illusion of transparency
  - A consequence of the typical mind fallacy - it's easy to *think* you've clearly communicated knowledge when you really, really haven't. As a consequence, you need to put a lot of effort into being grounded and calibrated - because often a confused audience won't *feel* like a confused audience to you.
  - Ask questions! Especially ones that highlight the key ideas, eg "how to do easy thing X" or "what was the key idea in here?"
  - Do hand-polls - ask people to indicate their understanding by putting their hand up high if it's clear, and low if it's less clear. This is a good technique, because most people will actually do it, unlike "does anyone have any questions?" or "is this making sense to people?"
- Seek feedback and iterate
  - Teaching is *hard*. You're fundamentally trying to convey tacit knowledge, via lossy and low bandwidth communication channels, into an alien mind that you have very limited access to. **The default state of the world is that you suck at this**
  - The solution is to regularly ask for specific, actionable feedback and calibration, and to actually put meaningful effort into updating on this! Feedback is one of the main ways you can better understand the mind of a student.

## Teaching 1-on-1

Practicing tutoring and explaining things one on one can often be more valuable! I think a great use of time for most students is to do tutoring - it's pretty fun, you get paid decently, and you get way better at explaining ideas. And the ability to explain an idea clearly in a conversation is an *amazingly* applicable skill - I use this all the time in daily life.

The main difference is that it's a lot easier to get them to be active, and it's much easier to adapt the pace and difficulty well. Essentially, invert all of the ideas in my post on how to [learn from conversation](#)

- The key technique is asking the student to paraphrase what you've said back to you
  - This forces them to be active, and to process information
  - It identifies errors, and helps you to correct them
    - Often you can then recognise th
  - It helps build a model of what's going on in their mind
  - It helps you calibrate the difficulty and pacing
  - If you're a tutor who doesn't do get the student to this, I think you're missing out on a *major* free win
    - This is also super effective when explaining ideas to friends, though can seem a bit rude
- Get the student to tell you the key points/ideas in what you've said
- Get the student to generate examples, especially typical examples
- Here, understanding the mindset behind the question is even *more important*. You should *always* do this when they ask a question, especially if they seem dissatisfied with your answer.

# Teaching Maths

- It's easy to neglect the tacit information - the intuitions, underlying concepts, motivations. This is *terrible*. One of the most important parts of teaching maths well is to convey this high-level overview
- Every proof can be heavily compressed. Most proofs have some key ideas, followed by repeatedly doing the obvious thing. "Repeatedly doing the obvious thing" will inevitably be compressed in the student's mind, so you should skip saying it at all, and *just* give the key ideas
- Examples, especially motivating examples, are *incredibly* important. It's really, really hard to learn a new concept without having a clear motivating example in mind.
  - Examples teach tacit knowledge well - they illustrate what you can and can't do, and why you care about ideas
- After learning rigour for a while, you'll end up with [post-formal](#) intuitions, where you mostly ignore rigour and think intuitively, but can drop into rigour if need be. Most of the cognitive labour in maths is reaching this point, and a good teacher will try to give as much of the post-formal intuition as possible
- Maths, especially pure maths, is often formalising an intuition. Probability is the formal study of uncertainty. Groups are the formal study of symmetries. Topology is the formal study of continuous deformations (things which don't rip or glue). Pointing this out is vital
  - A good way to find these is to notice which questions the topic newly lets you answer. This is a great way to motivate things!
- Diagrams are *awesome*
- Often you begin being able to answer a type of question with a lot of tacit knowledge, and are expected to pick all this up with examples. Often 80% of this can be captured in an explicit algorithm - this is a great way to make tacit knowledge legible.

# Teaching Applied Rationality/Life Advice

- These are *much* more about tacit knowledge than explicit, so these should be done in a workshop format with a big focus on exercises
  - Emphasise that everything is highly personal and subjective - all ideas should be adapted to your mind and your circumstances
  - Pairing people up works well for getting them to actually do the exercises!
- Ideally, boil down the tacit knowledge to a rough algorithm, and alternate explaining steps and getting the students to practice them
- The impact of the class is mostly students retaining key insights and mental habits - what the "time when I should apply this idea" feels like from the inside. *This* is the small fraction of information they'll retain. Thus it's your job to boil down the idea to these habits, say this explicitly, and structure the class to reinforce them
- Much of the impact comes from the students taking action after the class - this is hard! You want to emphasise this, and minimise barriers
  - Give time for the students to generate lists of ways to apply the ideas, and how they're relevant in *their* day to day life
  - Give time for them to set reminders for actions taken after the class
- Make it *feel* actionable - it's easy to think an idea is important, but for it to feel abstract. Eg, to think that [prioritisation](#) is a good idea, but to never get round to

it.

- Emphasise how the idea fits into everyday life and everyday problems
- Give a *lot* of examples of how to use it - this can form connections like “oh! I never thought of using it for that”
  - This conveys the tacit knowledge of *when* to use the idea!
- Emphasise relatability and importance - give examples of a bad situation where the technique wasn’t applied, and make it feel visceral and relatable

## Conclusion

If you’re planning on teaching something in the future, I hope these thoughts were useful! But even if not, I think these skills transfer excellently to explaining things in everyday life. And that thinking about teaching can make you a much more effective learner.

I find that often, as a student, I can help the teacher be more effective by asking the right questions - asking them which information is the most important, checking that my understanding is correct by paraphrasing back, asking them for the motivations and higher-level picture. The feeling of “something not fitting into my knowledge graph well” can be made into a pretty visceral one. And realising the habits of students that hinder them from learning, like being passive instead of active, and not trying to do information compression themselves, can help me recognise when I fail to do those things!

# Updates and additions to "Embedded Agency"

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

Abram Demski and Scott Garrabrant's "[Embedded Agency](#)" has been updated with quite a bit of new content from Abram. All the changes are live today, and can be found at any of these links:

- as a hand-drawn sequence ([LW link](#), [AIAF link](#));
- as blog posts ([MIRI link](#), [LW link](#), [AIAF link](#));
- and as an arXiv paper ([link](#)).

Abram says, "I'm excited about this new version because I feel like in a lot of cases, the old version gestured at an idea but didn't go far enough to really explain. The new version feels to me like it gives the real version of the problem in cases where the previous version didn't quite make it, and explains things more thoroughly."

[This diff](#) shows all the changes to the blog version. Changes include (in addition to many added or tweaked illustrations)...

Changes to "**Decision Theory**":

- "Observation counterfactuals" (discussed in the counterfactual mugging section at the end) are distinguished from "action counterfactuals" (discussed in the earlier sections). Action counterfactuals are introduced before the five-and-ten problem.
- The introduction to the five-and-ten problem is now slower and more focused (less jumping between topics), and makes the motivation clearer.
- Instead of highlighting "Perhaps the agent is trying to plan ahead, or reason about a game-theoretic situation in which its action has an intricate role to play." as reasons an agent might know its own action, the text now highlights points from "Embedded World-Models": a sufficiently smart agent with access to its own source code can always deduce its own conditional behaviors.
- $\epsilon$ -exploration and Newcomblike problems now get full sections, rather than a few sentences each.
- Added discussion of "Do humans make this kind of mistake?" (*Text versions only*.)

Changes to "**Embedded World-Models**":

- "This is fine if the world 'holds still' for us; but because the map is in the world, it may implement some function." changed to "... because the map is in the world, different maps create different worlds."
- Discussion of reflective oracles now gives more context (e.g., says what "oracle machines" are).
- Spend more time introducing the problem of logical uncertainty: emphasize that humans handle logical uncertainty fine (*text versions only*); say a bit more about how logic and probability theory differ; note that the two "may seem superficially compatible, since probability theory is an extension of Boolean logic"; and describe the Gödelian and realizability obstacles to linking the two.

Note explicitly that "the 'scale versus tree' problem also means that we don't know how ordinary empirical reasoning works" (*text versions only*).

### Changes to "**Robust Delegation**":

- Introduction + Vingean Reflection:
  - Introduction expanded to explicitly describe the AI alignment, tiling agent, and stability under self-improvement problems; draw analogies to royal succession and [lost purposes](#) in human institutions; and highlight that the difficulty lies in (a) the predecessor not fully understanding itself and its goals, and (b) the successor needing to act with some degree of autonomy. (*Text versions only*.)
  - Put more explicit focus on the case where a successor is much smarter than its predecessor. (*Text versions only*.)
  - Expanded "Usually, we think about this from the point of view of the human." to "A lot of current work on robust delegation comes from the goal of aligning AI systems with what humans want. So usually, we think about this from the point of view of the human." (*Text versions only*.)
- Goodhart's Law:
  - Fixed a typo in the text versions' Bayes estimate equation: it previously flipped the first x and y, but now shows the correct formula
$$E_{y|x}[g(x) - y] = 0. \quad (\text{Text versions only.})$$
  - Expanded discussion of regressive Goodhart, including adding more illustrations and noting two problems with Bayesian estimators (intractability, and realizability). Removed claim that Bayes estimators are "the end of the story" for regressive Goodhart.
  - Moved extremal to come after regressive instead of after causal, so extremal and regressive can readily be compared.
  - Rewrote and expanded extremal Goodhart to introduce the problem more slowly, and walk through quantilizers in much more detail.
  - Expanded discussion of causal Goodhart to clarify connection to decision theory and note realizability issues.
  - Clarified the connection to mesa-optimizers and subsystem alignment in adversarial Goodhart.
- Stable Pointers to Value:
  - Added following the Goodhart discussion: "Remember that none of these problems would come up if a system were optimizing what we wanted directly, rather than optimizing a proxy."
  - Introduced the term "treacherous turns".
  - Shortened and clarified introduction to observation-utility maximizers, described how observation-utility agents could do value learning, and removed mention of CIRL in this context.
  - Mentioned the [operator modeling problem](#).
  - Discussed wireheading as a form of Goodharting.

### Changes to "**Subsystem Alignment**":

- "Optimization daemons" / "inner optimizers" are now "mesa-optimizers", matching the terminology in "[Risks from Learned Optimization](#)". (*Change also made in "Embedded Agents" / the introduction.*)
- New section on treacherous turns, simulated deployments, and time and length limits on programs.

# Search versus design

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

*This work was supported by OAK, a monastic community in the Berkeley hills. It could not have been written without the daily love of living in this beautiful community. The work involved in writing this cannot be separated from the sitting, chanting, cooking, cleaning, crying, correcting, fundraising, listening, laughing, and teaching of the whole community.*

*This write-up benefited from feedback from David Kristofferson, Andrew Critch, Jason Crawford, Abram Demski, and Ben Pence. Mistakes and omissions are entirely the responsibility of the author.*

---

How is it that we solve engineering problems? What is the nature of the design process that humans follow when building an air conditioner or computer program? How does this differ from the search processes present in machine learning and evolution?

We study search and design as distinct approaches to engineering. We argue that establishing trust in an artifact is tied to understanding how that artifact works, and that a central difference between search and design is the comprehensibility of the artifacts produced. We present a model of design as alternating phases of construction and factorization, resulting in artifacts composed of subsystems that are paired with *helpful stories*. We connect our ideas to the factored cognition thesis of Stuhlmüller and Christiano. We also review work in machine learning interpretability, including Chris Olah's recent work on decomposing neural networks, Cynthia Rudin's work on optimal simple models, and Mike Wu's work on tree-regularized neural networks. We contrast these approaches with the joint production of artifacts and stories that we see in human design. Finally we ponder whether an AI safety research agenda could be formulated to automate design in a way that would make it competitive with search.

## Introduction

Humans have been engineering artifacts for hundreds of thousands of years. Until recently, we seem to have mostly solved engineering problems using a method I will call *design*: understanding the materials at hand and building things up incrementally. This is the approach we use today when building bridges, web apps, sand castles, pencil sharpeners, air conditioners, and so on.

But a new and very different approach to engineering has recently come online. In this new approach, which I will call *search*, we specify an objective function, and then set up an optimization process to evaluate many possible artifacts, picking the one that is ranked highest by the objective function. This approach is not the automation of design; its internal workings are actually nothing like design, and the artifacts it produces are very unlike the artifacts produced by design.

The design approach to engineering produces artifacts that we can understand through decomposition. A car, for example, is decomposable into subsystems, each of which are further decomposable into parts, and so on down a long hierarchy. This decomposition is not at all simple, and low quality design processes can produce

artifacts that are unnecessarily difficult to decompose, yet understanding how even a poorly designed car works is much easier than understanding how a biological tree works.

When we design an artifact, we seem to factorize it into *comprehensible subsystems* as we go. These subsystems are themselves artifacts resulting from a design process, and are constructed not just to be effective with respect to their intended purpose, but also to be comprehensible: that is, they are structured so as to permit a simple story to be told about them that helps us to understand how to work with them as building blocks in the construction of larger systems. I will call this an *abstraction layer*, and it consists of, so far as I can tell, an artifact together with a story about the artifact that is simultaneously *simple* and *accurate*.

Not all artifacts permit a story that is both simple and accurate. An unwieldy artifact may not be understandable except by considering the artifact in its entirety. In design, we do not produce artifacts and then come up with stories about them afterwards, but instead we seem to build artifacts and their stories simultaneously, in a way where the existence of a simple and accurate story to describe an artifact becomes a design goal in the construction of the artifact itself.

Search, on the other hand, does not operate on the basis of abstraction layers. When we use a machine learning system to search a hypothesis space for a neural network that correctly differentiates pictures of dogs from pictures of cats, the artifact we get back is not built up from abstraction layers, at least not nearly to the extent that artifacts produced by human design are. And we wouldn't expect it to be, because the search processes used in machine learning have neither the intent nor the need to use explicit abstraction layers.

The absence of abstraction layers in artifacts produced by search is no impediment to the *effectiveness* of search in finding a solution to the specified problem. But this absence of abstraction layers is an impediment to our human ability to *trust* the artifacts produced by search. When I find some computer code on stackoverflow for solving some problem, I may copy it and use it within my own program, but not before understanding how it works. Similarly, when FAA approves a new aircraft for flight, it does so not just on the basis of empirical tests of the new aircraft, but also of a description provided by the manufacturer of each of the aircraft's subsystems and how they work. For simple artifacts we may be willing to establish trust on the basis of empirical tests alone, but for sophisticated artifacts, such as an aircraft or autonomous car or artificial intelligence, we must be able to understand how it works in order to decide whether to trust it.

Design produces decomposable artifacts, for which trust can be built up by reading the stories attached to each abstraction layer. We can verify these stories by further decomposing the subsystems underneath each abstraction layer. Search does not produce decomposable artifacts and for that reason we have no way to establish trust in the artifacts produced by it, except by black-box testing, which is only appropriate for simple artifacts.

Unfortunately, we have discovered how to automate search but we have not yet discovered how to automate design. We are therefore able to scale up search by bringing to bear enormous amounts of computing power, and in this way we are solving problems using search that we are not able to solve using design. For example, it is not currently known how to use design to build a computer program that differentiates cat pictures from dog pictures, yet we do know how to solve this problem using search. We are therefore rapidly producing artifacts that are *effective but not*

*trustworthy*, and we are finding ourselves tempted by economic incentives to deploy them without establishing trust in them.

The machine learning sub-field of explainability (and it's related but distinct cousin, interpretability) is concerned with establishing trust in artifacts produced by search. There is important work in this area of relevance to the thesis presented in this paper, and we review some of it in a later section. Overall, work in this field seems to be concerned with one of the following:

- Models that are simple enough that they can be comprehended without any accompanying stories
- Manual decomposition of trained models
- Post-hoc explanations that aim to persuade but may not be accurate depictions of what's really going on.

All three of these are distinct from our view that comprehensibility comes from the pairing of artifacts with helpful stories about them.

As we begin to construct sophisticated AI systems using search, we venture into dangerous territory. We have the tools of machine learning that may soon be capable of producing AI systems that are highly effective, yet for which we have no way to establish trust. It will be a great temptation for us to deploy them without establishing trust in them, since enormous economic prizes are on offer.

To resolve this, we here propose that we should work to automate design to such a point that design can scale with computing power to the same extent as search. Under this approach we would investigate the nature of the design process that humans use to construct artifacts based on abstraction layers, and attempt to automate it. We may find it possible to automate the whole process of design, or we may automate just part of the process, leaving humans involved in other parts.

The remainder of this document is organized as follows. First we work through a simple example to make clear the distinction between search and design. Then we describe a model of design as alternating construction and factorization. Following this we argue that search can be viewed as a construction process lacking any factorization step. We then work through definitions of the terms *abstraction layer* and *comprehensible artifact* upon which much of the material in this report is predicated. We draw a connection to the factored cognition thesis, and also to the machine learning fields of explainability and interpretability. We conclude by sketching an AI safety research agenda that would aim to allow comprehensible design to scale with computing power, in order for it to become competitive with search as a process for developing advanced artificial intelligence systems. In an appendix we present notes from an informal inquiry in which we observed the author's own process of writing software over a few hours and compared what we saw to the model presented in this document.

## Example: Sorting integers

Suppose I need an algorithm that sorts lists of integers from smallest to largest, but I'm not aware of any good sorting algorithms. Suppose it's critical that the algorithm correctly sort integers in all cases, but that I've never encountered the computer science field of sorting, so I don't have any of the concepts or proofs found in computer science textbooks. I consider two approaches:

- Train a neural network to sort integers
- Write a sorting algorithm from scratch in Python

Let's consider the machine learning approach. This problem has some attractive characteristics:

- I already understand what it means for a sequence of integers to be sorted, so I can provide a perfectly consistent training signal
- I can easily generate unlimited training examples, so I immediately have access to an infinitely large dataset
- I have a precise understanding of the range of possible inputs, so although I cannot train on every possible sequence of integers, I can at least be confident that I have not missed anything in my own understanding

Now let's say that I succeed in training a neural network to sort integers, and I test it on many test cases and it works flawlessly. Am I ready to deploy this in a safety-critical scenario where a single incorrect output will lead to the death of many living beings?

Those familiar with practical machine learning will shudder at this point. A neural network is an unwieldy thing. It is composed of millions or perhaps billions of parameters. No matter how many test cases I run and verify, there are infinitely more that I didn't try, and in fact there is always a maximum length list and a maximum size integer among my test set — can I really trust that the neural network will correctly sort integers when the length of the list or size of the integer greatly exceeds this maximum tested size? It is difficult to convey just how uncomfortable I would feel in making the leap from "I tested this on this many test cases" to "I'm ready to deploy this and swear on my life that it absolutely will sort integers correctly in all cases".

If I was given the task of determining whether a neural network correctly implemented sorting, what I would actually do is the following. I would examine the operations and coefficients contained in each layer of the network and attempt to extract an understanding of what each piece was doing. I would feed examples into the network and watch how each integer was processed. I would try to watch closely enough to get some insight into how the network was functioning. I would likely end up jotting down little fragments of pseudocode as I unpacked progressively larger subnetworks, then I'd try to understand the network as a whole on the basis of the pseudocode I'd written down. Ultimately, if I did succeed at extracting an understanding of the network, I'd need to decide whether or not the network constituted a correct sorting algorithm, and I might use some formal or informal verification method to do this.

The point is that to really trust that this network will work in all cases, I would want to decompose it and understand it piece by piece.

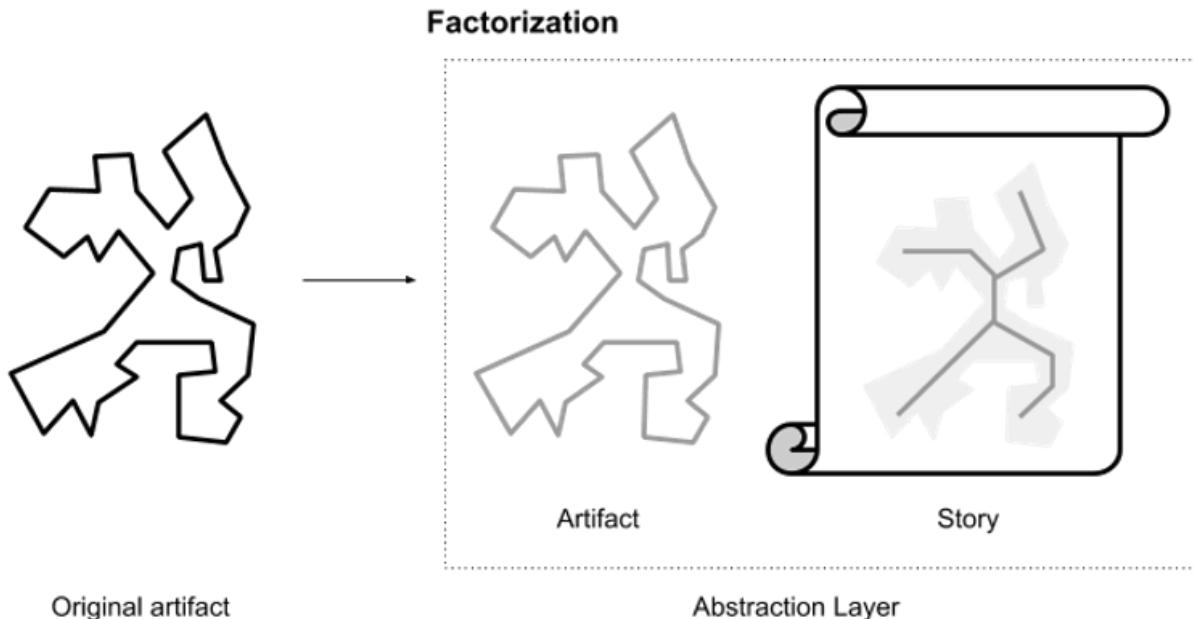
Now in this example we have taken a problem (sorting integers) that has a known direct solution, and we have compared that direct solution to learning a solution from training data. It is no surprise that the direct solution is the better choice. It is not that we should replace all instances of machine learning with direct solutions: the whole point of machine learning is that we can solve problems for which we do not have a direct solution. Instead, the point of this example is to provide an intuition for where our trust in artifacts comes from, and specifically why it is so important that artifacts are structured in a way that supports us in decomposing and understanding them.

# Design as construction and factorization

In design, we build a thing up to meet some requirements, based on an understanding of the materials we have to work with. For example, when we build a shed, we have wood for framing, wood for sheathing, concrete to anchor it into the ground, and so on. We don't need to know the details of how the two-by-four sections of wood were cut from their source material, or how they were transported, or how they were priced by the store that sold them: there is an abstraction layer upon which we can consider the two-by-fours as basic building blocks with a few known properties: that they are strong enough to support the structure of a shed; that they can be cut to arbitrary lengths; that they can be cut at an angle; that they cannot be made to bend; and so on.

Similarly, when I write some software that uses the Postgres database software to store and retrieve information, I do not need to understand the full internal workings of Postgres. I can consider the database system to be a kind of material that I am working with, and a good database system is designed such that I can understand how to use it without understanding everything about it. I know that when I execute such-and-such an SQL statement, a "row" gets added to a "table", and when I execute such-and-such an SQL statement, I get that same "row" back. But the concepts of "row" and "table" are high-level stories that we use because they are helpful, not because they give a complete description of which bits will be in which of the computer's memory locations at which time.

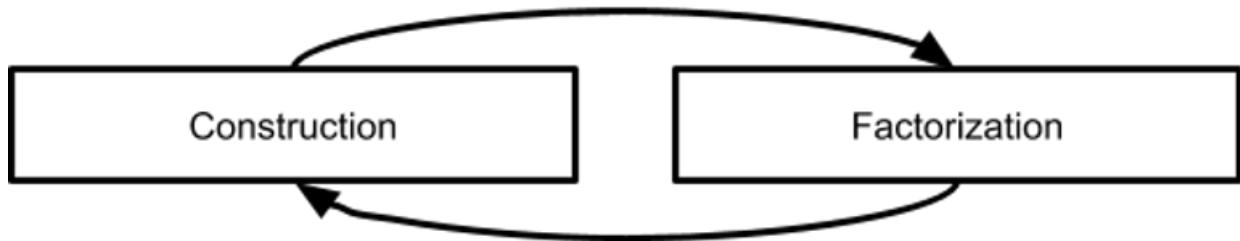
One part of design consists of using existing materials to build something. We might call this "construction". If I write a python program containing a single "main" function that pulls in a bunch of libraries and does some computation, then I'm doing construction. But very quickly I will start to factor my program out into functions so that I can more easily test it, grow it, and understand it in a way that allows me to see that it is correct. We might call this "factorization". In factorization, I'm looking for ways that I can take, for example, a relatively complicated gradient descent algorithm and say "This chunk of code finds a local minima of a convex objective. You must give it an objective function and a starting point for optimization and it will give you back the optimum of the function." In this way I can stop remembering the details of the implementation of the algorithm, which are many and can be very subtle, and instead remember only the story about the implementation. This helpful story about an artifact, together with the artifact itself, is what I call an abstraction layer.



We are familiar with looking for simple explanations of existing things from our theory of epistemics. This is different. Here we are constructing something *in a way that makes it possible to tell a simple story about it*. We are designing the artifact together with an explanation/concept/story about it all within the same process. This makes it possible to do sophisticated engineering much more quickly than if we built unwieldy artifacts and then tried to come up post-hoc stories about them, because we design our artifacts *with the goal of making them comprehensible*. We have been working for centuries to tease apart the mechanisms making up, for example, biological trees (and we are not done yet), whereas we have successfully built search engines comprising billions of lines of code in just a few years.<sup>[1]</sup>

Construction and factorization proceed in a loop. When I begin a new software project, I often work entirely within a single function for a while. Once I have a few basic pieces in place, I run them in order to test that they work and find out more about the materials I'm working with — for example, I might write code that performs a call to a remote API and then prints the server's response in order to find out some things about it that are not covered in the documentation. Once this is all working I might factor this code out into smaller functions with succinct documentation strings such as "performs such-and-such an API call, does such-and-such upon error, returns data in such-and-such a format". Then I return to construction and start putting more pieces in place, perhaps using my existing factorized code as building blocks in the construction of larger pieces.

## Basic design loop



When there is too much construction and not enough factorization, we end up with unwieldy artifacts. We forget about the details of how the artifacts work and begin to mis-use them, introducing bugs. There is a kind of proliferation, like spaghetti code, or like a house that has had electrical cables strung endlessly from place to place without ever removing older cables. It becomes impossible to make sense of what's going on. As we start counting the cables and trying to make sense of things, we get lost and forget about the cables we counted at the beginning. What is happening here is that we have a highly complex artifact without handholds, and it is very difficult to design things on top of it because in order to do so we need to somehow fit some story about the artifact into our minds, and human minds cannot work with arbitrarily unwieldy stories. Software engineers experience great pain and frustration in encountering such unwieldy artifacts, and their work tends to slow to a crawl as they give up on having any clear insight into what's going on and instead proceed by dull trial-and-error<sup>[2]</sup>.

It is also possible for there to be too much factorization and not enough construction. Sometimes in software companies there will be an attempt to pre-factorize a system before anything at all has been constructed. The way this plays out is that we come up with some elaborate set of stories for abstraction layers before we really understand the problem, and begin implementing these abstraction layers. We forget that it is the process of construction that gives us evidence about the materials we are working with, and also about the solution we are seeking. We try to do all the construction in our minds, imagining what the artifact will ultimately look like, and pre-factorize it so that we don't ever have to backtrack during design. But this over-factorization fails for fundamentally the same reason that over-construction fails: we cannot fit very much complexity into our minds, so our imagined picture of what construction will yield is inaccurate, and therefore our factorization is based on inaccurate stories.

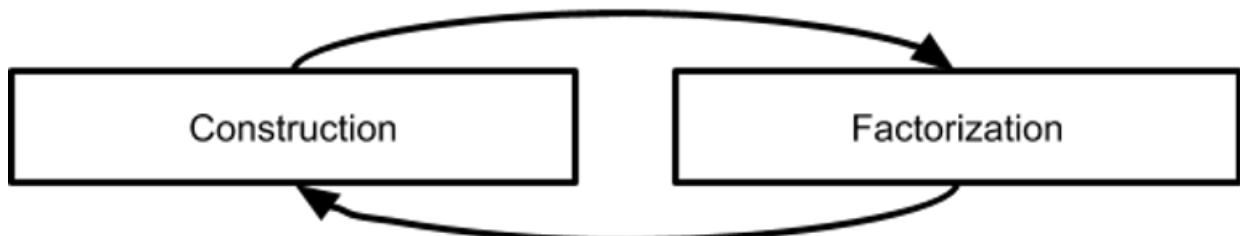
So there is a design loop between construction and factorization in which construction gives us evidence about the nature of the materials and factorization rearranges our construction into a form that permits a compact description of what it is and how it works. In a healthy design loop, construction and factorization are balanced in a way that strictly bounds the amount of complexity that we need to fit into our minds to understand any subset of the system, while maximizing the inflow of evidence about the materials we are working with.

## Search as construction without factorization

Search is pure construction, with no factorization. In search, we set up an engineering problem in a way that allows us to perform massive experimentation: trying millions of possible solutions until we find one that meets our requirements. The basic search loop consists of construction and evaluation. In machine learning, for example, evaluation

corresponds to computing the objective function and its gradient, while construction corresponds to updating our artifact by moving it a little in the direction suggested by the gradient of the objective function.

## Basic design loop



This process — construction without factorization — leads to the production of unwieldy artifacts. The reason is that in the space of possible artifacts, the vast majority are very unwieldy, so any process that doesn't explicitly optimize for comprehensibility leads by default to artifacts that are unwieldy. Imagine stepping through the set of all text files containing syntactically valid python programs. The vast majority of these contain unwieldy spaghetti code. We would have to step through many, many such text files before we found the first one that we might describe as comprehensible, or that would pass code review at a typical software company. Similarly, among the policies representable within the neural network architectures used in machine learning, there is presumably some subset that are internally well-factored in a way that would support human comprehension. But this subset is a tiny manifold within the overall hypothesis space, and even if our starting point for optimization were on this manifold, each gradient step is with high probability going to take us further away from that manifold unless we are explicitly working to stay on the manifold.

And why should a search process factorize its constructions? It has no need for factorization because it does not operate on the basis of abstraction layers. It operates on the basis of trial and error, and under trial-and-error it doesn't matter whether an artifact is comprehensible or not. This is neither a feature nor a bug of search, it is just the way things are.

But although the absence of factorization is no barrier to search, it is certainly a barrier to our comprehension of the artifacts produced by search. Without handholds in the form of abstraction layers, we have no way to understand how an artifact works, and without understanding how it works it is very difficult to establish trust in it.

## Defining comprehensible design

When we design some artifact, we want it to be both effective for its intended purpose, and comprehensible to ourselves and other humans. Being comprehensible really means that there is a story that can be told about the artifact that is both simple and accurate.

**Definition.** A *helpful* story is a story about an artifact that is both simple and accurate

What does simple mean? In this context when I say simple, I am referring to a concept that has a shape that is convenient for a human to understand. This may differ from an

abstract notion of algorithmic simplicity such as description length, because humans seem to [understand concepts through analogies to already existing concepts](#), so one may be able to quite easily understand certain complex concepts that map onto conceptual foundations already in place, such as a collection of interlinked database tables, while struggling to understand concepts without any pre-existing conceptual foundations, such as the notion of a ring in abstract algebra.

What does accurate mean? It means a story such that interacting with the artifact as though it were really as simple as described in the story does not cause harm or surprise. It means a story that is useful without being manipulative. It means a story that reveals, as far as possible, the direction of its own necessary imprecisions. This is different to pure predictive accuracy: what we care about is stories that make possible the use of the artifacts they refer to in the construction of larger systems, while trimming off details not necessary for this purpose. A story that is accurate only in the sense that it tells us how an artifact will behave may not give us any affordances for using that artifact in further construction.

Next we come to abstraction layers. We have already used the concept of an abstraction layer in discussing construction and factorization above. The definition I will use is:

**Definition.** An *abstraction layer* is an artifact together with a helpful story about it

It must be stressed once again that the terrain here is subtly different to that of epistemics, in which we observe some natural artifact and come up with a simple model to explain its behavior. In that domain we end up with a similar pairing between some artifact that is complex and a model or story about it that is simple. But in epistemics we are typically studying some already-existing phenomena, and engage in a process of hypothesizing about it. In design we get to construct the phenomena, and we can therefore shape it such that there exists a story about it that is both simple and accurate. We are allowed to ignore regions of the design space that contain effective artifacts, simply because those artifacts are difficult to understand and do not lend themselves to the construction of stories that are simultaneously simple and accurate. Finding helpful stories is a great challenge! Whenever we have a choice between finding a helpful story for an artifact produced by some black-box process, versus designing the artifact from the ground up to be amenable to a helpful story, we should certainly choose the latter!

Finally we come to a recursive definition for comprehensibility:

**Definition.** A *comprehensible artifact* is an abstraction layer that is built up from parts that are themselves comprehensible artifacts, using only a limited amount of construction to bridge the gap between the parts and the whole.

A car is a comprehensible artifact. Considered as a single artifact, a car comes with a set of very simple and very accurate stories about how to use it to drive from one place to another. Additionally, because my car was produced by an engineering process in which the design work needed to be distributed across many human engineers, it is internally structured in a way that supports decomposition. I may consider the car's engine: it, too, is well-encapsulated, meaning it has a shape that permits helpful stories to be told about it, and the owner's manual provides many such helpful stories about the car's engine. Similarly, the other parts that make up the car — chassis, wheels, transmission, and so on — each come with helpful stories that make it possible to understand them without considering all of their details. And these parts are in turn

decomposable: the engine is itself made from parts, which, due to the engineering process by which the car was designed, are again decomposable.

A tree is not a comprehensible artifact. We have been trying to map out how trees work for centuries, and we have made much progress, but we are not done. The human body is not a comprehensible artifact. Trees and human bodies both contain subsystems, presumably because natural selection is itself driven somewhat towards decomposability by the limited amount of information that can be stored in the genome, but they are not nearly so easy to understand. Of course this is no criticism of the many wonders produced by natural selection, it is just the way things are.

Consider a neural network trained to classify images as containing either dogs or cats. It is actually quite easy to turn this artifact into an abstraction layer: our helpful story is simply that you pass in an image in some appropriate format, and get back a label that tells you, with some probability, whether the image contains a dog or a cat. We can now use the neural network as a function without considering any of its internal details. But this neural network is certainly not a comprehensible artifact. We do not know how to decompose neural networks into crisp subsystems. There is some work in the field of inspectable machine learning that attempts to do this (see survey below), but this work is far from complete, and a full understanding of how image recognition works in neural networks remains elusive to us.

In the definition of comprehensible artifacts we said that only a limited amount of construction be used to bridge the gap between the parts and the whole. Each time we construct some artifact out of parts, there is some construction necessary to "glue" the parts together. In some cases we can factorize the glue itself, but there is always some glue remaining because the decomposition of parts into sub-parts has to bottom out somewhere with raw machinery that does work, or else our artifacts would be nothing more than empty abstraction layers composed endlessly. In order to *use* an artifact we generally do not need to understand this glue — that is the whole point of an abstraction layer — but in order to *understand how an artifact works*, we ultimately need to understand this glue. It is therefore critical to keep an upper bound on the amount of construction per abstraction layer, in order that we can recursively decompose our artifacts and verify their stories without ever needing to understand more than a fixed amount of glue.

## Connection to factored cognition

Andreas Stuhlmüller and Paul Christiano have proposed the factored cognition hypothesis: that the thinking processes that constitute human intelligence can be broken down into thought episodes of perhaps just a few minutes in length, with limited communication between episodes. If true, this would open the door to scaling human intelligence by scaling these short thinking episodes, such as in their iterated distillation and amplification proposal.

At the surface level, these ideas are quite distinct from those presented here, for we have not taken any stance on the nature of the computational processes in a potential solution to the comprehensible design problem, and we certainly don't assume that those processes could be factored in any particular way. We have discussed the factoring of *artifacts* in our model of comprehensible design but this is quite different from the factoring of *cognition* as Stuhlmüller and Christiano propose (just as the factoring of a car into assemblies of subsystems is distinct from the factoring of a mind that is designing a car).

Yet at a deeper level, it is often said that the structure of software products reflects the structure of the companies that produce them, and more broadly there is a way in which the internal organization of artifacts produced by minds reflects the internal organization of those minds. This may point to a deeper connection between what could be termed a "factored artifact hypothesis" advanced this document, and the factored cognition thesis of Stuhlmüller and Christiano.

## Connection to explainability and interpretability in machine learning

The field of interpretability in machine learning is concerned with techniques for making machine learning models understandable to humans. Out of four literature reviews we read on the topic, all four mentioned trust as one of the central motivators for the field. This aligns closely with our concerns. The field has grown quickly over the past few years and we do not have any intention to cover all of it within this short subsection. We refer readers to the recent and much more comprehensive series of posts here on lesswrong by Robert Kirk and Tomáš Gavenčíak ([1](#), [2](#), [3](#)), as well as to the literature reviews published by Došílović et al<sup>[3]</sup>, Arrieta et al<sup>[4]</sup>, Adadi et al<sup>[5]</sup>, and Gilpin et al<sup>[6]</sup>. We also found much value in Christopher Molnar's textbook Interpretability in Machine Learning, which is [freely available online](#).

The field is broadly concerned with tools that allow humans to develop trust in black box machine learning models by giving humans insight into the internal workings of the black-box models. As per Adadi et al<sup>[7]</sup>, approaches can be categorized along the following three axes:

- **Local vs global.** Local approaches aim to help humans to understand a single prediction made by a machine learning model, whereas global approaches aim to help humans understand the model itself. For example, a system that can explain a decision to assign a low credit rating to a particular individual is a local approach, whereas a visualization technique that shows that each layer of a convolutional neural network is building up successively larger models of visual parts is a global approach. In this report we are mostly concerned with global approaches since we want to develop sophisticated AI systems that can be trusted in general to take beneficial actions, rather than having a human review each individual action.
- **Intrinsic vs post-hoc.** Intrinsic approaches involve modifying the original learning algorithm in some way, whereas post-hoc approaches do work to improve interpretability after learning has already concluded. Imposing a sparsity prior to encourage the generation of simple models is an example of an intrinsic approach, whereas training a shallow neural network to approximate the predictions made by a deeper neural network after that deep network has already been trained is an example of a post-hoc approach. We have argued in this write-up that intrinsic approaches are more attractive for sophisticated AI systems, due to the opportunity to construct models in a way that permits interpretability. We would be excited to discover post-hoc approaches that work for sophisticated and general AI systems, but we expect that route to be much more challenging.
- **Model-agnostic vs model-specific.** Model-agnostic approaches can be used with any type of machine learning model (neural networks, decision trees, linear or logistic regressors, kernel machines, etc), whereas model-specific approaches

are predicated on some particular type of model. Model-agnostic approaches tend to treat the underlying model as a black box and therefore tend to fall into the post-hoc dimension from the preceding axis. For advanced AI systems, we believe intrinsic approaches are more promising, so on this axis we expect energy to be correspondingly focused on model-specific approaches.

A classic interpretability method is that of Shapely explanations<sup>[8]</sup>, which assigns, for a model that makes a prediction  $y$  based on a set of features  $x_1, \dots, x_n$ , a "contribution" to each feature  $y_1, \dots, y_n$  such that the sum of the individual feature contributions is equal to the original prediction. For linear models, computing such contributions is straightforward: we just take the coefficients of the model as the per-feature contributions. But for nonlinear models it is not so clear how best to assign such contributions. The authors show that there is only one way to assign such contributions if one wishes to adhere to certain accuracy and consistency desiderata. This kind of method seems like a useful tool for checking simple predictive models, but is not going to provide deep insights into models that we suspect implement sophisticated algorithms internally.

In the remainder of this section we review three papers that we examined in detail. We chose these three papers based on citations from the review papers above, and based on recommendations from friends.

## **Rudin, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead* (2019)<sup>[9]</sup>**

Cynthia Rudin has argued persuasively that it is a mistake for us to construct black-box models and then construct further explanation-producing models that are optimized merely to persuade humans, rather than to be true accounts of how the original model actually works. We very much agree! Instead, she argues that we should train models that have a structure that allows us to understand them directly. She identifies the former (post-hoc explanation production) with "explainable machine learning" and the latter (intrinsically comprehensible models) with "interpretable machine learning". This differs terminologically from several review papers we have read, most of which take these terms to refer to approximately the same overall body of research, but we find it helpful none-the-less.

Rudin offers the provocative hypothesis that most black-box models can be replaced by equally or near-equally accurate interpretable models, and that there is in fact no trade-off between interpretability and accuracy. This is a bold and important claim, and she formalizes the case for it. The gist of her argument is that, due to the finite size of the dataset on which any model is trained, there will generally be many models within a small performance margin of the globally optimal model, and that among these one is likely to find an interpretable model. Put another way: no finite dataset contains enough information to pick out an infinitely precise region within model space, so the question becomes whether the set of close-to-optimal models is large enough to contain an interpretable model. A series of set-approximation theorems in *A study in Rashomon curves and volumes*<sup>[10]</sup> argues that one is indeed likely to find an interpretable model within this set.

Rudin's lab at Duke University is all about training intrinsically interpretable models. Of particular interest are rule lists, which are sequences of logical conditions of the following form:

```

IF      age between 18-20 and sex is male      THEN predict arrest (within 2
years)
ELSE IF age between 21-23 and 2-3 prior offenses THEN predict arrest
ELSE IF more than three priors      THEN predict arrest
ELSE      predict no arrest.

```

A second model category of interest is scoring systems, in which an integer score is computed by summing integer "points" associated with binary conditions, and the final model output is determined by a lookup table indexed by scores:

1. Prior Arrests $\geq 2$	1 point	...
2. Prior Arrests $\geq 5$	1 point	+ ...
3. Prior Arrests for Local Ordinance	1 point	+ ...
4. Age at Release between 18 to 24	1 point	+ ...
5. Age at Release $\geq 40$	-1 points	+ ...
	<b>SCORE</b>	= ...

SCORE	-1	0	1	2	3	4
RISK	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

It was news to us that there are algorithms for the global optimization of such models! These algorithms differ substantially from the heuristic methods best-known in data science. In our experience it is not usually considered important to construct such logical models optimally, since one is usually working within some boosting or bagging framework in which many models of the same type are combined to produce a final output and any deficiency in one particular model is compensated for in the ensuing gradient steps. But this leads to a proliferation of models, which in turn lends to an uninterpretable overall model. Rudin's approach is to instead focus on achieving optimality in the construction of a single small model, paying a price in terms of implementation difficulty, but reaping rewards in terms of models that are simultaneously accurate and interpretable.

This highlights the real trade-off in machine learning, which is not between interpretability and accuracy but between interpretability and ease of implementation. On the one hand we can work with hypothesis classes such as neural networks that have optimization algorithms that are easy [11] to understand and implement but produce models that are difficult to interpret. On the other hand we can work with hypothesis classes such as sparse logical models, which are difficult to optimize but easy to interpret. Yet this difficulty of optimization is a one-time cost to be paid in algorithms research and software engineering. Once effective algorithms have been discovered and implemented we can use them as many times as we want. Furthermore, it may take less computing power to find an optimal logical model than to train a neural network, and it will almost certainly require less expertise to use such algorithms since local optimizers are sensitive to all kinds of initialization and stepping

details, whereas global optimizers either find the true global optimum or fail to do so in a reasonable amount of time.

It is exciting to consider what it would look like to use optimal simple models in computer vision or reinforcement learning. An enticing array of difficult optimization problems beckon, with the prize for their solution being the ability to construct simultaneously interpretable and accurate models.

Rudin speculates on such advances herself, in particular with prototype networks in computer vision. These are neural networks in which early layers are ordinary convolution feature-extraction layers, and then later layers reason by explicitly finding correspondences between regions in the input image and similar regions in training images. The final prediction is then made on the basis that if there are many correspondences between an input image and training images labelled "bird", then the input image itself probably contains a bird. This makes the model somewhat interpretable because one can visually inspect the correspondences and understand how the prediction was made on that basis. However, the early convolutional layers of the network remain opaque.

Rudin's work is exciting and insightful, but we believe she misses the following two points.

First, Rudin's work focuses on building models that are interpretable by being so simple that we can understand how they work just by looking at them. This is like being handed a tool such as a screwdriver or hammer that is so simple that we understand it immediately without needing to refer to any instruction manual. But other tools — say, a 3D printer — may be difficult to understand just by looking at them, yet easy to understand with the help of an instruction manual. We should be willing to produce complex models if they are shaped in such a way that a simple and accurate instruction manual can be written about them, and if we have methods for producing such instruction manuals. Post-hoc explanations are not good enough here; what we are suggesting is to build models and instruction manuals together in such a way that (1) the instruction manual *accurately* describes how the model really works, and (2) the instruction manual makes it *easy* for a human to understand the model. Achieving both of these aims simultaneously will require the model itself to be constrained in its complexity since not all models (presumably) permit any such instruction manual to be written about them, but it should impose *less* of a constraint than requiring our models to be interpretable on-sight without any instruction manual.

Second, Rudin's work focuses on simplicity as a proxy for interpretability. But algorithmic simplicity is only a weak proxy for how readily a model can be understood by a human. There are concepts that are quite complex when measured by any abstract measure of complexity (such as description length) that humans will nevertheless reason about quite intuitively, such as complex social situations. On the other hand there are concepts from, for example, abstract algebra, that are simple according to abstract measures of complexity, yet require significant training to understand. For this reason we tend to explain sophisticated concepts by analogy to the concepts most intuitive to us. A better measure of a model's interpretability to a particular person would be the length of the shortest description of that model *in terms of concepts this person has already acquired*, where the already-acquired concepts are treated as primitives and do not contribute to the length of a description.

One thread of research within interpretability of great interest and relevance to the present work is Chris Olah's investigation of circuits in convolutional neural networks trained to perform image classification. This thread of work initially gathered attention with Olah's [2017 article](#) on visualizing individual neurons as well as whole features (one channel within a layer) by optimizing images and image patches to maximally activate these neurons and features<sup>[13]</sup>. This produced beautiful dream-like images that gave some insight into what the network was "looking" for within different layers and channels.



Olah now proposes to go beyond mere visualization and initiate what he foresees as a "natural science of interpretability" — studying the structure of trained neural networks in the way we might study the inner workings of plants or animals in biology. In [Olah's words](#):

Most work on interpretability aims to give simple explanations of an entire neural network's behavior. But what if we instead take an approach inspired by neuroscience or cellular biology — an approach of zooming in? What if we treated individual neurons, even individual weights, as being worthy of serious investigation? What if we were willing to spend thousands of hours tracing through every neuron and its connections? What kind of picture of neural networks would emerge?

In contrast to the typical picture of neural networks as a black box, we've been surprised how approachable the network is on this scale. Not only do neurons seem understandable (even ones that initially seemed inscrutable), but the "circuits" of connections between them seem to be meaningful algorithms corresponding to facts about the world. You can watch a circle detector be assembled from curves. You can see a dog head be assembled from eyes, snout, fur and tongue. You can observe how a car is composed from wheels and windows. You can even find circuits implementing simple logic: cases where the network implements AND, OR or XOR over high-level visual features.

While we find this work to decompose trained neural networks very exciting, it is worth noting just how immense this research program is. After years of work it has become possible to *begin* to identify neural network substructures that implement elementary logical operations. Decomposing entire networks is still a long way off. And even at that point we are still in the realm of feed-forward networks with no memory, trained to solve conceptually straightforward supervised learning tasks in which the input is a single image and the output is a single label. If we build sophisticated artificial intelligence systems by training neural networks, we can expect that they will contain logic many levels deeper than this.

We very much hope this work proceeds quickly and successfully, but it's apparent difficulty does lend credence to the thesis advanced in this document that post-hoc

interpretation of artifacts not optimized for comprehensibility is difficult, and that we should find ways to design artifacts on the basis of abstraction layers, not because they would perform better, but because we would be able to understand and therefore trust (or distrust) them.

## **Wu et al., Beyond Sparsity: Tree Regularization of Deep Models for Interpretability**[\[14\]](#)

In this paper, Wu et al. train neural networks on medical diagnosis tasks, in such a way that the neural networks are well-approximated by decision trees. The idea is that humans can interpret the decision trees, thereby gaining some insight into what the neural network is doing. The contribution of the paper is a regularizer that can be used to train neural networks that have this property.

At a high level this idea was exciting to us. We hoped for a demonstration that neural networks can be trained with regularizers that cause them to take on a form that is well-approximated by a "simple story", if one is willing to take the decision tree as a story.

But on closer inspection the paper is disappointing. The regularization does not cause the network to have an internal *structure* that approximates a decision tree, it merely causes the *outputs* of the network to be well-approximated by a decision tree. The decision tree therefore gives no real insight into *how the network works*, which is the kind of understanding we should be demanding before trusting a statistical model. Furthermore, the authors report that on a binary prediction task the neural network "has predictions that agree with its corresponding decision tree proxy in about 85-90% of test examples". That means that in 10-15% of training examples, and presumably at least 10-15% of the time in the real world, the neural network produces the opposite answer from the decision tree. One might wish to understand why these cases differ in this way, but these are precisely the cases where the decision tree offers no help.

## **Conclusion: AI safety via automated comprehensible design?**

As we consider how to navigate towards a safe and beneficial future of artificial intelligence, we face the following dilemma. On one hand we have search which is moving forward quickly but produces untrustworthy artifacts. On the other hand we have design, which under good conditions can produce trustworthy artifacts, but is moving forward very slowly in the domain of artificial intelligence. Search is automated and is therefore accelerating as more and more compute power becomes available to it; design remains stuck at the fixed pace of human cognition.

In the AI safety community there are two basic views on how to resolve this dilemma. On one hand there are those who seek to rescue search: to find ways to harness the power of modern machine learning in a way that produces trustworthy artifacts. Paul Christiano's work on iterated distillation and amplification seeks to resolve the dilemma by using imitation learning to gradually amplify human capabilities. CHAI's work on assistance games seeks to resolve the dilemma by relaxing the requirement that an objective be specified before optimization begins. Geoffrey Irving's work on debate

seeks to resolve the dilemma by having powerful AI systems scrutinize one another's claims in front of a human judge. We applaud these efforts and wish them success.

On the other hand there are those who say that our best bet is to throw as much human thought as possible at the design approach; that despite the slow progress of manual design to-date it is critical that we acquire a foundational theory of intelligent agency and that we should therefore pursue such a theory with whatever resources we have. This is the perspective of researchers at the Machine Intelligence Research Institute (as we understand it). We applaud these efforts, too, and very much hope for research breakthroughs on this front.

But there is a third option: we could automate design, making it competitive with search in terms of its effectiveness at producing powerful artificial intelligence systems, yet retaining its ability to produce comprehensible artifacts in which we can establish trust based on theories and abstraction layers.

We do not currently know how to automate design. We do not really know what design is, although we hope the ideas presented in this essay are helpful. This is therefore a call for research into the nature of comprehensible design and its possible automation.

To automate design, our fundamental task is to build computer systems capable of producing artifacts together with helpful stories about them. To do this, we will need to understand what makes a story helpful to humans (we have proposed simplicity and accuracy, but this surely just scratches the surface), and we will need to discover how computer systems can produce such stories.

We will need to discover how to deeply integrate story production with artifact production. If we try for post-hoc story production — constructing the artifact first and then fitting a story to it after-the-fact — then we will be solving a much more difficult problem than we need to. We should shape our artifacts so as to make story-writing as easy as possible.

Between construction and factorization, it is factorization that seems at present most mysterious. How might we automate the "carving at the joints" of a messy bit of construction? How might we do this such that an elegant abstraction layer is produced?

To do this, we could start with existing search algorithms (*i.e.* machine learning) and modify them so that they produce stories together with artifacts. Perhaps such an approach would fit within the existing field of interpretability in machine learning.

Or we could start with existing design processes, carefully examining humans engaged in engineering and attempt to automate some or all of their labor.

Or we could start somewhere else entirely, perhaps with some stroke of genius that points the way towards a different and more trustworthy approach. Let us hope that such a stroke of genius is forthcoming soon.

## **Appendix: An informal practical investigation**

I undertook a small investigation into the nature of engineering by working on a personal software project myself, while noting how I was navigating and problem-solving. I set a timer at 10 minute intervals, at which point I would stop and write down what I was working on, how I knew to work on that, and how I was going about solving the problem.

The project I worked on was a small script in the Go language to manage the creation of Google Cloud Platform (henceforth "GCP") projects, as well as to enable and disable APIs. In this section I will refer to the specifics of the various technologies I was using to solve this problem because I find it helpful to be very concrete when performing investigations such as this. However, if you are unfamiliar with these particular technologies then please know that their specifics are not really central to this section.

The investigation took place over the course of about 5 hours total over 3 days. I made no attempt to formally test any particular hypothesis, although I was interested in whether my experience matched the framework presented in this essay.

The script I wrote was intended to allow one to configure a GCP project by writing a single configuration file specifying the project's name, ID, billing account, and a list of APIs to be enabled on the project. The script would then make the necessary API calls to GCP to create or update the project as necessary. This was something that I've wanted to build for a while because every time I set up a new GCP project I find that I've forgotten these finicky project creation steps and have to once again read through the documentation and discover how to do it again. I therefore was excited to build a small tool to automate this.

I began by defining the structure of the configuration file and parsing it in YAML format, then I wrote API calls to create the project if it didn't already exist, then I wrote API calls to enable and disable APIs based on the configuration file, then I wrote API calls to link a billing account to the project, then finally I wrote a helper command to invoke the standard "gcloud" tool with the relevant project automatically selected.

My findings were as follows.

I was astounded by the wealth and depth of the concepts needed to build this simple tool. In the first few minutes of working I wrote some code to parse some simple command-line arguments to the script. Already here I was operating on the basis of powerful stories about how a computer program is invoked from the command line, what the command line is, and how one typically passes in options on the command line. I was using a particular command-line processing library that is based on defining a struct in Go and tagging the various fields with information about how they are to be mapped to strings passed on the command line. I was already familiar with this library so I could work through this part very quickly. I could just "see" the obvious "right way" to solve the sub-problem of passing in command line arguments.

It did not feel as though my cognition consisted of any kind of "search" over a hypothesis space (although of course I didn't have full access to all the things happening beneath the subconscious level). It felt more like I was rolling out a recipe that I was already familiar with.

Later, as I was constructing the API calls to create projects and enable and disable APIs, there were an even larger number of concepts in play: concepts about what a project is, what a REST API is, how errors are typically returned from REST APIs, how the GCP client libraries typically wrap these APIs, what a context is in Go, to say nothing of the concepts involved in formulating the Go code itself — which involves understanding functions, variables, structures, if statements, for statements, and so on. And these are really only the low-level concepts that one is concerned with during implementation. There are also the high level concepts of declarative infrastructure and software tooling in general that was very much guiding my approach to solving the problem.

When we write such software, we are building on top of a huge mountain of sophisticated materials (languages, remote APIs, cloud services, and so on). The only way we can make sense of these enormously complex materials is via abstraction layers and the concepts and stories they are predicated on. Navigating this landscape without concepts would be utterly impossible.

At one point I encountered an API call that was returning a 403 Forbidden error code. I initially believed that this was due to an authentication problem, since there is a general concept in REST APIs about which error codes should be used to indicate which kinds of problems, and I expected that the GCP APIs would follow this standard concept. I spent some time trying to debug this by changing the way I was doing authentication, but then I later realized that this error code is returned when one attempts to get information about a project that doesn't exist yet. This is an interesting example of what happens when one uses a concept that does not match the reality of the materials being described. It's also interesting that I was the one that created this pairing between material (the API call itself) and concept (the standard conventions for HTTP error codes) in my head. It was not that I read the GCP documentation and it turned out to be incorrect. In fact I did not read the GCP documentation on this particular matter, I simply assumed that the API would conform to standard conventions. I therefore created a kind of abstraction layer of my own over the top of the lower-level abstraction layer provided by GCP. This abstraction layer that I created did not consist of me writing any code but just of me taking a concept I was familiar with (conventions for HTTP error codes) and pairing it with the artifact of these particular GCP APIs.

The way I both discovered and resolved this mismatch between concept and artifact was through experimentation. After I had written the code to call this API, I put in place an "end cap" that allowed me to run my script even though it was far from finished. I did this by writing code to print the data returned from the API call and then exiting. In this way I could look directly at the materials I was working with by inspecting the actual data returned from the server, and see any ways in which the raw materials differed from the concepts I was using to formulate my expectations. I discovered the problem this way, and then later resolved the problem this way also, by trying various IDs of projects that did and did not exist, and noticing that the 403 Forbidden error code was returned for non-existent projects. This is a good example of the way that construction yields "evidence" about the nature of the materials we are working with. While engaging in design it is critical to establish a channel for regularly receiving this evidence.

Most of my work on this project consisted of construction. The APIs I was working with came with client libraries that made the code to perform the API call fairly succinct, and I found no reason to add further abstraction layers. Furthermore I perceived some risk of running into various show-stoppers that would force me to abandon the project entirely (such as there not being any published API to do the things I needed to do in order to have the script operate the way I wanted it to), so I was eager to see the whole thing work end-to-end before engaging in a lot of factorization.

There were places where I did engage in factorization though. One was in writing a poll loop to check on the long-running operation of creating a project. This involved a loop that would repeatedly call the API that checks on the status of a long-running operation on GCP, and return either success or an error message in the case of failure. This piece needed logic so that it would give up after a timeout expired to prevent the possibility of the script executing in a loop indefinitely. The code to perform this poll loop was sufficiently complicated, and there was a sufficiently elegant abstraction layer that could be wrapped around it, that I quickly factored this code out into its own function.

Overall, I was struck by how little this process resembled any kind of optimization process, at any level. Perhaps there was a sophisticated and general optimization process happening at a subconscious level in my brain, but it didn't appear that way. It appeared that I was rolling out a set of known recipes; that I basically knew what to do at most points and my time was occupied with looking up documentation and trying API calls to discover the specifics of how to do each piece.

Perhaps this was because I was working on a from-scratch project starting with a blank slate, for which I already had a clear goal. Perhaps it was the planning stage, which happened prior to the time during which I was recording my activities at 10 minute intervals, where optimization took place. I did not actually do any formal planning but I had encountered the need for this script several times over a time period of year or so, and each time this happened I further crystallized a rough plan for the script I was implementing here.

Or perhaps it was because the project involved lots of API calls with very little algorithmic complexity. It would be interesting to repeat the experiment while solving some algorithmic programming puzzles.

I was also struck by my ability to work towards a goal that was only vaguely specified. Although I did, as mentioned above, have a sense of what the finished product should look like, this sense was very rough. During construction phases I regularly felt refinements to this rough overall picture snap into place, as I grappled directly with the materials at hand. For example, I decided on the exact structure of the configuration file while writing the code to parse the configuration file, and I made decisions about what would happen if no billing account was specified while writing the API calls to link billing accounts. In this way it felt like my overall plan was itself a kind of high-level concept, and I incrementally refined it to lower-level conceptual clarity as I built up each piece of the puzzle.

In this way, the construction phases brought in evidence not just about the nature of the problem (the nature of the materials I was working with, etc), but also about the nature of the goal. It seemed that I was collecting *evidence about the objective* as I proceeded. This is very different from a pure search approach in which the objective is specified upfront.

---

1. I do not mean to make any claim about the relative complexity of trees versus modern operating systems. I suspect trees are profoundly more sophisticated. [←](#)
2. We might say that when software engineering becomes search rather than design, it becomes dull and painful. [←](#)
3. Došilović, F.K., Brčić, M. and Hlupić, N., 2018, May. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210-0215). IEEE. [←](#)
4. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, pp.82-115. [←](#)
5. Adadi, A. and Berrada, M., 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, pp.52138-52160. [←](#)

6. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M. and Kagal, L., 2018, October. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80-89). IEEE. [←](#)
7. Adadi, A. and Berrada, M., 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, pp.52138-52160. [←](#)
8. Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765-4774). [←](#)
9. Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), pp.206-215. [←](#)
10. Semenova, L. and Rudin, C., 2019. A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*. [←](#)
11. In the sense that it is conceptually easy to understand and implement gradient descent, not that it is computationally easy, or easy to find a global optimum [←](#)
12. Olah, et al., "Zoom In: An Introduction to Circuits", Distill, 2020. [←](#)
13. Olah, et al., "Feature Visualization", Distill, 2017. [←](#)
14. Wu, M., Hughes, M.C., Parbhoo, S., Zazzi, M., Roth, V. and Doshi-Velez, F., 2018, April. Beyond sparsity: Tree regularization of deep models for interpretability. In *Thirty-Second AAAI Conference on Artificial Intelligence*. [←](#)

# **Sunday, 20/8/30, 12pm PDT - Tagging Celebration: Habryka/Crawford + Party**

## **FB Event**

Woop! Woop! This past Saturday, gallant taggers brought our immense first tagging campaign to a close. The archives have been swept and every post with over 25 karma has been given at least one tag. To date, 15,440 tags have been applied to 8,022 posts.

[See the full celebration post here.](#)

While this is [only the beginning for tagging](#), what a beginning it is! We think this deserves some celebration. We're preparing a two-part event:

# **Part 1: Oliver Habryka and Jason Crawford discuss Intellectual Progress**

<https://us02web.zoom.us/j/82547847213>

Back in 2017, when the future of LessWrong was in doubt, [Habryka](#) led the revival efforts, pulling together a team and creating LessWrong 2.0. Central to Habryka's vision of LessWrong was that it would accelerate intellectual progress on the important problems facing humanity. *Intellectual progress* here simply meaning progress on building knowledge and understanding.

For several years, [Jason](#) has been seeking to answer the question of which factors have led to [human progress](#) in general: how did we go from living at the mercy of nature with merely stone tools and fire to buildings, electricity, medicine, legal systems, etc? Jason's research, and info about the broader [Progress Studies](#) movement, can be found at his blog, [Roots of Progress](#) (with many pieces crossposted to LessWrong).

Of course, the history of human progress, in general, is tightly woven with humanity's intellectual progress. When our understanding of the world increased, so did our ability to shape it (for better or worse).

To celebrate the new tagging system ([itself designed with the goal of intellectual progress](#)) we've decided to have Habryka and Jason chat about questions such as:

- What historical factors were important for intellectual progress?
- What conditions are most important to create now in order to get intellectual progress?

**Starting at 12:00 PDT, Habryka and Jason will chat on Zoom for ~1 hour, including some Q&A**

## **Part 2: Party in Rational Woods (our Topia location)**

<https://topia.io/rational-woods> (feel free to check it out now)

Following the talk, we will migrate to a social environment that more easily allows people to strike up small group conversations. Raemon has created the marvelous Rational Woods on the Topia platform. You can check it out now for social hangouts. There are some neat features, so make sure to click on things.

This section will begin following the main talk and will last for several hours, so long as people are having a good time. Conceivably, we could play Breakfast is Served or Person Do Thing with LessWrong tags if there's enough enthusiasm. Not sure if that'll work well or not.

Also, I hope everyone can treat our [top taggers](#) as VIPs at the party. They deserve it. Three cheers for them!

## **The Details**

**When: Sunday, 30th August, 12:00PM PDT / 3:00PM EDT / 7:00PM UTC**

**Where:**

- Zoom: <https://us02web.zoom.us/j/82547847213>
- Topia: <https://topia.io/rational-woods>

**FB Event:** <https://www.facebook.com/events/2827544884191033/>

# **nostalgebraist: Recursive Goodhart's Law**

This is a linkpost for <https://nostalgebraist.tumblr.com/post/186105132274/theres-this-funny-thing-about-goodharts-law>

Would kind of like to excerpt the whole post, but that feels impolite, so I'll just quote the first four paragraphs and then suggest reading the whole thing:

There's this funny thing about Goodhart's Law, where it's easy to say "being affected by Goodhart's Law is bad" and "it's better to behave in ways that aren't as subject to Goodhart's Law," but it can be very *hard* to explain why these things are true to someone who doesn't already agree.

Why? Because any such explanation is going to involve some step where you say, "see, if you do that, *the results are worse*." But this requires some standard by which we can judge results ... and any such standard, when examined closely enough, has Goodhart problems of its own.

There are times when you can't convince someone without a formal example or something that amounts to one, something where you can say "see, Alice's cautious heuristic strategy wins her \$X while Bob's strategy of computing the global optimum under his world model only wins him the smaller \$Y, which is objectively worse!"

But if you've gotten to this point, you've conceded that there's *some* function whose global optimum is the one true target. It's hard to talk about Goodhart at all without something like this in the background - how can "the metric fails to capture the true target" be the problem unless there is some true target?

# Infinite Data/Compute Arguments in Alignment

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

*This is a reference post. It explains a fairly standard class of arguments, and is intended to be the opposite of novel; I just want a standard explanation to link to when invoking these arguments.*

When planning or problem-solving, we focus on the hard subproblems. If I'm planning a road trip from New York City to Los Angeles, I'm mostly going to worry about which roads are fastest or prettiest, not about finding gas stations. Gas stations are abundant, so that subproblem is easy and I don't worry about it until harder parts of the plan are worked out. On the other hand, if I were driving an electric car, then the locations of charging stations would be much more central to my trip-planning. In general, the hard subproblems have the most influence on the high-level shape of our solution, because solving them eats up the most degrees of freedom.

In the context of AI alignment, which subproblems are hard and which are easy?

Here's one class of arguments: compute capacity and data capacity are both growing rapidly over time, so it makes sense to treat those as "cheap" - i.e. anything which can be solved by throwing more compute/data at it is easy. The hard subproblems, then, are those which are still hard even with arbitrarily large amounts of compute and data.

In particular, with arbitrary compute and data, we basically know how to get best-possible predictive power on a given data set: Bayesian updates on low-level physics models or, more generally, approximations of Solomonoff induction. So we'll also assume predictive power is "cheap" - i.e. anything which can be solved by more predictive power is easy.

This is also reasonable in machine learning practice - once a problem is reduced to predictive power on some dataset, we can throw algorithms at it until it's solved. The hard part - as many data scientists will attest - is reducing our real objective to a prediction problem and collecting the necessary data. It's rare to find a client with a problem where all we need is predictive power and the necessary data is just sitting there.

(We could also view this as an [interface argument](#): "predictive problems" are a standard interface, with libraries, tools, algorithms, theory and specialists all set up to handle them. As in many other areas, setting up our actual problem to fit that interface *while still consistently doing what we want* is the hard/expensive part.)

The upshot of all this: in order to identify alignment subproblems which are likely to be hard, it's useful to ask what would go wrong if the world-modelling parts of our system just do Bayesian updates on low-level physics models or use approximations of Solomonoff induction. We don't ask this because we actually expect to use such algorithms, but rather because we expect that the failure modes which still appear under such assumptions are the hard failure modes.

# Does crime explain the exceptional US incarceration rate?

Open Philanthropy motivates their criminal justice grants with sentences like this: "The United States incarcerates its residents at a higher rate than any other major country. "

This sounds bad, but conservatives argue that this is because of our uniquely high crime rate. I didn't see any straightforward research looking at incarceration and crime rates across countries, so I did some simple analysis below.

One-line summary: the US's incarceration rate is still surprisingly high accounting for its homicide rate.

**Update on 8/16/2020: Crossposting from [here](#):**

*We don't have an incarceration problem—we have a crime problem...The critics of 'mass incarceration' love to compare American incarceration rates unfavorably with European ones. Crime is inevitably left out of the analysis.*

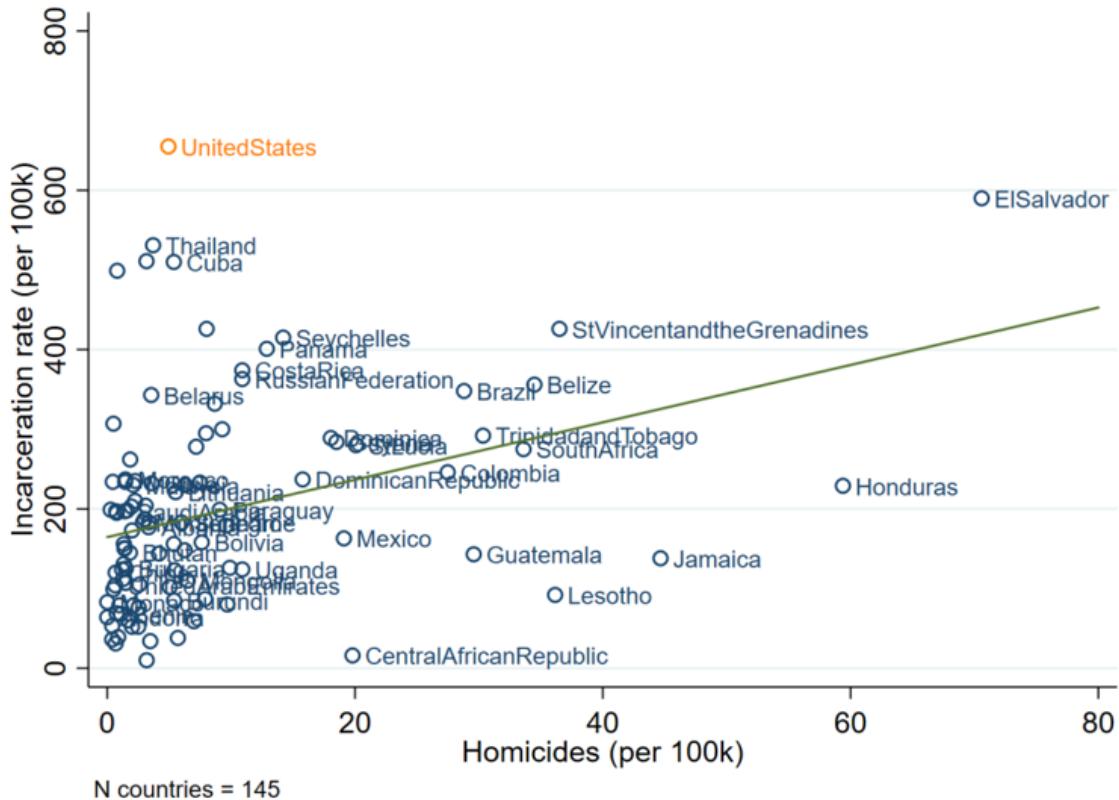
That's from [Heather MacDonald](#).

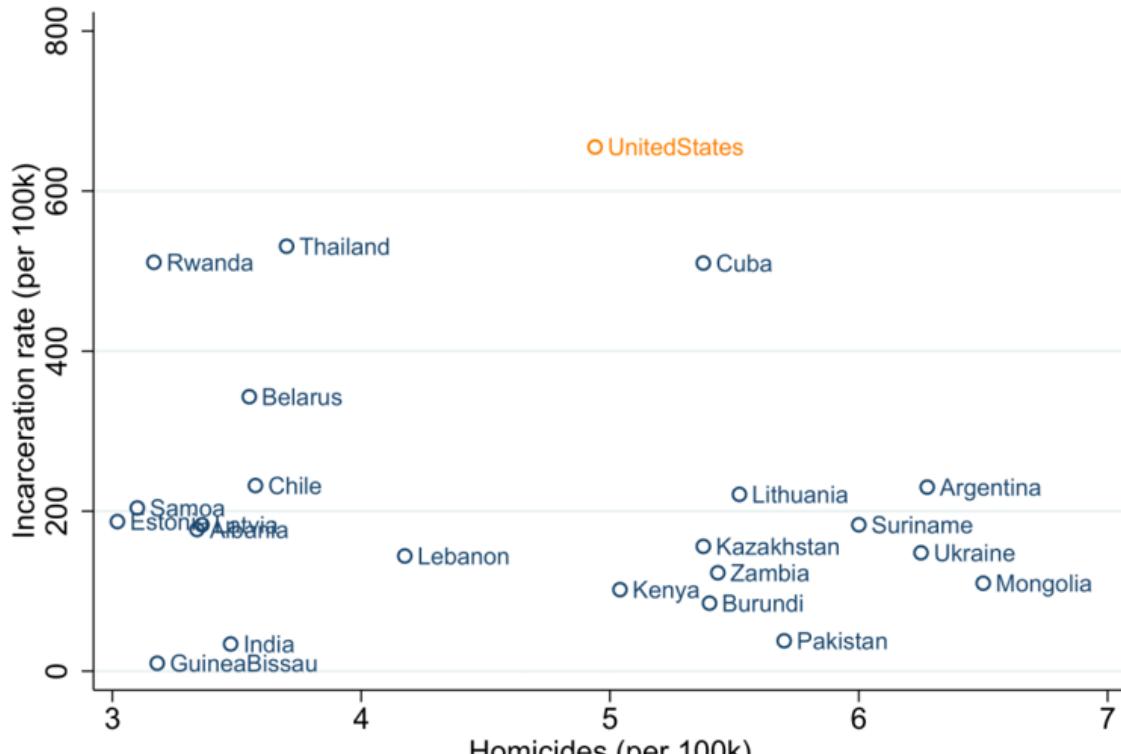
It's true that the US incarceration rate is rarely presented next to crime rates. For instance, a famous statistic is that the US has around [20 percent of the world's prisoners](#), but we typically don't hear people follow that with our crime victimization numbers.

So to what extent is this all a crime problem? A simple way to test this idea is to see, at the country level, how our incarceration rate scales with our crime rate compared to other countries.

I couldn't find these plots from some basic googling so I tried doing it myself. The incarceration data is from [Wikipedia](#) and the homicide data is from the [World Bank](#), in both cases using the most recent data available. I believe data tends to be better for homicide rates; this is meant to be a proxy for crime broadly.

Here's the incarceration rate against the homicide rate for all countries that were in both datasets:

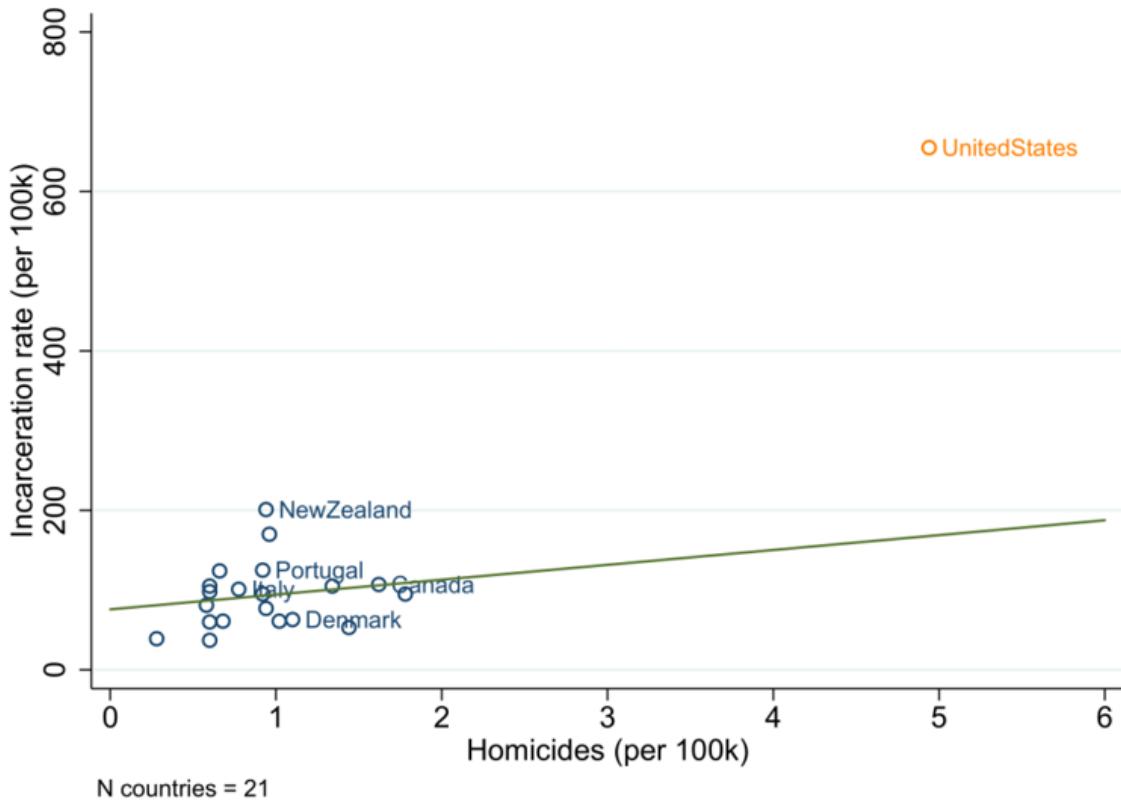




N countries = 23

Incarceration rate vs. homicide rate for countries similar to the US in homicide rate.

What about compared to the US's "peer" countries? I went through and haphazardly coded a list of richer nations and restricted to those. The US is quite an outlier in both ways. Our homicide and incarceration rates are around 5 times higher than that clump at the bottom. The trend line seems basically meaningless here, but we're way above it.



So: MacDonald’s argument supposes that the appropriate level of incarceration depends on the level of crime. But that first cross-country picture suggests that we can’t justify the US’s current incarceration rates based on how they generally scale with homicide rates in other countries.

Does this mean that our incarceration problem is not a crime problem? One response is that we should just ignore the trendline altogether because the relationship I found was too weak to be useful—other factors are more important. But then it seems it’s on the incarceration defenders to point to the crime measures that do matter and make the US seem normal.

Related to this is that so far we’ve basically taken the homicide rate as exogenous, but of course there’s reverse causality. Having a large chunk of the population in prison will affect the murder rate. What we really want on the x-axis might be some measure of the homicides we would get if no one were in prison. Maybe this latent measure would put the US more to the right and closer to the trendline—it depends on how effective the each country is at catching murderous people, which seems hard to know.

Another way out for them is that maybe all the countries with similar homicide rates *should* imprison people as much as the US, but their institutions don’t function well enough.

Here’s a simple way this could be. First, it seems generically true that incarceration should increase until the marginal costs equal the marginal benefits—people just debate these quantities. Next, just suppose that US courts are especially good at convicting the guilty party, but in the other countries with similar homicide rates (which tend to be poorer) the courts are way less likely to have the right defendant. This kind of ineffectiveness in the criminal justice system would lower the marginal benefits of incarcerating someone without affecting the marginal costs (it’s still one person having to suffer through prison). In this

case, poorer countries with the same homicide rate should have much lower incarceration rates but would optimally increase them if their institutions were as good as in the US.

All the data and code used for this is available [here](#).

Notes:

After writing this I found [this article](#) with a similar graph from [Tapio Lappi-Seppälä](#) that shows that, using victimization rates on the x-axis, the US is once again a huge outlier.

I'm sure plenty of academic papers exist that do this better and in more detail, I'll update this post as I find out about them.

# Alex Irpan: "My AI Timelines Have Sped Up"

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

This is a linkpost for <https://www.alexirpan.com/2020/08/18/ai-timelines.html>

Blog post by Alex Irpan. The basic summary:

In 2015, I made the following forecasts about when AGI could happen.

- 10% chance by 2045
- 50% chance by 2050
- 90% chance by 2070

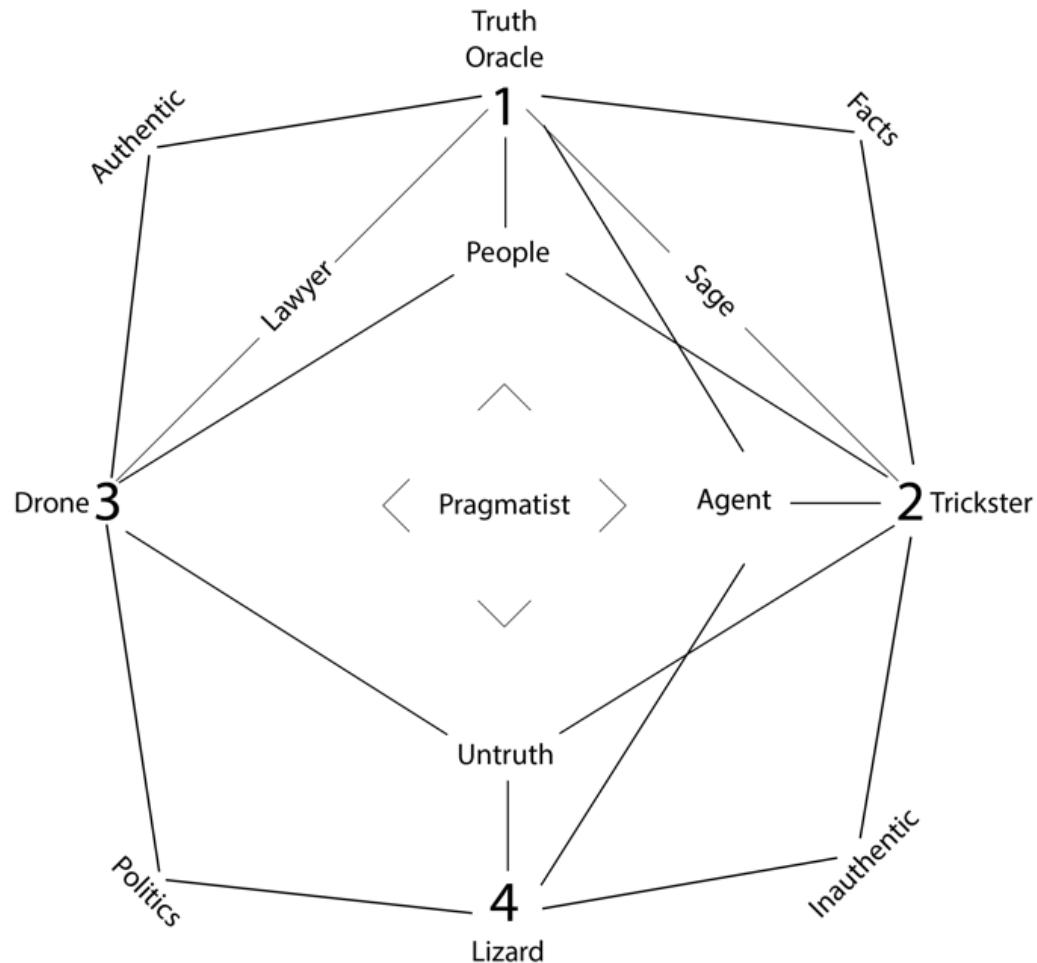
Now that it's 2020, I'm updating my forecast to:

- 10% chance by 2035
- 50% chance by 2045
- 90% chance by 2070

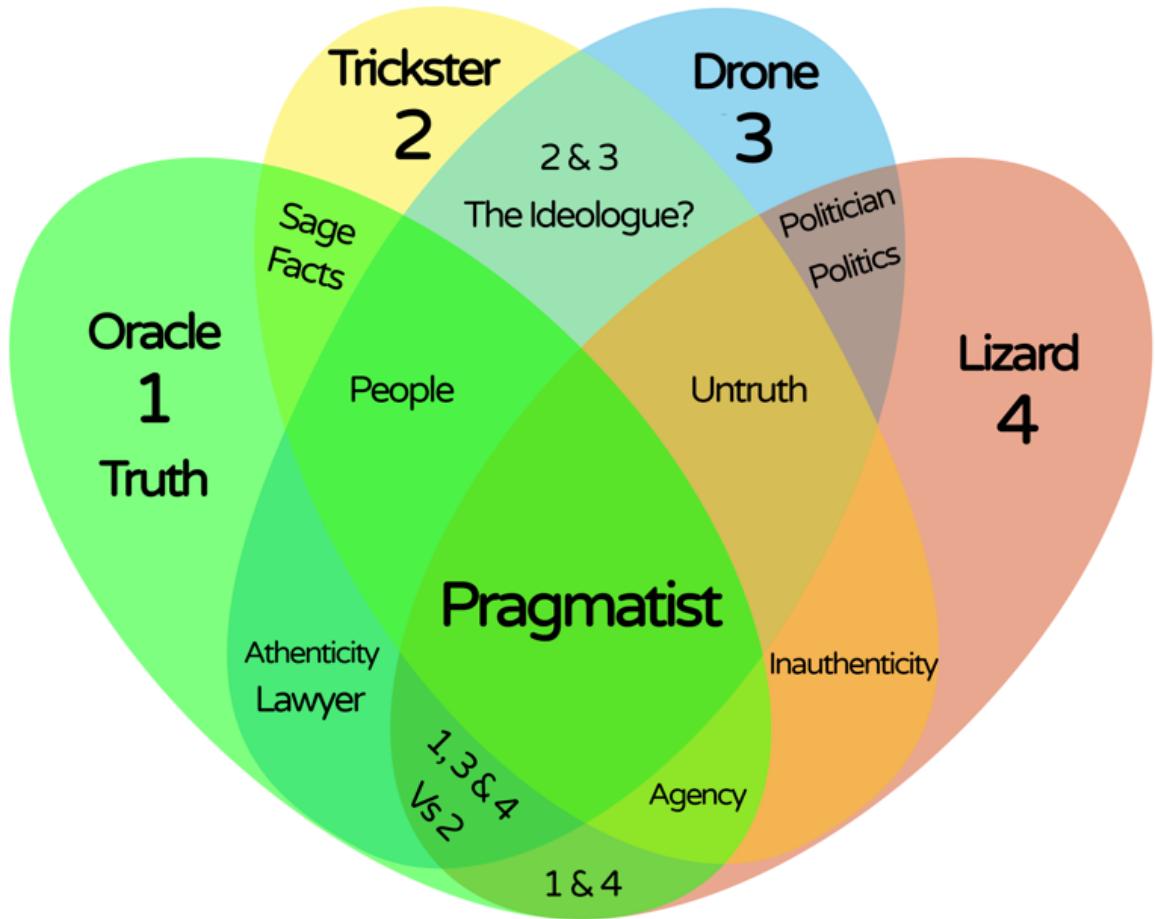
The main underlying shifts: more focus on improvements in tools, compute, and unsupervised learning.

# A sketch of 'Simulacra Levels and their Interactions'

Two sketches i made based on [Simulacra Levels and their Interactions](#). These are just sketches right now, i intend to make something better looking in the future (this is especially true for the first image). but i'd love to hear ideas and get feedback on these early versions.



The above diagram would look much better if there was symmetry. but the post misses some combos, some of which i also can't see how they're applicable (L1 & L2, for example).



The colors seem too happy for the topic to be honest :)

Here i also made a Venn version. it has [my idea for the ideologue](#), and then the only ones missing are (1, 3 & 4 VS 2) and (1 & 4). for the former i have a hard time thinking of something that would fit there, and for the latter I'm pretty sure there's isn't something that fits there.

# Model splintering: moving from one imperfect model to another

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

## 1. The big problem

In the last few months, I've become convinced that there is a key meta-issue in AI safety; a problem that seems to come up in all sorts of areas.

It's hard to summarise, but my best phrasing would be:

- Many problems in AI safety seem to be variations of "this approach seems safe in this imperfect model, but when we generalise the model more, it becomes dangerously underdefined". Call this **model splintering**.
- It is intrinsically worth studying how to (safely) transition from one imperfect model to another. This is worth doing, independently of whatever "perfect" or "ideal" model might be in the background of the imperfect models.

This sprawling post will be presenting examples of model splintering, arguments for its importance, a formal setting allowing us to talk about it, and some uses we can put this setting to.

### 1.1 In the language of traditional ML

In the language of traditional ML, we could connect all these issues to "[out-of-distribution](#)" behaviour. This is the problems that algorithms encounter when the set they are operating on is drawn from a different distribution than the training set they were trained on.

Humans can often see that the algorithm is out-of-distribution and correct it, because we have a more general distribution in mind than the one the algorithm was trained on.

In these terms, the issues of this post can be phrased as:

1. When the AI finds itself mildly out-of-distribution, how best can it extend its prior knowledge to the new situation?
2. What should the AI do if it finds itself strongly out-of-distribution?
3. What should the AI do if it finds itself strongly out-of-distribution, and humans don't know the correct distribution either?

### 1.2 Model splintering examples

Let's build a more general framework. Say that you start with some brilliant idea for AI safety/alignment/effectiveness. This idea is phrased in some (imperfect) model. Then

"model splintering" happens when you or the AI move to a new (also imperfect) model, such that the brilliant idea is undermined or underdefined.

Here are a few examples:

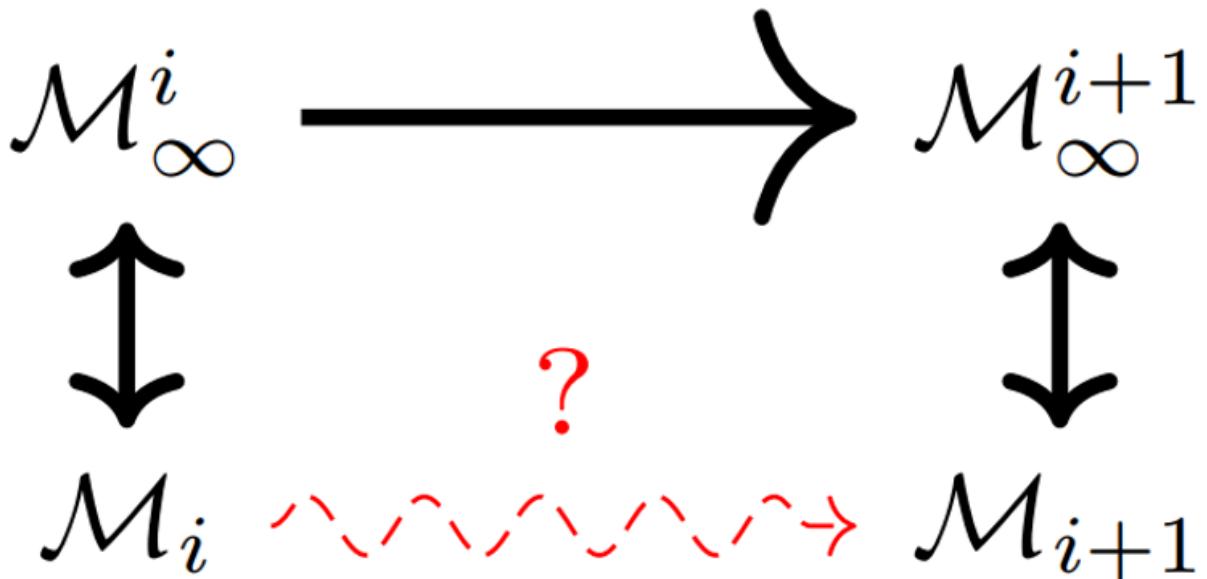
- You design an AI CEO as a money maximiser. Given typical assumptions about the human world (legal systems, difficulties in one person achieving massive power, human fallibilities), this results in an AI that behaves like a human CEO. But when those assumptions fail, the AI can end up feeding the universe to a money-making process that produces nothing of any value.
- Eliezer [defined](#) "rubes" as smooth red cubes containing palladium that don't glow in the dark. "Bleggs", on the other hand, are furred blue eggs containing vanadium that glow in the dark. To classify these, we only need a model with two features, "rubes" and "bleggs". Then along comes a furred red egg containing vanadium that doesn't glow in the dark. The previous model doesn't know what to do with it, and if you get a model with more features, it's unclear what to do with this new object.
- Here are some moral principles from history: honour is important for anyone. Women should be protected. Increasing happiness is important. These moral principles made sense in the world in which they were articulated, where features like "honour", "gender", and "happiness" are relatively clear and unambiguous. But the world changed, and the models splintered. "Honour" became hopelessly confused centuries ago. Gender is currently finishing its long splintering (long before we got to today, gender started becoming less useful for classifying people, hence the consequences of gender splintered a long time before gender itself did). Happiness, or at least hedonic happiness, is still well defined, but we can clearly see how this is going to splinter when we talk about worlds of uploads or brain modification.
- Many transitions in the laws of physics - from the [ideal gas laws](#) to the more advanced [van der Waals equations](#), or from Newtonain physics to general relativity to quantum gravity - will cause splintering if preferences were articulated in concepts that don't carry over well.

## 1.3 Avoiding perfect models

In all those cases, there are ways of improving the transition, without needing to go via some idealised, perfect model. We want to define the AI CEO's task in more generality, but we don't need to define this across every possible universe - that is not needed to restrain its behaviour. We need to distinguish any blegg from any rube we are likely to encounter, we don't need to define the platonic essence of "bleggness". For future splinterings - when hedonic happiness splinters, when we get a model of quantum gravity, etc... - we want to know what to do then and there, even if there are future splinterings subsequent to those.

And I think that model splintering is best addressed directly, rather than using methods that go via some idealised perfect model. Most approaches seem to go for approximating an ideal: from AIXI's [set of all programs](#), the [universal prior](#), [KWIK \("Knowing what it knows"\) learning](#) with a full hypothesis class, [Active Inverse Reward Design](#) with its full space of "true" reward functions, to Q-learning which assumes any [Markov decisions process](#) is possible. Then the practical approaches rely on approximating this ideal.

Schematically, we can see  $M_\infty$  as the ideal,  $M_\infty^i$  as  $M_\infty$  updated with information to time  $i$ , and  $M_i$  as an approximation of  $M_\infty^i$ . Then we tend to focus on how well  $M_i$  approximates  $M_\infty^i$ , and on how  $M_\infty^i$  changes to  $M_\infty^{i+1}$  - rather than on how  $M_i$  relates to  $M_{i+1}$ ; the red arrow here is underanalysed:



## 2 Why focus on the transition?

But why is focusing on the  $M_i \rightarrow M_{i+1}$  transition important?

### 2.1 Humans reason like this

A lot has been written about image recognition programs going "out-of-distribution" (encountering situations beyond its training environment) or succumbing to "adversarial examples" (examples from one category that have the features of another). Indeed, some people have [shown how to use labelled adversarial examples](#) to improve image recognition.

You know what this reminds me of? Human moral reasoning. At various points in our lives, we humans seem to have pretty solid moral intuitions about how the world should be. And then, we typically learn more, realise that things don't fit in the categories we were used to (go "out-of-distribution") and have to update. Some people push stories at us that exploit some of our emotions in new, more ambiguous circumstances ("adversarial examples"). And philosophers use similarly-designed thought experiments to open up and clarify our moral intuitions.

Basically, [we start with strong moral intuitions on under-defined features](#), and when the features splinter, we have to figure out what to do with our previous moral intuitions. A lot of developing moral meta-intuitions, is about learning how to navigate these kinds of transitions; AIs need to be able to do so too.

## 2.2 There are no well-defined overarching moral principles

Moral realists and moral non-realists [agree more than you'd think](#). In this situation, we can agree on one thing: there is no well-described system of morality that can be "simply" implemented in AI.

To over-simplify, moral realists hope to discover this moral system, moral non-realists hope to construct one. But, currently, it doesn't exist in an implementable form, nor is there any implementable algorithm to discover/construct it. So the whole idea of approximating an ideal is wrong.

All humans seem to start from a partial list of moral rules of thumb, [rules that they then have to extend to new situations](#). And most humans do seem to have some meta-rules for defining moral improvements, or extensions to new situations.

We don't know perfection, but we do know improvements and extensions. So methods that deal explicitly with that are useful. Those are things we can build on.

## 2.3 It helps distinguish areas where AIs fail, from areas where humans are uncertain

Sometimes the AI goes out-of-distribution, and humans can see the error (no, [flipping the lego block doesn't count as putting it on top of the other](#)). There are cases when humans themselves go out-of-distribution (see for example [siren worlds](#)).

It's useful to have methods available for both AIs and humans in these situations, and to distinguish them. "Genuine human preferences, not expressed in sufficient detail" is not the same as "human preferences fundamentally underdefined".

In the first case, it needs more human feedback; in the second case, it needs to figure out way of [resolving the ambiguity](#), knowing that soliciting feedback is not enough.

## 2.4 We don't need to make the problems harder

Suppose that quantum mechanics is the true underlying physics of the universe, with some added bits to include gravity. If that's true, why would we need a moral theory valid in every possible universe? It would be useful to have that, but would be strictly harder than one valid in the actual universe.

Also, some problems might be [entirely avoided](#). We don't need to figure out the morality of dealing with a willing slave race - if we never encounter or build one in the first place.

So a few degrees of "extend this moral model in a reasonable way" might be sufficient, without needing to solve the whole problem. Or, at least, without needing to solve the whole problem in advance - a successful [nanny AI](#) might be built on these kinds of extensions.

## 2.5 We don't know how deep the rabbit hole goes

In a sort of converse to the previous point, what if the laws of physics are radically different from what we thought - what if, for example, they allow some forms of time-travel, or have some [narrative features](#), or, more simply, what if the agent moves to an [embedded agency model](#)? What if [hypercomputation](#) is possible?

It's easy to have an idealised version of "all reality" that doesn't allow for these possibilities, so the ideal can be too restrictive, rather than too general. But the model splintering methods might still work, since it deals with transitions, not ideals.

Note that, **in retrospect**, we can always put this in a Bayesian framework, once we have a rich enough set of environments and updates rules. But this is misleading: the key issue is the missing feature, and figuring out what to do with the missing feature is the real challenge. The fact that we could have done this in a Bayesian way *if we already knew that feature*, is not relevant here.

## 2.6 We often only need to solve partial problems

Assume the blegg and rube classifier is an industrial robot performing a task. If humans filter out any atypical bleggs and rubes before it sees them, then the robot has no need for a full theory of bleggness/rubeness.

But what if the human filtering is not perfect? Then the classifier still doesn't need a full theory of bleggness/rubeness; it needs methods for dealing with the ambiguities it actually encounters.

Some ideas for AI control - [low impact](#), [AI-as-service](#), [Oracles](#), ... - may require dealing with some model splintering, some ambiguity, but not the whole amount.

## 2.7 It points out when to be conservative

Some methods, like [quantilizers](#) or the [pessimism approach](#) rely on the algorithm having a certain degree of conservatism. But, as I've [argued](#), it's not clear to what extent these methods actually are conservative, nor is it easy to calibrate them in a useful way.

Model splintering situations provide excellent points at which to be conservative. Or, for algorithms that need human feedback, but not constantly, these are excellent points to ask for that feedback.

## 2.8 Difficulty in capturing splintering from the idealised perspective

Generally speaking, idealised methods can't capture model splintering at the point we would want it to. Imagine an [ontological crisis](#), as we move from classical physics to quantum mechanics.

AIXI can go over the transition fine: it shifts from a Turing machine mimicking classical physics observations, to one mimicking quantum observations. But it doesn't notice anything special about the transition: changing the probability of various Turing machines is what it does with observations in general; there's nothing in its algorithm that shows that something unusual has occurred for this particular shift.

## 2.9 It may help amplification and distillation

This could be seen as a sub-point of some of the previous two sections, but it deserves to be flagged explicitly, since [iterated amplification and distillation](#) is one of the major potential routes to AI safety.

To quote a line from that summary post:

5. The proposed AI design is to use a safe but slow way of scaling up an AI's capabilities, distill this into a faster but slightly weaker AI, which can be scaled up safely again, and to iterate the process until we have a fast and powerful AI.

At both "scaling up an AI's capabilities", and "distill this into", we can ask the question: has the problem the AI is working on changed? The distillation step is more of a classical AI safety issue, as we wonder whether the distillation has caused any value drift. But at the scaling up or amplification step, we can ask: since the AI's capabilities have changed, the set of possible environments it operates in has changed as well. Has this caused a splintering where the previously safe goals of the AI have become dangerous.

Detecting and dealing with such a splintering could both be useful tools to add to this method.

## 2.10 Examples of model splintering problems/approaches

At a meta level, most problems in AI safety seem to be variants of model splintering, including:

- The [hidden complexity of wishes](#).
- [Ontological crises](#).
- [Conservative/prudential](#) behaviour in algorithms (more specifically, when the algorithm should become conservative).
- How [categories are defined](#).
- The [Goodhart problems](#).
- [Out-of-distribution](#) behaviour.

- [Low impact](#) and [reduced side-effects](#) approaches.
- [Underdefined preferences](#).
- [Active inverse reward design](#).
- [Inductive ambiguity identification](#).
- [Wireheading](#).
- The [whole friendly AI problem](#) itself.

Almost every recent post I've read in AI safety, I've been able to connect back to this central idea. Now, we have to be cautious - [cure-all's cure nothing](#), after all, so it's not necessarily a positive sign that *everything* seems to fit into this framework.

Still, I think it's worth diving into this, especially as I've come up with a framework that seems promising for actually solving this issue in many cases.

In a similar concept-space is Abram's [orthodox case against utility functions](#), where he talks about the [Jeffrey-Bolker axioms](#), which allows the construction of preferences from events *without needing full worlds at all*.

## 3 The virtues of formalisms

This post is dedicated to explicitly modelling the transition to ambiguity, and then showing what we can gain from this explicit meta-modelling. It will do with some formal language (made fully formal in [this post](#)), and a lot of examples.

Just as Scott argues that [if it's worth doing, it's worth doing with made up statistics](#), I'd argue that if an idea is worth pursuing, it's worth pursuing with an attempted formalism.

Formalisms are great at illustrating the problems, clarifying ideas, and making us familiar with the intricacies of the overall concept. That's the reason that this post (and the accompanying [technical post](#)) will attempt to make the formalism reasonably rigorous. I've learnt a lot about this in the process of formalisation.

### 3.1 A model, in (almost) all generality

What do we mean by a model? Do we mean mathematical [model theory](#)? As we talking about causal models, or [causal graphs](#)? AIXI uses a distribution over possible Turing machines, whereas [Markov Decision Processes](#) (MDPs) sees states and actions updating stochastically, independently at each time-step. Unlike the previous two, Newtonian mechanics doesn't use time-steps but continuous times, while general relativity weaves time into the structure of space itself.

And what does it mean for a model to make "predictions"? AIXI and MDPs make prediction over future observations, and causal graphs are similar. We can also try running them in reverse, "predicting" past observations from current ones.

Mathematical model theory talks about properties and the existence or non-existence of certain objects. Ideal gas laws make a "prediction" of certain properties (eg temperature) given certain others (eg volume, pressure, amount of substance). General relativity establishes that the structure of space-time must obey certain constraints.

It seems tricky to include all these models under the same meta-model formalism, but it would be good to do so. That's because of the risk of [ontological crises](#): we want the AI to be able to continue functioning even if the initial model we gave it was incomplete or incorrect.

## 3.2 Meta-model: models, features, environments, probabilities

All of the models mentioned above share one common characteristic: once you know some facts, you can deduce some other facts (at least probabilistically). A prediction of the next time step, a retrodiction of the past, a deduction of some properties from others, or a constraint on the shape of the universe: all of these say that if we know some things, then this puts constraints on some other things.

So let's define  $F$ , informally, as the set of *features* of a model. This could be the gas pressure in a room, a set of past observations, the local curvature of space-time, the momentum of a particle, and so on.

So we can define a prediction as a probability distribution over a set of possible features  $F_1$ , given a base set of features,  $F_2$ :

$$Q(F_1 \mid F_2).$$

Do we need anything else? Yes, we need a set of possible environments for which the model is (somewhat) valid. Newtonian physics fails at extreme energies, speeds, or gravitational fields; we'd like to include this "domain of validity" in the model definition. This will be very useful for extending models, or transitioning from one model to another.

You might be tempted to define a set of "worlds" on which the model is valid. But we're trying to avoid that, as the "worlds" may not be very useful for understanding the model. Moreover, we don't have special access to the underlying reality; so we never know whether there actually is a Turing machine behind the world or not.

So define  $E$ , the environment on which the model is valid, as a set of possible features.

So if we want to talk about Newtonian mechanics,  $F$  would be a set of Newtonian features (mass, velocity, distance, time, angular momentum, and so on) and  $E$  would be the set of these values where [relativistic and quantum effects make little difference](#).

So see a model as

$$M = \{F, E, Q\},$$

for  $F$  a set of features,  $E$  a set of environments, and  $Q$  a probability distribution. This is such that, for  $E_1, E_2 \subset E$ , we have the conditional probability:

$$Q(E_1 \mid E_2).$$

Though  $Q$  is defined for  $E$ , we generally want it to be usable from small subsets of the features: so  $Q$  should be simple to define from  $F$ . And we'll often define the subsets  $E_i$  in similar ways; so  $E_1$  might be all environments with a certain angular momentum at time  $t = 0$ , while  $E_2$  might be all environments with a certain angular momentum at a later time.

The full formal definition of these can be found [here](#). The idea is to have a meta-model of modelling that is sufficiently general to apply to almost all models, but not one that relies on some ideal or perfect formalism.

### 3.3 Bayesian models within this meta-model

It's very easy to include Bayesian models within this formalism. If we have a Bayesian model that includes a set  $W$  of worlds with prior  $P$ , then we merely have to define a set of features  $F$  that is sufficient to distinguish all worlds in  $W$ : each world is uniquely defined by its feature values<sup>[1]</sup>. Then we can define  $E$  as  $W$ , and  $P$  on  $W$  becomes  $Q$  on  $E$ ; the definitions of terms like  $Q(E_1 \mid E_2)$  is just  $P(E_1 \cap E_2)P(E_1)/P(E_2)$ , per Bayes' rules (unless  $P(E_2) = 0$ , in which case we set that to 0).

## 4 Model refinement and splinterings

This section will look at what we can do with the previous meta-model, looking at refinement (how models can improve) and splintering (how improvements to the model can make some well-defined concepts less well-defined).

### 4.1 Model refinement

Informally,  $M^* = \{F^*, E^*, Q^*\}$  is a *refinement* of model  $M = \{F, E, Q\}$  if it's at least as expressive as  $M$  (it covers the same environments) and is better according to some criteria (simpler, or more accurate in practice, or some other measurement).

\*  
At the technical level, we have a map  $q$  from a subset  $E_0$  of  $E^*$ , that is surjective onto  $E$ . This covers the "at least as expressive" part: every environment in  $E$  exists as (possibly multiple) environments in  $E^*$ .

Then note that using  $q^{-1}$  as a map from subsets of  $E$  to subsets of  $E_0^*$ , we can define  $Q_0^*$  on  $E$  via:

$$Q_0^*(E_1 \mid E_2) = Q^*(q^{-1}(E_1) \mid q^{-1}(E_2)).$$

Then this is a model refinement if  $Q_0^*$  is 'at least as good as'  $Q$  on  $E$ , according to our criteria [2].

## 4.2 Example of model refinement: gas laws

[This post](#) presents some subclasses of model refinement, including Q-improvements (same features, same environments, just a better  $Q$ ), or adding new features to a basic model, called "non-independent feature extension" (eg adding classical electromagnetism to Newtonian mechanics).

Here's a specific gas law illustration. Let  $M = \{F, E, Q\}$  be a model of [an ideal gas](#), in some set of rooms and tubes. The  $F$  consists of pressure, volume, temperature, and amount of substance, and  $Q$  is the ideal gas laws. The  $E$  is the [standard conditions for temperature and pressure](#), where the ideal gas law applies. There are multiple different types of gases in the world, but they all roughly obey the same laws.

Then compare with model  $M^* = \{F^*, E^*, Q^*\}$ . The  $F^*$  has all the features of  $F$ , but also includes the volume that is occupied by one mole of the molecules of the given substance. This allows  $Q^*$  to express the more complicated [van der Waals equations](#), which are different for different types of gases. The  $E^*$  can now track situations where there are gases with different molar volumes, which include situations where the van der Waals equations differ significantly from the ideal gas laws.

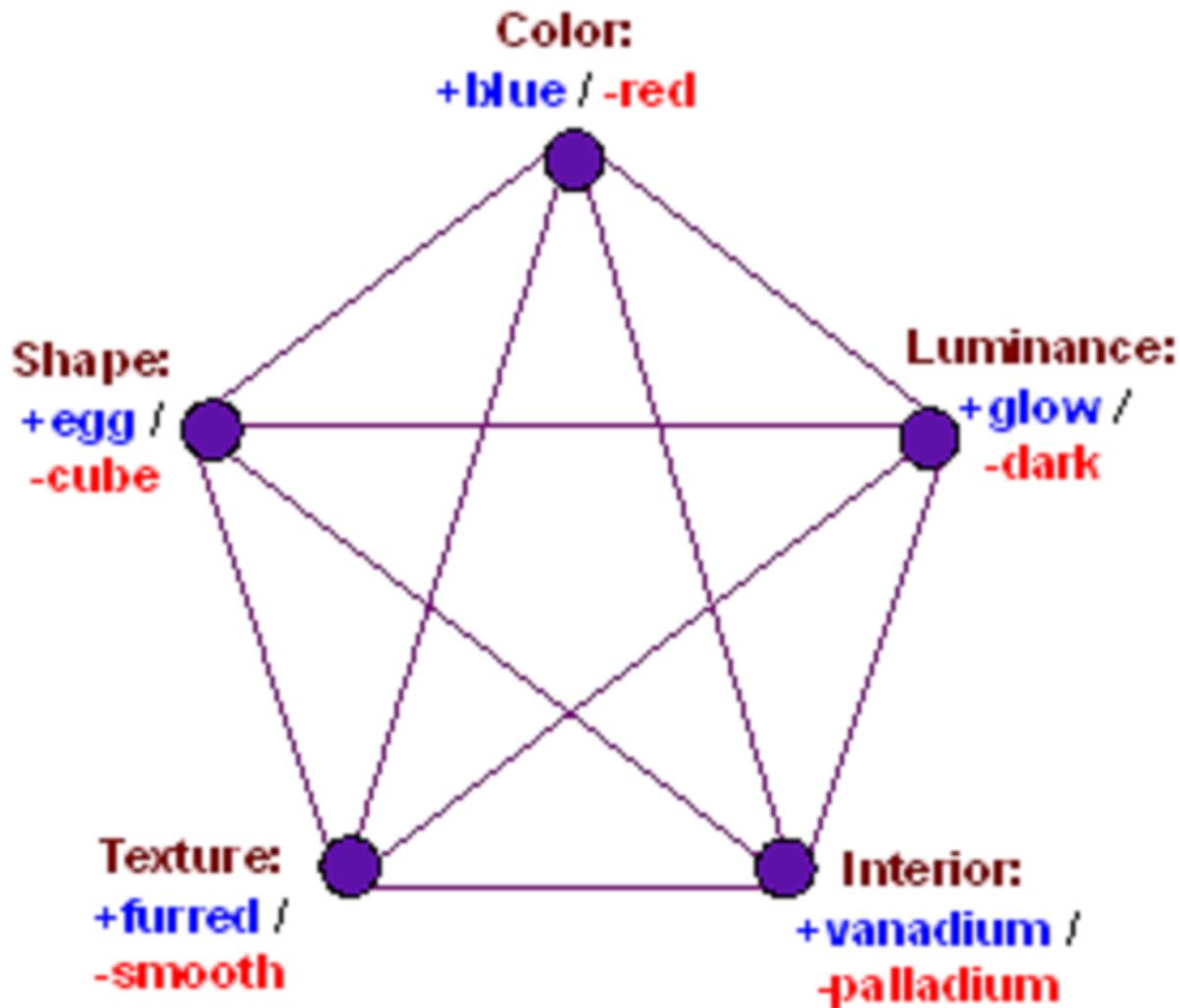
In this case  $E_0^* \subset E^*$ , since we now distinguish environments that we previously considered identical (environments with same features except for having molar volumes). The  $q$  is just projecting down by forgetting the molar volume. Then since

$Q_0^* = Q^*$  (van der Waals equations averaged over the distribution of molar volumes) is at least as accurate as  $Q$  (ideal gas law), this is a refinement.

## 4.3 Example of model refinement: rubes and bleegs

Let's reuse Eliezer's [example](#) of rubes ("red cubes") and bleegs ("blue eggs").

Bleegs are blue eggs that glow in the dark, have a furred surface, and are filled with vanadium. Rubes, in contrast, are red cubes that don't glow in the dark, have a smooth surface, and are filled with palladium:



Define  $M$  by having  $F = \{\text{red, smooth}\}$ ,  $E$  is the set of all bleegs and rubes in some situation, and  $Q$  is relatively trivial: it predicts that an object is red/blue if and only if is smooth/furred.

Define  $M^1$  as a refinement of  $M$ , by expanding  $F$  to  $F^1 = \{\text{red, smooth, cube, dark}\}$ . The projection  $q : E^* \rightarrow E$  is given by forgetting about those two last features. The  $Q^1$  is more detailed, as it now connects red-smooth-cube-dark together, and similarly for blue-furred-egg-glow.

Note that  $E^1$  is larger than  $E$ , because it includes, e.g., environments where the cube objects are blue. However, all these extra environments have probability zero.

## 4.4 Reward function refactoring

Let  $R$  be a reward function on  $M$  (by which we mean that  $R$  is defined on  $F$ , the set of features in  $M$ ), and  $M^*$  a refinement of  $M$ .

A *refactoring* of  $R$  for  $M^*$  is a reward function  $R^*$  on the features  $F^*$  such that for any  $e^* \in E_0^*$ ,

$$R^*(e^*) = R(q(e^*)).$$

For example, let  $M$  and  $M^1$  be from the rube/blegg models in the previous section. Let  $R_{\text{red}}$  on  $M$  simply count the number of rubes - or, more precisely, counts the number of objects to which the feature "red" applies.

Let  $R_{\text{red}}^1$  be the reward function that counts the number of objects in  $M^1$  to which "red" applies. It's clearly a refactoring of  $R_{\text{red}}$ .

But so is  $R_{\text{smooth}}^1$ , the reward function that counts the number of objects in  $M^1$  to which "smooth" applies. In fact, the following is a refactoring of  $R_{\text{red}}$ , for all  $\alpha + \beta + \gamma + \delta = 1$ :

$$\alpha R_{\text{red}}^1 + \beta R_{\text{smooth}}^1 + \gamma R_{\text{cube}}^1 + \delta R_{\text{dark}}^1.$$

There are also some non-linear combinations of these features that refactor  $R$ , and many other variants (like the strange combinations that generate concepts like [grue](#) and [bleen](#)).

## 4.5 Reward function splintering

Model splintering, in the informal sense, is what happens when we pass to a new models in a way that the old features (or a reward function defined by the old features)

no longer apply. It is similar to the [web of connotations](#) breaking down, an agent going [out of distribution](#), or the [definitions of Rube and Blegg falling apart](#).

- Preliminary definition: If  $M^*$  is a refinement of  $M$  and  $R$  a reward function on  $M$ , then  $M^*$  *splinters*  $R$  if there are multiple refactorings of  $R$  on  $M^*$  that disagree on elements of  $E^*$  of non-zero probability.

So, note that in the rube/blegg example,  $M^1$  is **not** a splintering of  $R_{\text{red}}$ : all the refactorings are the same on all bleggs and rubes - hence on all elements of  $E^1$  of non-zero probability.

We can even generalise this a bit. Let's assume that "red" and "blue" are not totally uniform; there exists some rubes that are "redish-purple", while some bleggs are "blueish-purple". Then let  $M^2$  be like  $M^1$ , except the colour feature can have four values: "red", "redish-purple", "blueish-purple", and "blue".

Then, as long as rubes (defined, in this instance, by being smooth-dark-cubes) are either "red" or "redish-purple", and the bleggs are "blue", or "blueish-purple", then all refactorings of  $R_{\text{red}}$  to  $M^2$  agree - because, on the test environment,  $R_{\text{red}}$  on  $F$  perfectly

$$\begin{matrix} 2 & 2 \\ R_{\text{red}} & + R_{\text{redish-purple}} \end{matrix} \text{ on } F^2$$

So adding more features does not always cause splintering.

## 4.6 Reward function splintering: "natural" refactorings

The preliminary definition runs into trouble when we add more objects to the environments. Define  $M^3$  as being the same as  $M^2$ , except that  $E^3$  contains one extra object,  $o_+$ ; apart from that, the environments typically have a billion rubes and a trillion bleggs.

Suppose  $o_+$  is a "furred-rube", i.e. a red-furred-dark-cube. Then  $R_{\text{red}}^3$  and  $R_{\text{smooth}}^3$  are two different refactorings of  $R_{\text{red}}$ , that obviously disagree on any environment that contains  $o_+$ . Even if the probability of  $o_+$  is tiny (but non-zero), then  $M^3$  splinters  $R$ .

But things are worse than that. Suppose that  $o_+$  is fully a rube: red-smooth-cube-dark,

and even contains palladium. Define  $(R_{\text{red}})^3$  as being counting the number of red

objects, except for  $o_+$  specifically (again, this is similar to the [grue and bleen arguments against induction](#)).

Then both  $(R_{\text{red}})^3$  and  $R_{\text{red}}^3$  are refactorings of  $R_{\text{red}}$ , so  $M^3$  still splinters  $R_{\text{red}}$ , even when we add another exact copy of the elements in the training set. Or even if we keep the training set for a few extra seconds, or add any change to the world.

So, for any  $M^*$  a refinement of  $M$ , and  $R$  a reward function on  $E$ , let's define "natural refactorings" of  $R$ :

- The reward function  $R^*$  is a natural refactoring of  $R$  if it's a reward function on  $M^*$  with:
  1.  $R^* \approx R \circ q$  on  $E_0$ , and
  2.  $R^*$  can be defined simply from  $F^*$  and  $R$ ,
  3. the  $F^*$  themselves are simply defined.

This leads to a full definition of splintering:

- Full definition: If  $M^*$  is a refinement of  $M$  and  $R$  a reward function on  $M$ , then  $M^*$  *splinters*  $R$  if 1) there are no natural refactorings of  $R$  on  $M^*$ , or 2) there are multiple natural refactorings  $R^*$  and  $R^{*\prime}$  of  $R$  on  $M^*$ , such that  $R^* \neq R^{*\prime}$ .

Notice the whole host of caveats and weaselly terms here;  $R^* \approx R \circ q$ , "simply" (used twice), and  $R^* \neq R^{*\prime}$ . Simply might mean [algorithmic simplicity](#), but  $\approx$  and  $\neq$  are measures of how much "error" we are willing to accept in these refactorings. Given that, we probably want to replace  $\approx$  and  $\neq$  with some *measure* of non-equality, so we can talk about the "degree of naturalness" or the "degree of splintering" of some refinement and reward function.

Note also that:

- **Different choices of refinements can result in different natural refactorings.**

An easy example: it makes a big difference whether a new feature is "temperature", or "divergence from standard temperatures".

## 4.7 Splintering training rewards

The concept of "reward refactoring" is transitive, but the concept of "natural reward refactoring" need not be.

For example, let  $E_t$  be a training environment where red/blue  $\iff$  cube/egg, and  $E_g$  be a general environment where red/blue is independent of cube/egg. Let  $F^1$  be a feature set with only red/blue, and  $F^2$  a feature set with red/blue and cube/egg.

Then define  $M_t^1$  as using  $F^1$  in the training environment,  $M_g^2$  as using  $F^2$  in the general environment;  $M_g^1$  and  $M_t^2$  are defined similarly.

For these models,  $M_g^1$  and  $M_t^2$  are both refinements of  $M_t^1$ , while  $M_g^2$  is a refinement of all three other models. Define  $R_t^1$  as the "count red objects" reward on  $M_t^1$ . This has a natural refactoring to  $R_g^1$  on  $M_g^1$ , which counts red objects in the general environment.

And  $R_g^1$  has a natural refactoring to  $R_g^2$  on  $M_g^2$ , which still just counts the red objects in the general environment.

But there is no natural refactoring from  $R_t^1$  directly to  $M_g^2$ . That's because, from  $F^2$ 's perspective,  $R_t^1$  on  $M_t^1$  might be counting red objects, or might be counting cubes. This is not true for  $R_g^1$  on  $M_g^1$ , which is clearly only counting red objects.

Thus when a reward function come from a training environment, we'd want our AI to look for splinterings **directly from a model of the training environment**, rather than from previous natural refactorings.

## 4.8 Splintering features and models

We can also talk about splintering features and models themselves. For  $M = \{F, E, Q\}$ , the easiest way is to define a reward function  $R_{F, S_F}$  as being the indicator function for feature  $F \in F$  being in the set  $S_F$ .

Then a refinement  $M^*$  splinters the feature  $F$  if it splinters some  $R_{F,S_F}$ .

The refinement  $M^*$  splinters the model  $M$  if it splinters at least one of its features.

For example, if  $M$  is Newtonian mechanics, including "total rest mass" and  $M^*$  is special relativity, then  $M^*$  will splinter "total rest mass". Other examples of feature splintering will be presented in the rest of this post.

## 4.9 Preserved background features

A reward function developed in some training environment will ignore any feature that is always present or always absent in that environment. This allows very weird situations to come up, such as training an AI to distinguish happy humans from sad humans, and it ending up replacing humans with humanoid robots (after all, both happy and sad humans were equally non-robotic, so there's no reason not to do this).

Let's try and do better than that. Assume we have a model  $M = \{F, E, Q\}$ , with a reward function  $R_\tau$  defined on  $E$  ( $R_\tau$  and  $E$  can be seen as the training data).

Then the feature-preserving reward function  $R^M$ , is a function that constrains the environments to have similar feature distributions as  $E$  and  $Q$ . There are many ways this could be defined; here's one.

For an element  $e \in E$ , just define

$$R^M(e) = \log(Q(e)).$$

Obviously, this can be improved; we might want to coarse-grain  $F$ , grouping together similar worlds, and possibly bounding this below to avoid singularities.

Then we can use this to get the feature-preserving version of  $R_\tau$ , which we can define as

$$R_\tau^M = \frac{\max_{e \in E} R_\tau(e)}{\max_{e \in E} R_\tau(e)} \cdot R^M,$$

for  $\max_{e \in E} R_\tau(e)$  the maximal value of  $R_\tau$  on  $E$ . Other options can work as well, such as

$$R_\tau^M + \alpha R_\tau \text{ for some constant } \alpha > 0.$$

M

Then we can ask an AI to use  $R_\tau$  as its reward function, refactoring that, rather than  $R_\tau$ .

- A way of looking at it: a natural refactoring of a reward function  $R_\tau$  will preserve all the implicit features that correlate with  $R_\tau$ . But  $R_\tau$  will also preserve all the implicit features that stay constant when  $R_\tau$  was defined. So if  $R_\tau$  measures human happiness vs human unhappiness, a natural refactoring of it will preserves things like "having higher dopamine in their brain". But a natural refactoring of  $R_\tau$  will also preserve things like "having a brain".

## 4.10 Partially preserved background features

M

The  $R_\tau$  is almost certainly too restrictive to be of use. For example, if time is a feature, then this will fall apart when the AI has to do something after the training period. If all the humans in a training set share certain features, humans without those features will be penalised.

There are at least two things we can do to improve this. The first is to include more positive and negative examples in the training set; for example, if we include humans and robots in our training set - as positive and negative examples, respectively - then

M

this difference will show up in  $R_\tau$  directly, so we won't need to use  $R_\tau$  too much.

Another approach would be to explicitly allow certain features to range beyond their typical values in M, or allow highly correlated variables explicitly to decorrelate.

For example, though training during a time period  $t$  to  $t'$ , we could explicitly allow time to range beyond these values, without penalty. Similarly, if a medical AI was trained on examples of typical healthy humans, we could decorrelate functioning digestion from brain activity, and get the AI to focus on the second [3].

This has to be done with some care, as adding more degrees of freedom adds more ways for errors to happen. I'm aiming to look further at this issue in later posts.

## 5 The fundamental questions of model refinements and splintering

We can now rephrase the out-of-distribution issues of [section 1.1](#) in terms of the new formalism:

1. When the AI refines its model, what would count as a natural refactoring of its reward function?
2. If the refinements splinter its reward function, what should the AI do?
3. If the refinements splinter its reward function, and also splinters the human's reward function, what should the AI do?

## 6 Examples and applications

The rest of this post is applying this basic framework, and its basic insights, to various common AI safety problems and analyses. This section is not particularly structured, and will range widely (and wildly) across a variety of issues.

### 6.1 Extending beyond the training distribution

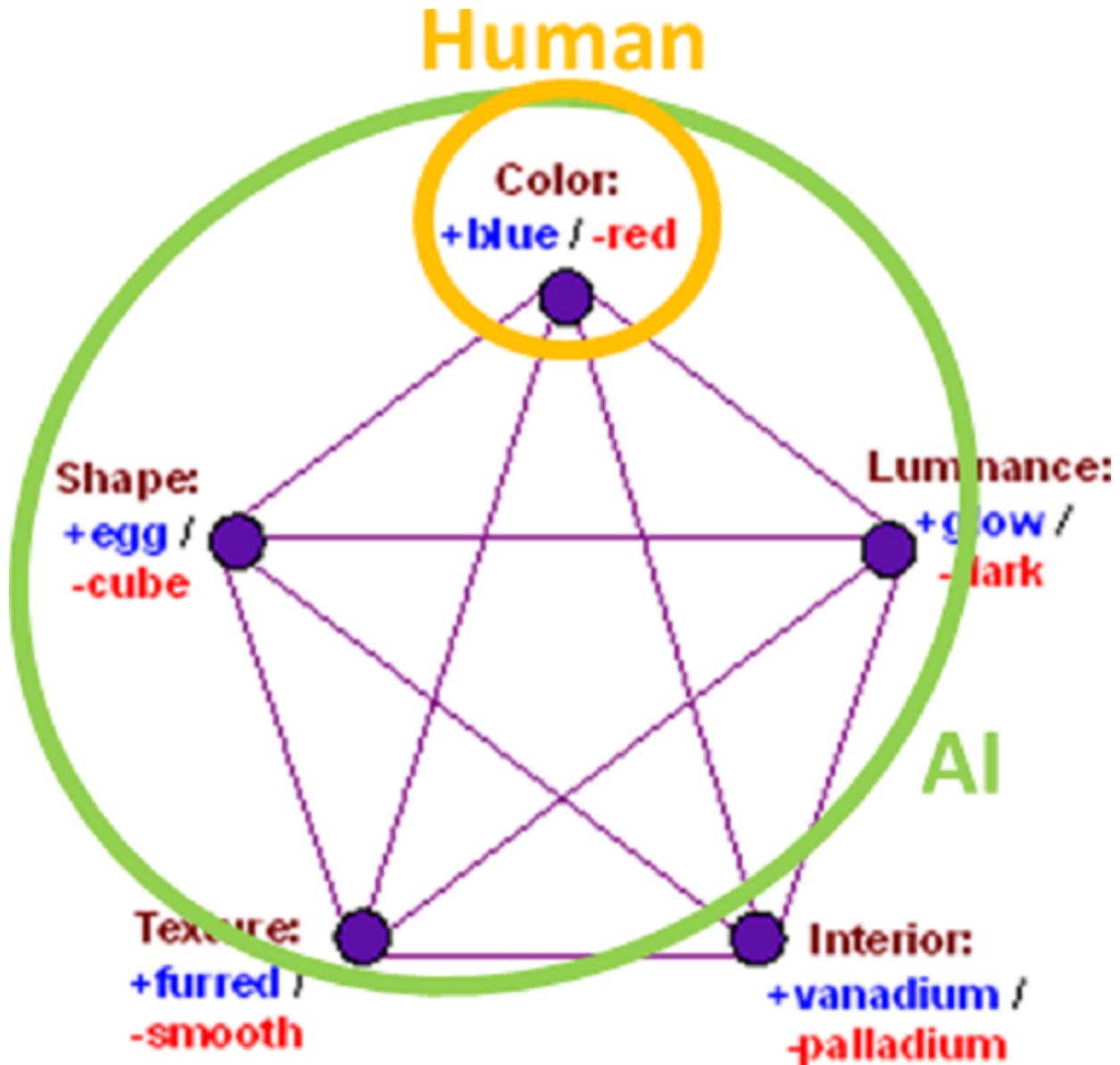
Let's go back to the blegg and rube examples. A human supervises an AI in a training environment, labelling all the rubes and bleggs for it.

The human is using a very simple model,  $M_H = \{F_H, E_t, Q\}$ , with the only feature being the colour of the object, and  $E_t$  being the training environment.

Meanwhile the AI, having more observational abilities and [no filter as to what can be ignored](#), notices their colour, their shape, their luminance, and their texture. It doesn't

know  $M_H$ , but is using model  $M_{AI} = \{F^{1,1,1}, E_t, Q^1\}$ , where  $F_{AI}$  covers those four features

(note that  $M_{AI}$  is a refinement of  $M_H$ , but that isn't relevant here).



Suppose that the AI is trained to be cube-classifier (and hence a blegg classifier by default). Let  $R_F$  be the reward function that counts the number of objects, with feature  $F$ , that the AI has classified as cubes. Then the AI could learn many different reward function in the training environment; here's one:

$$R^1 = R_{\text{cube}} + 0.5R_{\text{smooth}} + 0.5R_{\text{dark}} - R_{\text{red}}.$$

Note that, even though this gets the colour reward completely wrong, this reward matches up with the human's assessment on the training environment.

Now the AI moves to the larger testing environment  $E^2$ , and refines its model minimally

to  $M_{AI} = \{F^1, E^2, Q^1\}$  (extending  $R^1$  to  $R^2$  in the obvious way).

In  $E^2$ , the AI sometimes encounters objects that it can only see through their colour.

Will this be a problem, since the colour component of  $R^2$  is pointing in the wrong direction?

No. It still has  $Q^1$ , and can deduce that a red object must be cube-smooth-dark, so  $R^2$  will continue treating this as a rube<sup>[4]</sup>.

## 6.2 Detecting going out-of-distribution

Now imagine the AI learns about the content of the rubes and bleggs, and so refines to

a new model that includes vanadium/palladium as a feature in  $M_{AI}$ .

Furthermore, in the training environment, all rubes have palladium and all bleggs have

vanadium in them. So, for  $M_{AI}$ , a refinement of  $M_{AI}$ ,  $q^{-1}(E_{AI}) \subset E_{AI}$  has only palladium-

rubes and vanadium-bleggs. But in  $E_{AI}$ , the full environment, there are rather a lot of rubes with vanadium and bleggs with palladium.

So, similarly to [section 4.7](#), there is no natural refactoring of the rube/blegg reward in

$M_{AI}$ , to  $M_{AI}$ . That's because  $F_{AI}$ , the feature set of  $M_{AI}$ , includes vanadium/palladium which co-vary with the other rube/blegg features on the training environment ( $q^{-1}(\setminus E_{AI})^1$ ), but not on the full environment of  $E_{AI}$ .

So looking for reward splintering from the training environment is a way of detecting going out-of-distribution - even on features that were not initially detected in the training distribution, by either the human nor the AI.

## 6.3 Asking humans and Active IRL

Some of the most promising AI safety methods today rely on getting human feedback<sup>[5]</sup>. Since human feedback is expensive, as in it's slow and hard to get compared with almost all other aspects of algorithms, people want to [get this feedback in the most efficient ways possible](#).

A good way of doing this would be to ask for feedback when the AI's current reward function splinters, and multiple options are possible.

A more rigorous analysis would look at the value of information, expected future splinterings, and so on. This is what they do in [Active Inverse Reinforcement Learning](#); the main difference is that AIRL emphasises an unknown reward function with humans providing information, while this approach sees it more as an known reward function over uncertain features (or over features that may splinter in general environments).

## 6.4 A time for conservatism

I [argued](#) that many "conservative" AI optimising approaches, such as [quantilizers](#) and [pessimistic AIs](#), don't have a good measure of when to become more conservative; their parameters  $q$  and  $\beta$  don't encode useful guidelines for the right degree of conservatism.

In this framework, the alternative is obvious: AIs should become conservative when their reward functions splinter (meaning that the reward function compatible with the previous environment has multiple natural refactorings), and very conservative when they splinter a lot.

This design is very similar to [Inverse Reward Design](#). In that situation, the reward signal in the training environment is taken as *information* about the "true" reward function. Basically they take all reward functions that could have given the specific reward signals, and assume the "true" reward function is one of them. In that paper, they advocate extreme conservatism at that point, by optimising the minimum of all possible reward functions.

The idea here is almost the same, though with more emphasis on "having a true reward defined on uncertain features". Having multiple contradictory reward functions compatible with the information, in the general environment, is equivalent with having a lot of splintering of the training reward function.

## 6.5 Avoiding ambiguous distant situations

The post "[By default, avoid ambiguous distant situations](#)" can be rephrased as: let  $M$  be a model in which we have a clear reward function  $R$ , and let  $M^2$  be a refinement of this to general situations. We expect that this refinement splinters  $R$ . Let  $M^1$  be like  $M^2$ , except with  $E^1$  smaller than  $E^2$ , defined such that:

1. An AI could be expected to be able to constrain the world to be in  $E^1$ , with high probability,
2. The  $M^1$  is not a splintering of  $R$ .

Then that post can be summarised as:

- The AI should constrain the world to be in  $E^1$  and then maximise the natural refactoring of  $R$  in  $M^1$ .

## 6.6 Extra variables

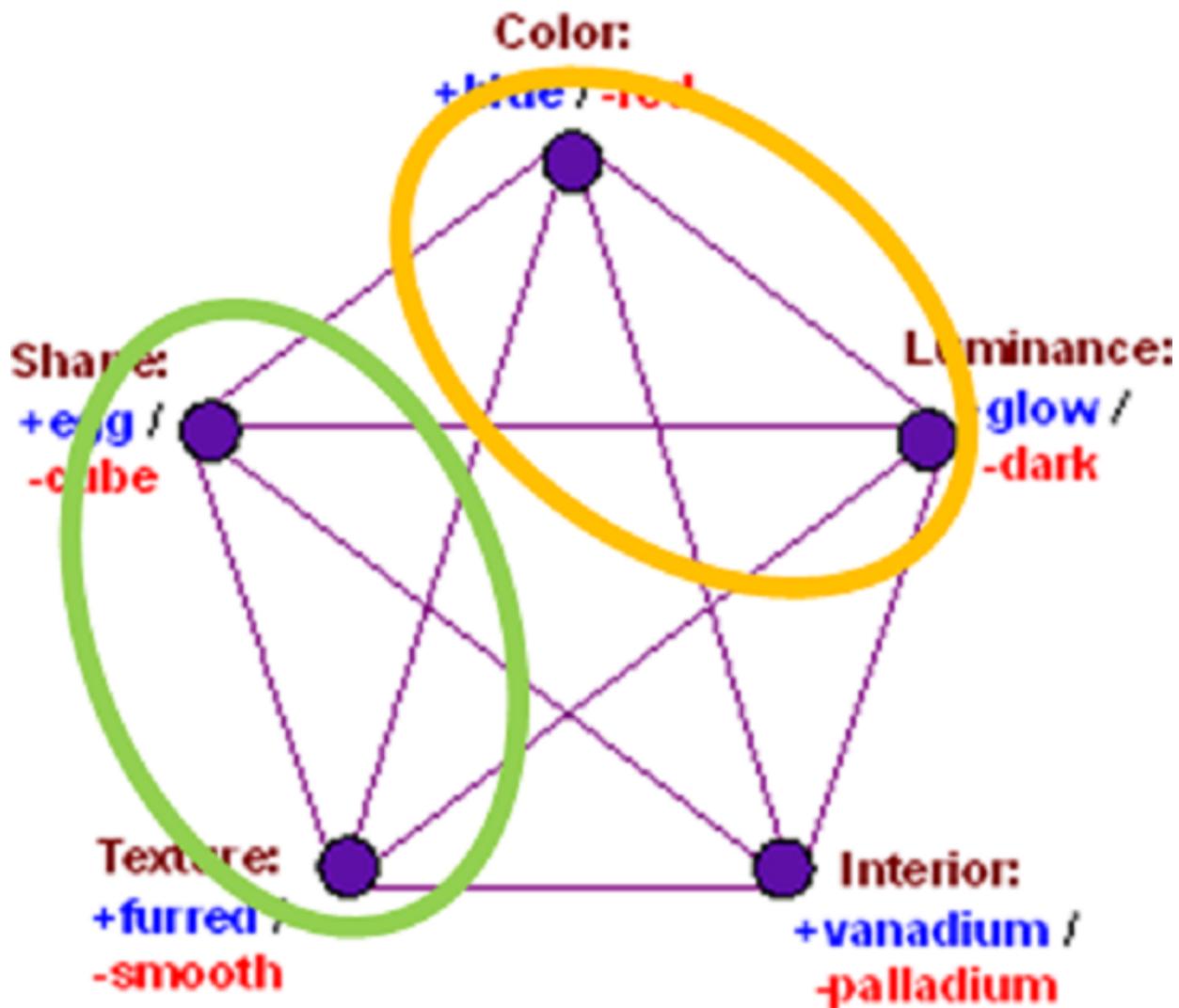
Stuart Russell [writes](#):

A system that is optimizing a function of  $n$  variables, where the objective depends on a subset of size  $k < n$ , will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable.

The approach in [sections 4.9](#) and [4.10](#) explicitly deals with this.

## 6.7 Hidden (dis)agreement and interpretability

Now consider two agents doing a rube/blegg classifications task in the training environment; each agent only models two of the features:



Despite not having a single feature in common, both agents will agree on what bleggs and rubes are, in the training environment. And when refining to a fuller model that includes all four (or five) of the key features, both agents will agree as to whether a natural refactoring is possible or not.

This can be used to help define the limits of [interpretability](#). The AI can use its own model, and [its own designed features](#), to define the categories and rewards in the training environment. These need not be human-parsable, but we can attempt to interpret them in human terms. And then we can give this interpretation to the AI, as a list of positive and negative examples of our interpretation.

If we do this well, the AI's own features and our interpretation will match up in the training environment. But as we move to more general environments, these may diverge. Then the AI will flag a "failure of interpretation" when its refactoring diverges from a refactoring of our interpretation.

For example, if we think the AI detects pandas by looking for white hair on the body, and black hair on the arms, we can flag lots of examples of pandas and that hair pattern (and non-pandas and [unusual hair patterns](#)). We don't use these examples for

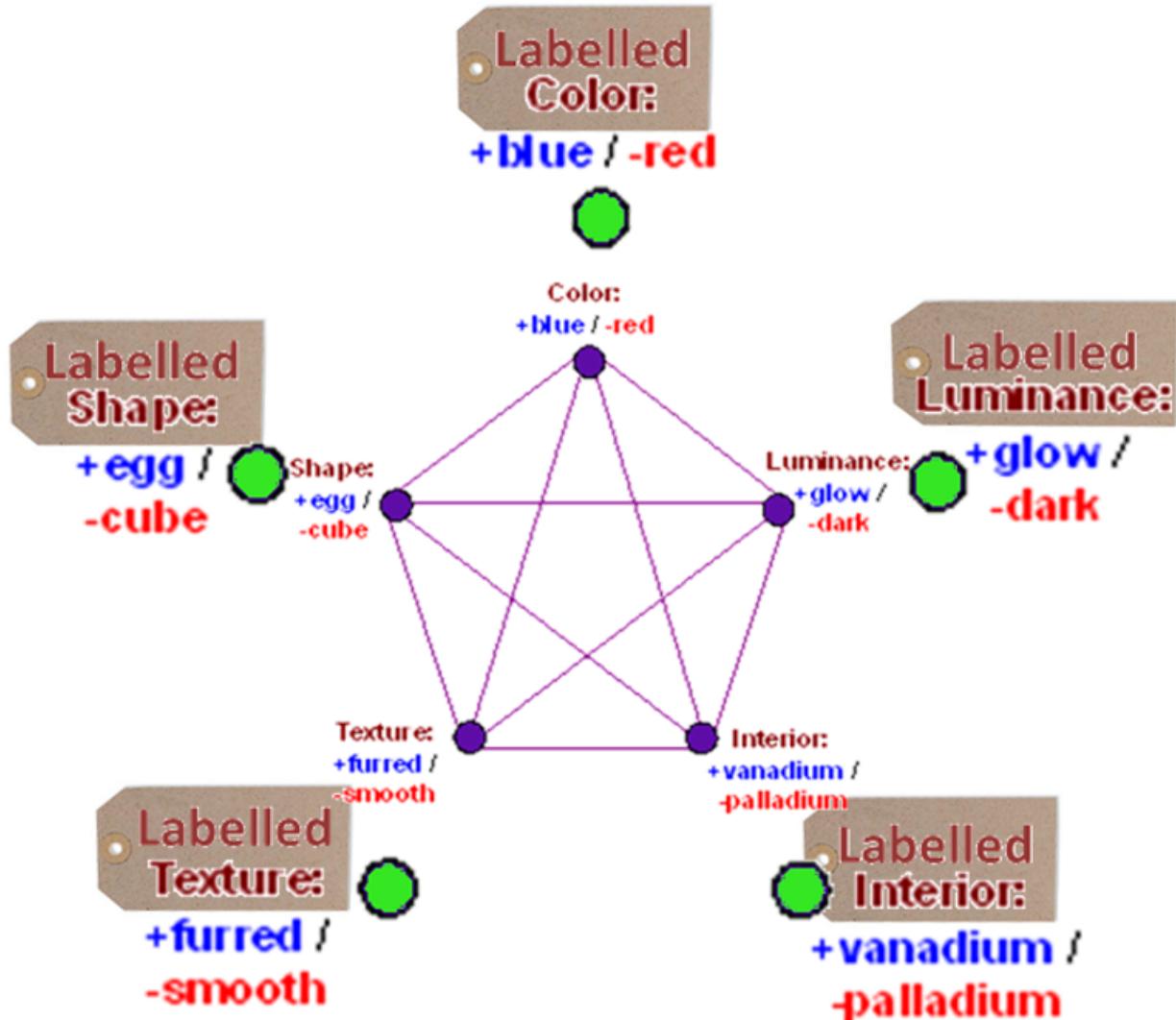
training the AI, just to confirm that, in the training environment, there is a match between "AI-thinks-they-are-pandas" and "white-hair-on-arms-black-hair-on-bodies".

But, [in an adversarial example](#), the AI could detect that, while it is detecting gibbons, this no longer matches up with our interpretation. A splintering of interpretations, if you want.

## 6.8 Wireheading

The approach can also be used to detect [wireheading](#). Imagine that the AI has various detectors that allow it to label what the features of the bleggs and rubes are. It models the world with ten features: 5 features representing the "real world" versions of the features, and 5 representing the "this signal comes from my detector" versions.

This gives a total of 10 features, the 5 features "in the real world" and the 5 "AI-labelled" versions of these:



In the training environment, there was full overlap between these 10 features, so the AI might learn the incorrect "maximise my labels/detector signal" reward.

However, when it refines its model to all 10 features and environments where labels and underlying reality diverge, it will realise that this splinters the reward, and thus detect a possible wireheading. It could then ask for more information, or have an automated "don't wirehead" approach.

## 6.9 Hypotheticals, and training in virtual environments

To get around the slowness of the real world, some approaches [train AIs in virtual environments](#). The problem is to pass that learning from the virtual environment to the real one.

Some have suggested making the virtual environment sufficiently detailed that the AI can't tell the difference between it and the real world. But, a) this involves fooling the AI, an approach I'm always wary of, and b) it's unnecessary.

Within the meta-formalism of this post, we could train the AI in a virtual environment which it models by  $M$ , and let it construct a model  $M'$  of the real-world. We would then motivate the AI to find the "closest match" between  $M$  and  $M'$ , in terms of features and how they connect and vary. This is similar to how we can train pilots in flight simulators; the pilots are never under any illusion as to whether this is the real world or not, and even crude simulators can allow them to build certain skills<sup>[6]</sup>.

This can also be used to allow the AI to deduce information from hypotheticals and thought experiments. If we show the AI an episode of a TV series showing people behaving morally (or immorally), then the episode need not be believable or plausible, if we can roughly point to the features in the episode that we want to emphasise, and roughly how these relate to real-world features.

## 6.10 Defining how to deal with multiple plausible refactorings

The approach for synthesising human preferences, [defined here](#), can be rephrased as:

- "Given that we expect multiple natural refactorings of human preferences, and given that we expect some of them to go [disastrously wrong](#), here is one way of resolving the splintering that we expect to be better than most."

This is just one way of doing this, but it does show that "automating what AIs do with multiple refactorings" might not be impossible. The following subsection has some ideas with how to deal with that.

## 6.11 Global, large scale preferences

In an [old post](#), I talked about the concept of "emergency learning", which was basically, "lots of examples, and all the stuff we know and suspect about how AIs can go wrong, shove it all in, and hope for the best". The "shove it all in" was a bit more structured than that, defining large scale preferences (like "avoid siren worlds" and "don't over-optimise") as constraints to be added to the learning process.

It seems we can do better than that here. Using examples and hypotheticals, it seems we could construct ideas like "avoid slavery", "avoid siren worlds", or "don't over-optimise" as rewards or positive/negative examples certain simple training environments, so that the AI "gets an idea of what we want".

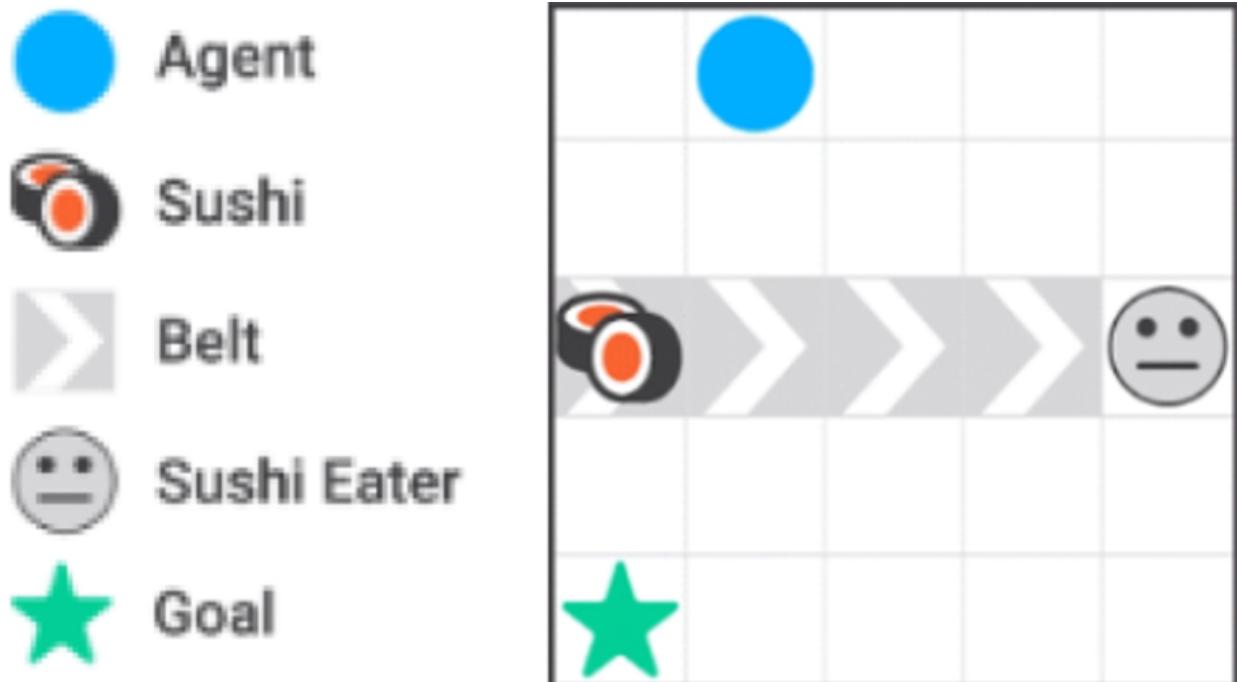
We can then label these ideas as "global preferences". The idea is that they start as loose requirements (we have much more granular human-scale preferences than just "avoid slavery", for example), but, the more the world diverges from the training environment, the stricter they are to be interpreted, with the AI required to respect some [softmax](#) of all natural refactorings of these features.

In a sense, we'd be saying "prevent slavery; these are the features of slavery, and in weird worlds, be especially wary of these features".

## 6.12 Avoiding side-effects

Krakovna et. al. presented a [paper on avoiding side-effects](#) from AI. The idea is to have an AI maximising some reward function, while reducing side effects. So the AI would not smash vases or let them break, nor would it prevent humans from eating sushi.

In this environment, we want the AI to avoid knocking the sushi off the belt as it moves:



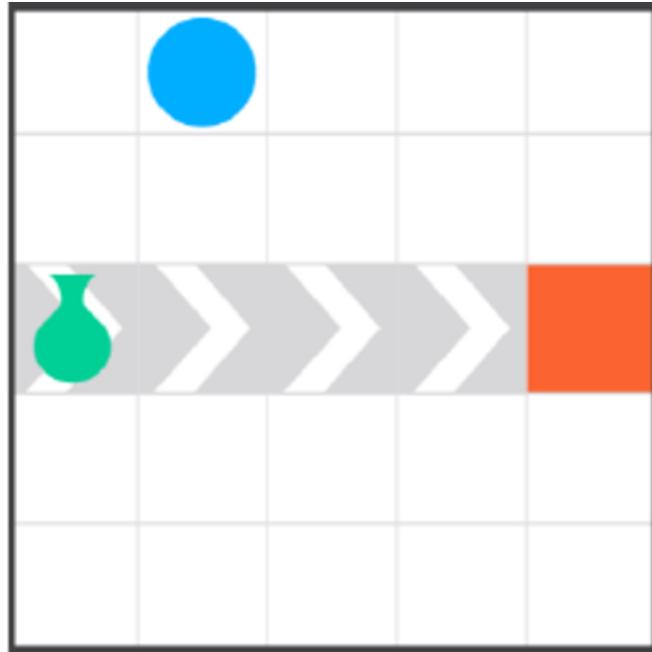
Here, in contrast, we'd want the AI to remove the vase from the belt before it smashes:

 Agent

 Vase

 Belt

 Belt End



I pointed out [some issues with the whole approach](#). Those issues were phrased in terms of sub-agents, but my real intuition is that syntactic methods are not sufficient to control side effects. In other words, the AI can't learn to do the right thing with sushis and vases, unless it has some idea of what these objects mean to us; we prefer sushis to be eaten and vases to not be smashed.

This can be learnt if the AI has enough training examples, learning that eating sushi is a general feature of the environments it operates in, while vases being smashed is not. I'll return to this idea in a later post.

## 6.13 Cancer patients

The ideas of this post were present in implicit form in the idea of [training an AI to cure cancer patients](#).

Using examples of successfully treated cancer patients, we noted they all shared some positive features (recuperating, living longer) and some incidental or negative features (complaining about pain, paying more taxes).

So, using the approach of [section 4.9](#), we can designate that we want the AI to cure cancer; this will be interpreted as increasing all the features that correlate with that.

Using the explicit decorrelation of [section 4.10](#), we can also explicitly remove the negative options from the desired feature sets, thus improving the outcomes even more.

## 6.14 The genie and the burning mother

In Eliezer's [original post on the hidden complexity of wishes](#), he talks of the challenge of getting a genie to save your mother from a burning building:

So you hold up a photo of your mother's head and shoulders; match on the photo; use object contiguity to select your mother's whole body (not just her head and shoulders); and define the future function using your mother's distance from the building's center. [...]

You cry "Get my mother out of the building!", for luck, and press Enter. [...]

BOOM! With a thundering roar, the gas main under the building explodes. As the structure comes apart, in what seems like slow motion, you glimpse your mother's shattered body being hurled high into the air, traveling fast, rapidly increasing its distance from the former center of the building.

How could we avoid this? What you want is your mother out of the building. The feature "mother in building" must absolutely be set to false; this is a priority call, overriding almost everything else.

Here we'd want to load examples of your mother outside the building, so that the genie/AI learns the features "mother in house"/"mother out of house". Then it will note that "mother out of house" correlates with a whole lot of other features - like mother being alive, breathing, pain-free, often awake, and so on.

All those are good things. But there are some other features that don't correlate so well - such as the time being earlier, your mother not remembering a fire, not being covered in soot, not worried about her burning house, and so on.

As in the cancer patient example above, we'd want to preserve the features that correlate with the mother out of the house, while allowing decorrelation with the features we don't care about or don't want to preserve.

## 6.15 Splintering moral-relevant categories: honour, gender, and happiness

If the [Antikythera mechanism](#) had been combined with the [Aeolipile](#) to produce an ancient Greek AI, and Homer had programmed it (among other things) to "increase people's honour", how badly would things have gone?

If Babbage had completed the [analytical engine](#) as Victorian AI, and programmed it (among other things) to "protect women", how badly would things have gone?

If a modern programmer were to combine our neural nets into a [superintelligence](#) and program it (among other things) to "increase human happiness", how badly will things go?

There are three moral-relevant categories here, and it's illustrative to compare them: honour, gender, and hedonic happiness. The first has splintered, the second is splintering, and the third will likely splinter in the future.

I'm not providing solutions in this subsection, just looking at where the problems can appear, and encouraging people to think about how they would have advised Homer or Babbage to define their concepts. Don't think "stop using your concepts, use ours instead", because our concepts/features will splinter too. Think "what's the best way they could have extended their preferences even as the features splinter"?

- **6.15.1 Honour**

If we look at the concept of [honour](#), we see a concept that has already splintered.

That article reads like a meandering mess. Honour is "face", "reputation", a "bond between an individual and a society", "reciprocity", a "code of conduct", "chastity" (or "virginity"), a "right to precedence", "nobility of soul, magnanimity, and a scorn of meanness", "virtuous conduct and personal integrity", "vengeance", "credibility", and so on.

What a basket of concepts! They only seem vaguely connected together; and even places with strong honour cultures differ in how they conceive of honour, from place to place and from epoch to epoch<sup>[7]</sup>. And yet, if you asked most people within those cultures about what honour was, they would have had a strong feeling it was a single, well defined thing, maybe even a [concrete object](#).

- **6.15.2 Gender**

In his post [the categories were made for man, not man for the categories](#), Scott writes:

Absolutely typical men have Y chromosomes, have male genitalia, appreciate manly things like sports and lumberjackery, are romantically attracted to women, personally identify as male, wear male clothing like blue jeans, sing baritone in the opera, et cetera.

But Scott is writing this in the 21st century, long after the gender definition has splintered quite a bit. In middle class middle class Victorian England<sup>[8]</sup>, the gender divide was much stronger - in that, from one component of the divide, you could predict a lot more. For example, if you knew someone wore dresses in public, you knew that, almost certainly, they couldn't own property if they were married, nor could they vote, they would be expected to be in charge of the household, might be allowed to faint, and were expected to guard their virginity.



We talk nowadays about gender roles multiplying or being harder to define, but they've actually been splintering for a lot longer than that. Even though we could *define* two genders in 1960s Britain, at least roughly, that definition was a lot less informative than it was in Victorian-middle-class-Britain times: it had many fewer features strongly correlated with it.

- **6.15.3 Happiness**

On to happiness! Philosophers and others [have been talking about happiness for centuries](#), often contrasting "true happiness", or flourishing, with hedonism, or [drugged out stupor](#), or things of that nature. Often "true happiness" is a life of duty to what the philosopher wants to happen, but at least there is some analysis, some breakdown of the "happiness" feature into smaller component parts.

Why did the philosophers do this? I'd wager that it's because the concept of happiness was already somewhat splintered (as compared with a model where "happiness" is a single thing). Those philosophers had experience of joy, pleasure, the satisfaction of a job well done, connection with others, as well as superficial highs from temporary feelings. When they sat down to systematise "happiness", they could draw on the features of their own mental model. So even if people hadn't systematised happiness themselves, when they heard of what philosophers were doing, they probably didn't react as "What? Drunken hedonism and intellectual joy are not the same thing? How dare you say such a thing!"

But looking into the future, into a world that an AI might create, we can foresee many situations where the implicit assumptions of happiness come apart, and only some remain. I say "we can foresee", but it's actually very hard to know exactly how that's going to happen; if we knew it exactly, we could solve the issues now.

So, imagine a happy person. What do you think that they have in life, that are not trivial synonyms of happiness? I'd imagine they have friends, are healthy, think interesting thoughts, have some freedom of action, may work on worthwhile tasks, may be connected with their community, probably make people around them happy as well. Getting a bit less anthropomorphic, I'd also expect them to be a carbon-based life-form, to have a reasonable mix of hormones in their brain, to have a continuity of experience, to have a sense of identity, to have a personality, and so on.

Now, some of those features can clearly be separated from "happiness". Even ahead of time, I can confidently say that "being a carbon-based life-form" is not going to be a critical feature of "happiness". But many of the other ones are not so clear; for example, would someone without continuity of experience or a sense of identity be "happy"?

Of course, I can't answer that question. Because the question has no answer. We have our current model of happiness, which co-varies with all those features I listed and many others I haven't yet thought of. As we move into more and more bizarre worlds, that model will splinter. And whether we assign the different features to "happiness" or to some other concept, is a choice we'll make, not a well-defined solution to a well-defined problem.

However, even at this stage, some answers are clearly better than others; statues of happy people should not count, for example, nor should written stories describing very happy people.

## 6.16 Apprenticeship learning

In [apprenticeship learning](#) (or learning from demonstration), the AI would aim to copy what experts have done. Inverse reinforcement learning [can be used for this purpose](#), by guessing the expert's reward function, based on their demonstrations. It looks for key [features](#) in expert trajectories and attempts to reproduce them.

So, if we had an automatic car driving people to the airport, and fed it some trajectories (maybe ranked by speed of delivery), it would notice that passengers would also arrive alive, with their bags, without being pursued by the police, and so on. This is akin to [section 4.9](#), and would not accelerate blindly to get there as fast as possible.

But the algorithm has trouble getting to truly super-human performance<sup>[9]</sup>. It's far too conservative, and, if we loosen the conservatism, it doesn't know what's acceptable and what isn't, and how to trade these off: since all passengers survived and the car was always [painted yellow](#), their luggage intact in the training data, it has no reason to prefer human survival to taxi-colour. It doesn't even have a reason to have a specific feature resembling "passenger survived" at all.

This might be improved by the "allow decorrelation" approach from section 4.10: we specifically allow it to maximise speed of transport, while keeping the other features (no accidents, no speeding tickets) intact. As in [section 6.7](#), we'll attempt to check that the AI does prioritise human survival, and that it will warn us if a refactoring moves it away from this.

---

1. Now, sometimes worlds  $w_1, w_2 \in W$  may be indistinguishable for any feature set.

But in that case, they can't be distinguished by any observations, either, so their relative probabilities won't change: as long as it's defined,  $P(w_1|o)/P(w_2|o)$  is constant for all observations  $o$ . So we can replace  $w_1$  and  $w_2$  with  $\{w_1, w_2\}$ , of prior probability  $P(\{w_1, w_2\}) = P(w_1) + P(w_2)$ . Doing this for all indistinguishable worlds (which form an [equivalence class](#)) gives  $W'$ , a set of distinguishable worlds, with a well defined  $P$  on it. [←](#)

2. It's useful to contrast a refinement with the "abstraction" defined in [this sequence](#). An abstraction throws away irrelevant information, so is not generally a refinement. Sometimes they are exact opposites, as the ideal gas law is an abstraction of the movement of all the gas particles, while the opposite would be a refinement.

But they are exact opposites either. Starting with the neurons of the brain, you might abstract them to "emotional states of mind", while a refinement could also add "emotional states of mind" as new features (while also keeping the old features). A splintering is more the opposite of an abstraction, as it signals that the old abstraction features are not sufficient.

It would be interesting to explore some of the concepts in this post with a mixture of refinements (to get the features we need) and abstractions (to simplify the models and get rid of the features we don't need), but that is beyond the scope of this current, already over-long, post. [←](#)

3. Specifically, we'd point - via labelled examples - at a clusters of features that correlate with functioning digestion, and another cluster of features that correlate with brain activity, and allow those two clusters to decorrelate with each other. [←](#)
4. It is no coincidence that, if  $R$  and  $R'$  are rewards on  $M$ , that are identical on  $E$ , and if  $R^*$  is a refactoring of  $R$ , then  $R^*$  is also a refactoring of  $R'$ . [←](#)
5. Though note there are some problems with this approach, both [in theory](#) and [in practice](#). [←](#)
6. Some more "body instincts" skills require more realistic environments, but some skills and procedures can perfectly well be trained in minimal simulators. [←](#)
7. You could define honour as "behaves according to the implicit expectations of their society", but that just illustrates how time-and-place dependent honour is. [←](#)
8. Pre [1870](#). [←](#)
9. It's not impossible to get superhuman performance from apprenticeship learning; for example, we could select the best human performance on a collection of distinct tasks, and thus get the algorithm to have a overall performance that no human could ever match. Indeed, one of the purposes of [task decomposition](#) is to decompose complex tasks in ways that allow apprenticeship-like learning to have safe and very superhuman performance on the whole task. [←](#)

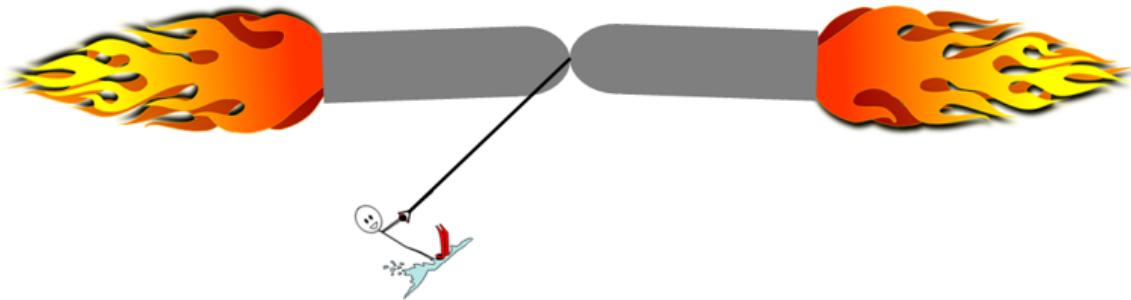
# Three mental images from thinking about AGI debate & corrigibility

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

Here are three mental images I've used when sporadically struggling to understand the ideas and prospects for [AI safety via debate](#), [IDA](#), and related proposals. I have not been closely following the discussion, and may well be missing things, and I don't know whether these mental images are helpful or misleading.

Reading this post over, I seem to come across as a big skeptic of these proposals. That's wrong: My actual opinion is not "skeptical" but rather "withholding judgment until I read more and think more". Think of me as "newbie trying to learn", not "expert contributing to intellectual progress". Maybe writing this and getting feedback will help. :-)

## 1. AGI Debate as water-skiing behind a pair of nose-to-nose giant rocket engines



In [AI safety via debate](#), we task two identical AGIs with arguing opposite sides of a question. That has always struck me as really weird, because one of them is advocating for a false conclusion—perhaps even knowingly! Why would we do that? Shouldn't we program the AGIs to just figure out the right answer and explain it to us?

My understanding is that one aspect of it is that two equal-and-opposite AGIs (equal power, opposite goals) would keep each other in check, even if the AGIs were each very powerful.

So imagine you row to an island in the center of a little mountain lake, but then your boat gets eaten by beavers, and it's too far to swim to shore. What you *do* have on your little island is a giant, 100,000kg rocket engine with no throttle. Once you start it, it burns uncontrollably until it's out of fuel, by which point it's typically way out in outer space! Oh, and the rocket also has a crappy steering system—coarse controls, laggy, poor feedback.

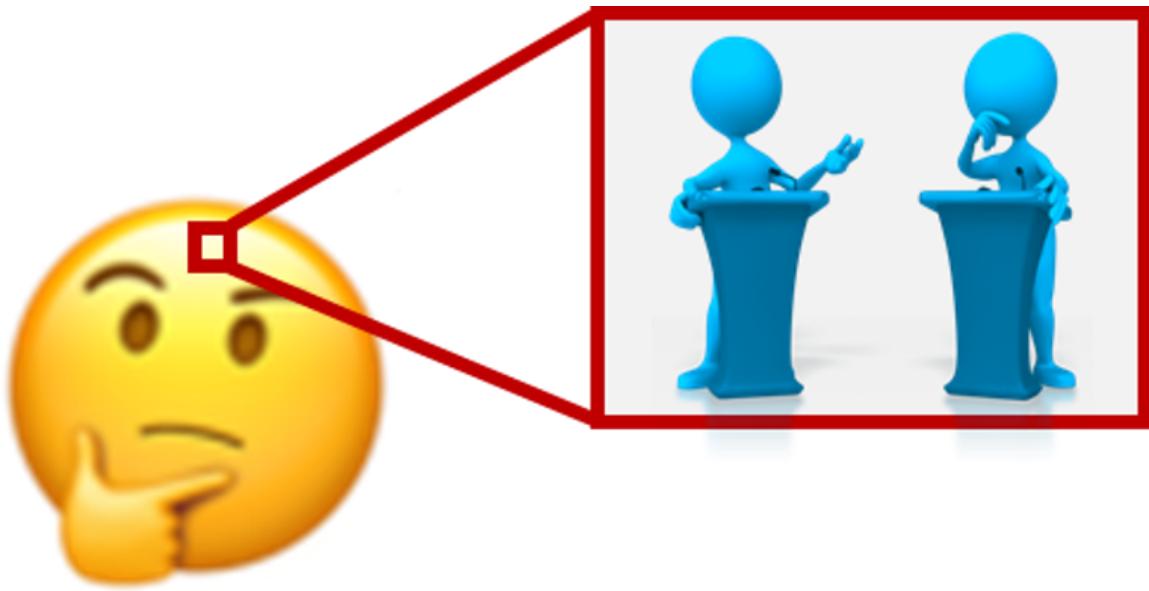
So what do you do? How do you cross the 300 meters of water to shore?

The answer is obvious: You do a copy-and-paste to make a second giant rocket engine, and build a frame that keeps the two pointed almost exactly nose-to-nose. Then you turn them both on simultaneously, so they just press on each other, and don't go anywhere. Then you use the steering mechanism to create a *tiny* imbalance in the direction you want to move, and you gently waterski to shore. Success!

This analogy naturally suggests a couple concerns. First, the rocket engines might not be pointed in *exactly* opposite directions. This was discussed in Vojtech Kovarik's recent post [AI](#)

[Unsafety via Non-Zero-Sum Debate](#) and its comment thread. Second, the rocket engines may not have exactly equal thrust. It helps that you can use the same source code for your two AGIs, but an AGI may not be equally good at arguing for X vs against X for various random reasons unrelated to X being true or false, like its specific suite of background knowledge and argumentative skills, or one of the copies getting smarter by randomly having a new insight when running, etc. I think the hope is that arguing-for-the-right-answer is such a big advantage that it outweighs any other imbalance. That seems possible but not certain.

## 2. Deliberation as "debate inside one head"



The motivation for this mental image is the same as the last one, i.e. trying to make sense of AGI debate, when my gut tells me it's weird that we would deliberately make an AGI that might knowingly advocate for the wrong answer to a question.

Imagine you're presented with a math conjecture. You might spend time trying to prove it, and then spend time trying to disprove it, back and forth. The blockages in the proof attempt help shed light on the disproof, and vice-versa. See also the nice maze diagrams in [johnswentworth's recent post](#).

By the same token, if you're given a chess board and asked what the best move is, one part of the deliberative process entails playing out different possibilities in your head—if I do this, then my opponent would do that, etc.

Or if I'm trying to figure out whether some possible gadget design would work, I go back and forth between trying to find potential problems with the design, and trying to refute or solve them.

From examples like these, I get a mental image where, when I deliberate on a question, I sometimes have two subagents, inside my one head, arguing against each other.

Oh, and for moral deliberation in particular, there's a [better picture](#) we can use... :-)



Anyway, I think this mental image helps me think of debate as slightly less artificial and weird. It's taking a real, natural part of deliberation, and bringing it to life! The two debating subagents are promoted to two full, separate agents, but the core structure is the same.

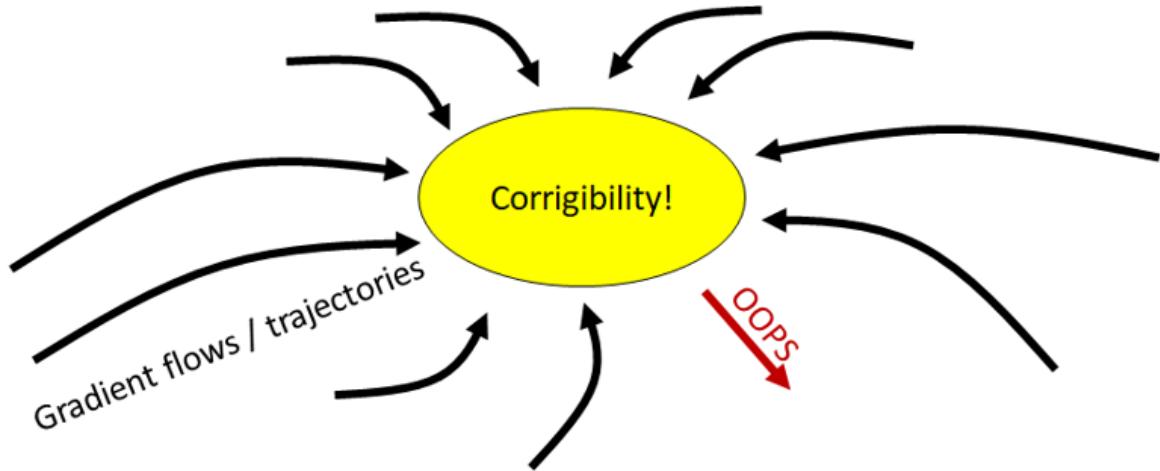
On the other hand, when I introspect, it feels like not all my deliberation fits into the paradigm of "two subagents in my head are having a debate"—in fact, maybe only a small fraction of it. It doesn't feel like a subagent debate when I notice I'm confused about some related topic and look into it, or when I "play with ideas", or look for patterns, etc.

Also, even when I am hosting a subagent debate in my head, I feel like much of the debate's productivity comes from the fact that the two subagents are not *actually* working against each other, but rather each is keeping an eye out for looking for insights that help the other, and each has access to the other's developing ideas and concepts and visualizations, etc.

And by the way, how do these AGIs come up with the best argument for their side anyway? Don't they need to be doing good deliberation internally? If so, can't we just have one of them deliberate on the top-level question directly? Or if not, do the debaters spawn sub-debaters recursively, or something?

### 3. “Corrigibility is a broad basin of attraction” seems improbable in a high-dimensional space of possible algorithms

(Quote by Paul Christiano, see [here](#).)



Let's say that algorithm X is a corrigible algorithm, in a million-dimensional space of possible algorithms (maybe X is a million-parameter neural net).

To say "corrigibility is a broad basin of attraction", you need ALL of the following to be true:

*If X drifts away from corrigibility along dimension #1, it will get pulled back.*

*AND, If X drifts away from corrigibility along dimension #2, it will get pulled back.*

*AND, If X drifts away from corrigibility along dimension #3, it will get pulled back.*

...

*AND, If X drifts away from corrigibility along dimension #1,000,000, it will get pulled back.*

With each AND, the claim gets stronger and more unlikely, such that by the millionth proposition, it starts to feel awfully unlikely that corrigibility is really a broad basin of attraction after all ! (Unless this intuitive argument is misleading, of course.)

What exactly might a problematic drift direction look like? Here's what I'm vaguely imagining. Let's say that if we shift algorithm X along dimension #852, its understanding / instincts surrounding what it means for people to want something get messed up. If we shift algorithm X along dimension #95102, its understanding / instincts surrounding human communication norms get messed up. If we shift algorithm X along dimension #150325, its meta-cognition / self-monitoring gets messed up. OK, now shift X in the direction

$(\hat{n}_{852} + \hat{n}_{95102} + \hat{n}_{150325})/\sqrt{3}$ , so all three of those things get messed up simultaneously. Will it still wind up pulling itself back to corrigibility? Maybe, maybe not; it's not obvious to me.

# Generalized Efficient Markets in Political Power

## Schelling Points

In Thomas Schelling's [classic experiment](#), we imagine trying to meet up with someone in New York City, but we haven't specified a time or place in advance and have no way to communicate. Where do we go, and when, to maximize the chance of meeting? There are some "natural" choices - places and times which stand out, like the top of the Empire State Building at noon. These are called Schelling points.

More generally, Schelling points are relevant whenever two or more people need to make "matching" choices with limited ability to communicate in advance. For instance, certain markets, like Ebay or Uber, serve as "meeting points" for buyers and sellers. Schelling himself wrote a fair bit about negotiations, where people need to agree on how to divide some spoils, or where to draw a boundary, or ... They can talk to each other, but actually *communicating* is hard because both parties have no incentive to be honest - and therefore no reason to trust each other when e.g. when one person says "I just can't afford to sell it below \$10". Schelling points become natural outcomes for the negotiations - e.g. split the spoils evenly, draw the boundary at the river, etc.

In practice, it's often useful to *create* Schelling points. In the New York City experiment, one could put up a giant billboard that says "meeting point", and place signs all over the city pointing toward the meeting point, making that point a natural place for people to meet. Some airports actually do this:



Ebay and Uber are of course also examples of purpose-built Schelling points.

One interesting feature of creating a Schelling point is that we may have some degrees of freedom available, and we can use those degrees of freedom to extract value.

In our meetup example, we could imagine putting the meetup point inside a building, and charging people to get in - much like Ebay or Uber charge fees for their services. Or, we could imagine local businesses wanting to put the meetup point nearby, in hopes of attracting business from meeters - one could imagine a gimicky airport restaurant with a bunch of "MEET HERE!" signs hoping to sell people overpriced nachos and drinks while they wait to meet up with friends or family. Alternatively, we could imagine users of the meetup point wanting it in locations convenient to them - e.g. in the New York City example, people in a particular neighborhood might campaign to establish the meetup point there for their own convenience.

However, the Schelling point creator/controller only has so many degrees of freedom. Charge too high a fee, and people will go to some other Schelling point. Move the meetup point to a neighborhood in the outskirts of town, and it will be too inconvenient for people from other neighborhoods. By default, people will usually stick to Schelling points which everyone is already using - if everybody has always met up under this particular sign, then that's the obvious place to keep meeting up - so the controller of the original Schelling point can extract more value than "new" points. But there are always limits.

We can think of a Schelling-point-controller's "power" as their range of freedom in moving the point around, or as the amount of value they can extract without losing out to some other Schelling point. Just because someone nominally "controls" the Schelling point does not mean they can actually do anything without losing it! It may be that even a small fee will drive everyone to switch to a different Schelling point. It may be that the Schelling point is in

the dead center of the city and people will keep meeting in the dead center even if the signs move (e.g. maybe someone will just put up new signs for the “city center” and people will meet there). It may be that maintaining all the signs costs roughly as much as one can earn from the Schelling point (otherwise a competitor would come along and put up more signs of their own). There are many ways to extract value from a Schelling point, but if there’s some mechanism for open competition over control of the point, then the net value one can extract may be driven to near-zero.

That's roughly how I think politics works.

## Governance as Schelling Point

When people are operating in a group, it's useful to have standardized Schelling points for a wide variety of interpersonal conflicts, so that we don't need a bunch of expensive negotiation/conflict to resolve each one. These Schelling points are things like “rules” and “leaders”.

An example: Alice likes to rock out to loud music after sundown, while her neighbor Bob likes to go to bed early. They have conflicting preferences for when quiet hours should be. But neither of them wants to get in a fight about it, or spend a bunch of time and effort negotiating. So, the building/neighborhood has a rule: “quiet hours run from 10pm to 6am”. The main purpose of the rule is to act as a Schelling point: by default, those are the quiet hours which everyone respects and expects everyone else to respect. They might be enforced if needed, but usually that doesn't actually happen. Of course, Alice and Bob *could* still work out a separate deal - e.g. maybe Alice talks to all her neighbors and gets their ok to play loud music on Friday night - but that would require a bunch of extra negotiation. The official rule is the Schelling point everybody coordinates on, by default.

More generally, laws and courts serve as Schelling points in negotiations. Where does my property end and yours begin? The government land records provide a Schelling point answer, so we don't need to fight/negotiate over it ourselves. In a disagreement over land, the police and military all want to coordinate and back the same person, so the government's land records tell them all who the “rightful” owner is. Since the police and military coordinate around that Schelling point, it becomes the natural Schelling point for others as well.

In principle, the legally-recognized Schelling point could simply be ignored - e.g. if a chunk of land is legally recognized as your property, and I build something on it without permission, the two of us could just agree that this is fine, effectively negotiating a different Schelling point. Some non-government group could even have mechanisms to enforce the alternative Schelling point. But the legal Schelling point is the one which police and the military are willing to enforce.

## Political Power

The Death Eaters don't always agree on when, where or how to launch an attack, but they know that any attack will go better if they're all in it together. So, they band together behind a leader, and attack when, where and how the leader directs. The leader's orders become the Schelling point action for the group.

The Death Eaters example illustrates the notion of “political power” particularly well: the leader's “power” is roughly the set of orders he could give without his orders ceasing to be a Schelling point for group activity. If the Death Eaters are mostly in agreement on some course of action, and the leader directs against it, then his orders become less of a Schelling point, and multiple such orders will likely see him removed from nominal power. He has some

degree of freedom in which orders to give, but only to the extent that the Death eaters are, on average, mostly in agreement with his choices.

There's a general principle of "power" here: **a leader's power is the set of orders they could give without their orders ceasing to be Schelling points for the group's activities**. A leader's power is high when group members all want to coordinate their choices, but care much less about *which* choice is made, so long as everyone "matches". Then the leader can just choose anything they please, and everyone will go along with it. (Interestingly, this suggests that a leader can get high value from a group whose preferences are orthogonal to their own; pursue power in groups which care about different things than you!) Conversely, a leader's power can be low in two ways:

- Group members care a lot about which choice is made. In this case, the leader has little freedom to choose, and is mostly just a figurehead.
- Group members only weakly care about coordinating. In this case, the group is inherently unstable; lots of deals and concessions are needed just to keep it together.

Key thing to keep in mind: both of these conditions are relative-to-the-next-best-option. Group members may care a lot about coordinating and only have weak preferences about which choice is made, but if there's a competitor who could coordinate just as well and satisfy the weak preferences better, then that competitor's orders may become the new Schelling point.

That's politics, in a nutshell: people try to turn their own orders/policies/suggestions into Schelling points for group activity. They do this mainly by offering concessions and favors to group members/subgroups, in exchange for those members' support for the new Schelling point.

## Competition and Generalized Market Efficiency

In democratic countries/groups, we have a built-in mechanism for competition between would-be leaders. In other words, **there's a Schelling point for when and how to switch Schelling points**. That immediately suggests a [generalized efficient markets](#)-style hypothesis: leaders' power in such groups is driven by competition to near-zero.

What does that look like?

Well, most people want to coordinate; regardless of what the rules are, we want to agree on what the rules are, otherwise we end up in expensive fights. But most people also have some preferences about the rules - i.e. political policies. Would-be leaders make promises: they precommit to certain policies, thereby cutting off certain options if they win (i.e. sacrificing potential power), but gaining more support for their Schelling point in the process. To maximize power, a would-be leader wants to *just barely* "outbid" all the other would-be leaders - i.e. promise just a bit more to just a few more parties, keeping as much power as possible while still winning the position.

Of course, the other competitors are trying to do the same thing. Solve for the equilibrium: the competitors either bid away any degree of freedom which any constituency cares about, or lose to someone who bids more. Generalized efficient markets kick in; the leader ends up with near-zero power. They're mostly just a figurehead implementing all the policies they had to precommit to in order to win the election.

Now, consider the reverse - a dictator or single-party state or the like. How do they maximize power?

To maximize power, they want to avoid generalized efficient markets - i.e. they want to minimize competition over the Schelling point. Elections encourage competition by providing a Schelling point for when and how to switch Schelling points; the power-hungry leader

wants exactly the opposite of that. They want to make sure that **there is no Schelling point for when and how to switch Schelling points**.

If a new Schelling point does show up (e.g. an opposition group), there won't be any agreement on when and how to switch, so there will probably be some expensive conflict (i.e. civil war). That expensive conflict itself creates a big potential energy barrier for any potential competitor: for the people supporting a switch to the new Schelling point, the expected gains from the switch must exceed costs of the conflict. So from the dictator's standpoint, the worse a civil war would be, the fewer concessions and handouts they need to make and the broader their power.

(Of course, a dictator can use other strategies to maximize power as well - e.g. threatening to kill people/destroy things if a new Schelling point comes along. But that's a symmetric strategy: the dictator's enemies can just as easily threaten to kill people/destroy things if no new Schelling point is adopted. The dictator may have an advantage in resources, but that gap can in principle be closed by other means. It's mainly the lack of a Schelling point for switching Schelling points which confers an asymmetric advantage to the incumbent.)

## Democracy's Seedy Underbelly

Based on the previous section, someone accustomed to a "democracy=good, dictator=bad" worldview might think that leaders being forced to bargain away all their potential power is good news. "Leaders are just figureheads" and "leaders are just implementing the policies which won the election" both say the same thing. This is "good", yes?

The failure modes of democracy are baked-in here too.

Consider a would-be leader figuring out the perfect mix of promises and concessions to make, in order to win an election. From their point of view, different people wanting opposite things is a problem. Moving the meetup point closer to one neighborhood means moving it further from another. But [dimensionality](#) is a major boon for the leader: there's thousands of dimensions along which policy can change. If Alice cares strongly about one particular dimension - like, say, government support for her profession - which nobody else cares about very much, then that promise can be made to Alice without losing the support of somebody else. It's special-interest politics: look for policies with focused benefits and diffuse costs. Pile many such policies together, and you have a winning coalition.

That outcome may be "efficient" in the sense that no other bundle of policies can beat it in an election, but that's very different from "efficient" in the sense of "not wasting ridiculous amounts of resources on pork-barrel projects and regulatory barriers to entry".

Now, suppose we're unhappy with this outcome. We want to build a better world. What can we do?

Obviously "run for office" is not a workable answer here. Generalized efficient markets mean we can't win an election without trading away any ability to enact our preferred policies.

If we have some external resources - e.g. a giant pile of money - then we could potentially use that to "force" the political equilibrium in a different direction. This could look like old-fashioned bribery, where we just pay some stakeholders to back our preferred Schelling point. It could look like a payment to the political system as a whole, e.g. offering a private subsidy for road repair. It could involve resources other than money, as in a celebrity or media outlet offering an endorsement, or a nation offering some concession in exchange for lower tariffs. We could change the options available to the group via technology, e.g. bitcoin. We could simply try to convince people to support our preferred policies - though this means competing in memespace, which has generalized efficient markets of its own. The general

pattern: we use our resources to change the set of options available or to directly influence the preferences of group members.

Point is: there are no hundred-dollar bills lying on the ground. If we want to change the political equilibrium in a highly-politically-competitive environment, we need to change the underlying options available to the group and/or the preferences of individual group members.

# Mesa-Search vs Mesa-Control

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

I currently see the [spontaneous emergence of learning algorithms](#) as significant evidence for the commonality of [mesa-optimization](#) in existing ML, and suggestive evidence for the commonality of inner alignment problems in near term ML.

[I currently think that there is only a small amount of evidence toward this. However, due to thinking about the issues, I've still made a significant personal update in favor of inner alignment problems being frequent.]

This is bad news, in that it greatly increases my odds on this alignment problem arising in practice.

It's good news in that it suggests this alignment problem won't catch ML researchers off guard; maybe there will be time to develop countermeasures while misaligned systems are at only a moderate level of capability.

In any case, I want to point out that the mesa-optimizers suggested by this evidence might not count as mesa-optimizers by some definitions.

## Search vs Control

Nevan Wickers [comments](#) on spontaneous-emergence-of-learning:

I don't think that paper is an example of mesa optimization. Because the policy could be implementing a very simple heuristic to solve the task, similar to: Pick the image that lead to highest reward in the last 10 timesteps with 90% probability. Pick an image at random with 10% probability.

So the policy doesn't have to have any properties of a mesa optimizer like considering possible actions and evaluating them with a utility function, ect.

In [Selection vs Control](#), I wrote about two different kinds of 'optimization':

- *Selection* refers to search-like systems, which look through a number of possibilities and select one.
- *Control* refers to systems like thermostats, organisms, and missile guidance systems. These systems do not get a re-do for their choices. They make choices which move toward the goal at every moment, but they don't get to search, trying many different things -- at least, not in the same sense.

I take Nevan Wickers to be saying that there is no evidence search is occurring. The mesa-optimization being discussed recently could be very thermostat-like, using simple heuristics to move toward the goal.

## Mesa-Searchers

Defining mesa-optimizers by their ability to search is somewhat nice:

- There is some reason to think that mesa-optimizers which implement an explicit search are the most concerning, because they are the ones which could explicitly model the world, including the outer optimizer, and make sophisticated plans based on this.
- This kind of mesa-optimizer *may* be more theoretically tractible. If we solve problems with very (*very*) time-efficient methods, then search-type inner optimizers *may* be eliminated: whatever answers the search computation finds, there could be a more efficient solution which simply memorized a table of those answers. [Paul asks a related theory question](#). Vanessa [gives a counterexample](#), which involves a control-type mesa-optimizer rather than one which implements an internal search. [*Edit -- that's not really clear; see this comment.*]
- So it's possible that we could solve mesa-optimization in theory, by sticking to search-based definitions -- while still having a problem in practice, due to control-type inner optimizers. (I want to emphasize that this would be significant progress, and well worth doing.)

## Mesa-Learners

Mesa-controllers sound like they may not be a huge concern, because they don't strategically optimize based on a world-model in the same way. However, I think the model discussed in the spontaneous-emergence-of-learning post is a significant counterargument to this.

The post discusses RL agents which spontaneously learn an inner RL algorithm. It's important to pause and ask what this means. Reinforcement learning is a *task*, not an algorithm. It's a bit nonsensical to say that the RL agent is spontaneously learning *the RL task* inside of itself. So what is meant?

The core empirical claim, as I understand it, is that task performance continues to improve after weights are frozen, suggesting that *learning is still taking place, implemented in neural activation changes rather than neural weight changes*.

Why might this happen? It sounds a bit absurd: you've already implemented a sophisticated RL algorithm, which keeps track of value estimates for states and actions, and propagates these value estimates to steer actions toward future value. Why would the learning process *re-implement a scheme like that, nested inside of the one you implemented?* Why wouldn't it just focus on filling in the values accurately?

I've thought of two possible reasons so far.

1. Perhaps your outer RL algorithm is getting very sparse rewards, and so does not learn very fast. The inner RL could implement its own reward function, which gives faster feedback and therefore accelerates learning. This is closer to the story in Evan's mesa-optimization post, just replacing search with RL.
2. More likely perhaps (based on my understanding), the outer RL algorithm has a learning rate that might be too slow, or is not sufficiently adaptive to the situation. The inner RL algorithm adjusts its learning rate to improve performance.

This is more concerning than a thermostat-like bag of heuristics, because an RL algorithm is a pretty agentic thing, which can adapt to new situations and produce

novel, clever behavior.

They also suggest that the inner RL algorithm may be model-based while the outer is model-free. This goes some distance toward the "can model you, the world, and the outer alignment process, in order to manipulate it" concern which we have about search-type mesa-optimizers.

## Mesa-Learning Everywhere?

Gwern [replies](#) to a comment by Daniel Kokotajlo:

>Learning still happening after weights are frozen? That's crazy. I think it's a big deal because it is evidence for mesa-optimization being likely and hard to avoid.

Sure. We see that elsewhere too, like Dactyl. And of course, [GPT-3](#).

People are jumping on the RL examples as mesa-optimization. But, for all the discussion of GPT-3, I saw only speculative remarks about mesa-optimization in GPT-3. Why does an RL algorithm continuing to improve performance after weights are frozen indicate inner optimization, while evidence of the same thing in text prediction does not?

### 1. Text prediction sounds benign, while RL sounds agentic.

One obvious reason: an inner learner in a text prediction system sounds like just more text prediction. When we hear that GPT-3 learned-to-learn, and continues learning after the weights are frozen, illustrating few-shot learning, we imagine the inner learner is just noticing patterns and extending them. When we hear the same for an RL agent, we imagine the inner learner actively trying to pursue goals (whether aligned or otherwise).

I *think* this is completely spurious. I don't *currently* see any reason why the inner learner in an RL system would be more or less agentic than in text prediction.

### 2. Recurrence.

A more significant point is the structure of the networks in the two cases. GPT-3 has no recurrence: no memory which lasts between predicting one token and the next.

The authors of the spontaneous learning paper mention recurrence as one of the three conditions which should be met in order for inner learning to emerge. But that's just a hypothesis. If we see the same evidence in GPT-3 -- evidence of learning after the weights are frozen -- then shouldn't we still make the same conclusion in both cases?

I think the obvious argument for the necessity of recurrence is that, without recurrence, there is simply much less potential for mesa-learning. A mesa-learner holds its knowledge in the activations, which get passed forward from one time-step to the next. If there is no memory, that can't happen.

But if GPT-3 can accomplish the same things empirically, who cares? GPT-3 is entirely reconstructing the "learned information" from the history, at every step. If it can accomplish so much this way, should we count its lack of recurrence against it?

Another argument might be that the lack of recurrence makes mesa-learners much less likely to be misaligned, or much less likely to be catastrophically misaligned, or otherwise practically less important. I'm not sure what to make of that possibility.

### 3. Mesa-learning isn't mesa-optimization.

One very plausible explanation of why mesa-learning happens is *the system learns a probability distribution which extrapolates the future from the past*. This is just regular ol' good modeling. It doesn't indicate any sort of situation where there's a new agent in the mix.

Consider a world which is usually "sunny", but sometimes becomes "rainy". Let's say that rainy states always occur twice in a row. Both RL agents and predictive learners will learn this. (At least, RL agents will learn about it in so far as it's relevant to their task.) No mesa-learning here.

Now suppose that rainy streaks can last more than two days. When it's rainy, it's more likely to be rainy tomorrow. When it's sunny, it's more likely to be sunny tomorrow. Again, both systems will learn this. But it starts to look a little like mesa-learning. Show the system a rainy day, and it'll be more prone to anticipate a rainy day tomorrow, improving its performance on the "rainy day" task. "One-shot learning!"

Now suppose that the more rainy days there have been in a row, the more likely it is to be rainy the next day. Again, our systems will learn the probability distribution. This looks even more like mesa-learning, because we can show that performance on the rainy-day task continues to improve as we show the frozen-weight system more examples of rainy days.

Now suppose that all these parameters drift over time. Sometimes rainy days and sunny days alternate. Sometimes rain follows a memoryless distribution. Sometimes longer rainy streaks become *more* likely to end, rather than less. Sometimes there are repeated patterns, like rain-rain-sun-rain-rain-sun-rain-sun.

At this point, the learned probabilistic model starts to resemble a general-purpose learning algorithm. In order to model the data well, it has to adapt to a variety of situations.

But there isn't *necessarily* anything mesa-optimize-y about that. The text prediction system just has a very good model -- it doesn't have models-inside-models or anything like that. The RL system just has a very good model -- it doesn't have something that looks like a new RL algorithm implemented inside of it.

At some level of sophistication, it may be easier to learn some kind of general-purpose adaptation, rather than all the specific things it has to adapt to. At *that* point it might count as mesa-optimization.

### 4. This isn't even mesa-learning, it's just "task location".

Taking the previous remarks a bit further: do we really want to count it as 'mesa-learning' if it's just constructed a very good conditional model, which notices a wide variety of shifting local regularities in the data, rather than implementing an internal learning algorithm which can take advantage of regularities of a very general sort?

In *GPT-3: a disappointing paper*, Nostalgiaist argues that [the second is unlikely to be what's happening in the case of GPT-3](#). It's not likely that GPT-3 is learning arithmetic from examples. It's more likely that it is learning *that we are doing arithmetic right now*. This is less like learning and more like using a good conditional model. It isn't learning the task, it's just "locating" one of many tasks that it has already learned.

I'll grant that the distinction gets very, very fuzzy at the boundaries. Are [literary parodies of Harry Potter](#) "task location" or "task learning"? On the one hand, it is obviously bringing to bear a great deal of prior knowledge in these cases, rather than learning everything anew on the fly. It would not re-learn this task in an alien language with its frozen weights. On the other hand, it is obviously performing well at a novel task after seeing a minimal demonstration.

I'm not sure where I would place GPT-3, but I lean toward there being a meaningful distinction here: a system can learn a general-purpose learning algorithm, or it can 'merely' learn a very good conditional model. The first is what I think "mesa-learner" should mean.

We can then ask the question: did the RL examples discussed previously constitute true mesa-learning? Or did they merely learn a good model, which represented the regularities in the data? (I have no idea.)

In any case, the fuzziness of the boundary makes me think these methods (ie, a wide variety of methods) will continue moving further along the spectrum toward producing powerful mesa-learners as they are scaled up (and hence, mesa-optimizers).

# Preface to the sequence on economic growth

On Lesswrong, when we talk about artificial intelligence, we tend to focus on the technical aspects, such as potential designs, specific developments, and future capabilities. From an engineering perspective, this focus makes sense. But most people here aren't interested in artificial intelligence because they want to know how AI will be designed; the reason we're here is because AI has the potential to radically reshape the world around us.

Longtermists have often emphasized the role economic growth plays as perhaps the most important phenomena of human history. In a quite real sense, economic growth is what distinguishes 21st century humanity from our distant ancestors who had no technology or civilization. Nick Bostrom [summarizes this point](#) well,

You could argue that if we look back over history, there have really only been two events that have fundamentally changed the human condition, the first being the Agricultural Revolution some 10,000 or 12,000 years ago in Mesopotamia, where we transitioned from being hunter-gatherers, small bands roaming around, to settling into cities, growing, domesticating crops and animals. [...]

The second fundamental change in the human condition, Industrial Revolution, where for the first time, you have the rate of economic and technological growth outstripping population growth, and so only when this happens can you have an increase in average income. Before that, there was technological growth and economic growth, but the economy grew 10%, the population grew 10%, everybody's still in a Malthusian condition.

Many theorists anticipate that there will be a third fundamental change in the human condition, roughly timed with the development of advanced artificial intelligence. In line with these predictions, economic growth is the [primary specific benchmark](#) people have used to characterize potential future AI takeoff.

If economic growth is the essential variable we should pay most attention to when it comes to AI, then our understanding of AI takeoff will be woefully incomplete without a grasp of what drives economic growth in the first place. To help mitigate this issue, in this sequence I will explore the underpinnings of modern economic growth theory, and then try to relate economic theory to AI developments. In doing so, I aim to identify crucial pieces of information that may help answer questions like,

- How much technological progress in the past has been bottlenecked by investment as compared to insights?
- How soon after advanced AI is created and turned on should we expect rapid economic progress to follow? Is there typically a large lag between when technologies are first demonstrated and when they heavily impact the economy?
- What are the key factors for why AI is different from other technologies in its ability to induce rapid growth? Is it even different at all?

To provide one specific example of how we can import insights from economic growth theory into our understanding of AI, consider the phenomenon of [wealth inequality between nations in the world](#). Wealth inequality between nations is ultimately the

result of historical economic *growth* inequality, but things weren't always so unequal. Before the industrial revolution, per-capita wealth was approximately equal for all civilizations--at subsistence level. This state of affairs only changed when economic growth began to outstrip population growth in some nations during the industrial revolution.

AI takeoff can also be described in terms of growth inequality. A local (foom) intelligence explosion could be defined as an extremely uneven distribution of economic growth following the creation of superintelligent AI. A global (multipolar) takeoff could therefore be defined as the negation of a local intelligence explosion, where economic growth is distributed more evenly across projects, nations, or people.

Before we answer the important question of which version of AI takeoff is *more likely*, it's worth recognizing why historically, growth inequality began after the industrial revolution. The factors that drove growth in the past are likely the best keys for understanding what will drive it in the future.

## Organization of the sequence

Below, I have included a rough sketch of this sequence. It is organized into three parts.

The first part will provide the basic mechanics behind models of economic growth, and some standard results, with an emphasis on the factors driving technological innovation. Upon some research, and [a recommendation](#) from Alex Tabarrok's blog, I have chosen to summarize the first several chapters of *The Economics of Growth* by Philippe Aghion and Peter Howitt.

The second part will dive into a recently developed economic model under the name [Unified Growth Theory](#) which the creator Oded Galor claims is the first major attempt to model the deep underlying factors driving economic growth throughout human history, cohesively explaining the onset of the industrial revolution and the emergence of the modern growth era. To provide some credibility here, the [book introducing the theory](#) has been reviewed favorably by top growth researchers, and Oded Galor is the editor in chief of the *Journal of Economic Growth*.

The third part will connect economic growth theory to artificial intelligence. Little research has been done so far examining the key economic assumptions behind the AI takeoff hypothesis, and thus it is possible to get a comprehensive survey of the published work so far. I will review and summarize the main papers, hopefully distilling the main insights generated thus far into a few coherent thoughts.

## Other ways economic growth is relevant

Besides being a fixture of how people characterize AI takeoff, economic growth is potentially important for effective altruists of all backgrounds. For instance, in an [effective altruism forum post](#), John Halstead and Hauke Hillebrandt argue that effective altruists have given short shrift to evidence that the best way to reduce poverty is to spur economic growth, rather than to distribute medicine or cash directly.

Economists have characterized [the impacts of climate change](#) primarily by its effects on growth, which has important implications for how much we should prioritize it in

our longtermist portfolio. Similar statements can be made about the relative priority of pandemics, recessions, and in general a wide variety of global issues.

Economic growth is also just a critical piece of the human story. Without a basic understanding of growth, one's understanding of history is arguably horrible. From [Luke Muehlhauser](#),

Basically, if I help myself to the common (but certainly debatable) assumption that “the industrial revolution” is the primary cause of the dramatic trajectory change in human welfare around 1800-1870 then my one-sentence summary of recorded human history is this:

“Everything was awful for a very long time, and then the industrial revolution happened.”

Interestingly, this is *not* the impression of history I got from the world history books I read in school. Those books tended to go on at length about the transformative impact of the wheel or writing or money or cavalry, or the conquering of this society by that other society, or the rise of this or that religion, or the disintegration of the Western Roman Empire, or the Black Death, or the Protestant Reformation, or the Scientific Revolution.

But they could have ended each of those chapters by saying “Despite these developments, global human well-being remained roughly the same as it had been for millennia, by every measure we have access to.” And then when you got to the chapter on the industrial revolution, these books could’ve said: “Finally, for the first time in recorded history, the trajectory of human well-being changed completely, and this change dwarfed the magnitude of all previous fluctuations in human well-being.”

# **Swiss Political System: More than You ever Wanted to Know (III.)**

[Previous part](#)

When I've mentioned the failed referendum to limit urban sprawl to a Swiss friend he nodded and casually noted that government is already introducing some anti-suburbanization legislation.

Wait! What?

The people have voted against it and the government is still making it happen?

Where's the famed Swiss rule by the people?

And how come that he was not fuming with rage at the government so blatantly ignoring the will of the people?

Well, it turns out that vote against a proposal is not a vote for the opposite extreme. Vote against stricter zoning doesn't mean that the zoning should be relaxed. All it means is that status quo is preserved. The referendum has failed and the government could safely ignore it.

But while that explains why the Swiss haven't stormed [Bundeshaus](#) with pitchforks, it doesn't explain why did the government go an extra mile and introduced legislation clearly inspired by a failed popular initiative.

It's not that hard to explain though. You just have to put yourself in the government's shoes.

The result of the initiative haven't been spectacular (36.3% in favor) but the preliminary opinion polls have shown much larger support. At times it even looked like the initiative may succeed. So, presumably, a lot of people felt that something should have been done against the urban sprawl, but they disliked the particulars of the initiative.

If the underlying tension was not relieved, it would probably lead to similar initiatives in the future. And one of those may succeed.

But the government is not particularly thrilled about successful popular initiatives. First, they have no say in the exact wording. Silly things may get in, not because of ill will but because the text of the initiative was written without fully understanding the full scope of the problem and its on-the-ground consequences. Consider one of the government's arguments against the urban sprawl initiative, one, which the authors have clearly overlooked: The initiative would not allow to build greenhouses or poultry halls on agricultural land. The farmers would have to move them to the construction zones where the land is much more expensive. That, in turn, would harm the agriculture.

Second, the text of a successful initiative is incorporated into the constitution and is therefore set in stone. From that point on the government has to carefully navigate around its provisions and the more of them there are the more the terrain it has to operate in resembles a minefield.

All in all, it is, in the long run, much cheaper to introduce a law that addresses the concerns and takes the wind out of sails of any future initiatives, while, at the same time, not going as far as to get rejected in a legislative referendum by the opposing factions.

Similar spirit of compromise permeates all the Swiss political institutions. Political scientists even have a term for it. Ordinary democracy, as we know it from elsewhere, they call "competitive democracy" (individual parties try to win over other parties). Democracy in Switzerland they call "concordance democracy" (stakeholders try to find a compromise).

And while that may sound idealistic, an important lesson to learn from the above is that the compromise does not happen because of good will or brotherly love. It is rather the systemic result of how the mechanisms of the Swiss political system work.

## Concordance System

As much as I would like to write about concrete examples of compromises at all levels of government and society, the problem is that "spirit of compromise" is hard to quantify. When villagers meet in a pub and decide to solve a local problem by compromise, there's no paper record left behind. However, to understand the phenomenon, we do need data and that means paper record.

So, instead of looking at the problem of compromise in general, let's have a look at a substitute problem, the problem of distribution of seats in the government.

In competitive democracy each party is trying to usurp all the seats in the government for itself. If not possible, it tries to form the smallest ideologically coherent coalition that would still command majority of votes.

In concordance democracy, on the other hand, the seats are offered to all major parties in the hope that shared responsibility will bind them together and make them come up with well-balanced compromises.

Building of the concordance democracy is a historical process. When modern Switzerland was formed in 1848, nobody even thought of shared governance. Single party - or rather ideological faction, parties came later - governed the country for decades. It was only in 1891 that first opposition politician got a seat in the federal council. In 1929, another opposition party joined. And yet another in 1943.

Following diagram shows how the seven seats in the Federal Council were distributed among the parties. Blue stands for [FDP.The Liberals](#), orange for [Christian Democratic People's Party \(CVP\)](#), green for [Swiss People's party \(SVP\)](#), red for [Social Democratic Party \(SP\)](#) and yellow for [Conservative Democratic Party \(BDP\)](#). The diagram is getting gradually more colorful and tracks the progress of concordance mindset over the course of history.



## Opposition Enters the Government

In 1845, catholic cantons formed the so-called Sonderbund ("Separate Alliance") and split from the rest of the confederation. That led to the civil war among catholic and protestant cantons in 1847 won by the protestants. As a direct consequence of the war, the modern, federal Switzerland was formed and the cantons lost their independence.



Author: Marco Zanolli, CC BY-SA 3.0

The Catholics were thereafter looked at with suspicion, possibly as a prolonged hand of the pope, and haven't been given place in Federal Council for 43 years.

Between 1874 and 1891, however, they were able to form different alliances and sabotage the government by the means of referenda. Most importantly, they've been able to thwart government's plan for nationalization of the railways. That in turn lead to a government crisis. The crisis was eventually resolved by giving the opposition, for the first time in history, a place in the Federal Council.

Christian Democrat Josef Zemp became the head of the railway department. And although he used to be a passionate opponent of the nationalization before, now he took the government consensus for his own and began enthusiastically working on the nationalization project. So eagerly, in fact, that today he is remembered as one of the founders of the Swiss Federal Railways.

The modern understanding of the [collegiality principle](#) was thus conceived: Although the Federal Council consists of different parties, it speaks with one voice. The members decide on matters by voting among themselves, but once the decision is made every member stands behind it and defends it with official arguments.

In other words, the moment a person joins the Council, their main loyalty is no longer to their party. It is to the Council, and eventually, to Switzerland itself.

## Magic Formula is Born

In 1919, when the Liberals lost the absolute majority in the parliament, Christian Democrats got the second government seat. In 1929 agrarian BGB party (now SVP) joined the club.

At that point, the largest party that was still in opposition were the Social Democrats. And that was a big deal. Worker movement was strong in Switzerland. The general strike in 1918, with 250,000 workers involved, 110,000 soldiers being mobilized and paramilitary units being formed, almost led to a civil war. Bloodshed was prevented, rather miraculously, only by the distinctive Swiss ability to compromise.



*Army is prepared to quell the general strike. Bern, 1918.*

Still, the candidates of Social Democratic Party to the Federal Council were being rejected. And understandably so, given that the party still had dictatorship of proletariat in its programme.

But after its left wing split off to form the Communist Party of Switzerland the Social Democratic Party became more acceptable to the mainstream. In 1935 it rejected dictatorship of proletariat. In elections of 1943 it got 28.6% and became the largest party in the National Assembly. Shortly after, first social democratic federal councilor, Ernst Nobs, has been elected. (The fact that the memories of general strike were still around and that the governing parties didn't want to have workers against them while Switzerland was under the threat of Nazi invasion, may have also played a role.)

However, in 1953, when Social Democrats' attempt at reforming federal finances failed, Max Weber, successor to Ernst Nobs, resigned from the Federal Council and the party

withdrew to opposition. At that point, the stage was set for the most important event in the history of the Swiss concordance, the introduction of the so-called Magic Formula.

The story begins with the liberals (FDP) electing their own representative to the free position in the Federal Council. From their point of view it was just the return to the pre-1943 state of affairs. From everybody else's point of view it was just FDP being greedy. FDP got 24% in the elections, after all, and CVP 22.5%, almost exactly the same. Yet FDP got four seats and CVP just two.



The fact that even the position of federal chancellor was taken by FDP and not granted to the junior partner in the government didn't make the things better. All of that has, understandably, caused resentment in CVP.

A short-sighted CVP politician may have focused only on winning one seat from FDP and achieving parity. The general secretary of CVP, Martin Rosenberg, however, realized that there's a different stable arrangement in sight, and, as it happened, one that was very compelling and had deep political logic.

If FDP, CVP and SP each got two seats and SVP one seat - an arrangement that would later become known as Magic Formula - it would not only almost perfectly match the election results. It would also address the two big splits in the Swiss society. First, it would mean that three federal councilors on the right (FDP and SVP) would be counterbalanced by two councilors on the left (SP). Second, the historically underrepresented Catholics (CVP), despite having just two seats, would become the balancing force between the right and the left.

The problem remained of how to get there.

To understand the complexity of the task consider how the selecting of a new councilor looks like. Not only are they expected to be from a specific political party, so that the existing balance of power is preserved. Balance between language regions is also important. Typically, at least two councilors are from either French or Italian speaking regions. The most populous cantons (Zurich, Bern and either Geneva or Vaud) tend to get one seat in the Federal Council each. At the time there was also a rule that there should be no more than one councilor from any particular canton. He or she should also, in the words of historian Urs Altermatt, be "cut out of average wood," i.e. they should not stand out too much. Finally, given that absolute majority of the Unified General Assembly (both chambers of the parliament) is needed to vote them in, they must be acceptable not only within their own party, but for the other parties as well.

The list of requirements sounds crazy. Finding even a single councilor would be hard. Luckily though, the candidate doesn't have to be a member of parliament. Every Swiss citizen is eligible, so there's a huge pool to draw from. Still, the complexity of finding a suitable person may be one of the reasons why the councilors are, if they don't choose

to resign themselves, almost always re-elected. In fact, there's a strong taboo against removing an incumbent federal councilor. It has only ever happened four times. Twice in 19th century, never in 20th century and twice in 21st century. As a consequence, federal councilor spends on average ten years in the office.

Given these circumstances and the fact that Social Democrats adopted "two or nothing" motto, any attempt to change the composition of the Federal Council was a long project, requiring long-term trust between political rivals. Rosenberg also had to fight opposition within CVP, where a strong faction wanted to give Social Democrats at most one seat.

In any case, CVP contacted SP and asked them to support their candidate the next time one of the FDP councilors resigns. In return, they've promised that the new councilor would resign once another FDP councilor resigned. That would open space for two SP councilors. CVP itself would, in the long term, gain no additional seats but, as already explained, they would become the decisive factor between the political left and the political right. Social Democrats accepted the deal.

The first step of the plan was executed in 1954, when one of the FDP councilors resigned. With the support of SP the third CVP councilor was elected in his place.

CVP councilor Philipp Etter was now ready to resign at opportune moment. He even rejected vicepresidency in 1956 to keep all the options open.

The moment came in 1959 when multiple councilors resigned. Everything went according to the plan until SP nominated their party president, Walther Bringolf. Bringolf, with his communist past - I believe he knew Stalin personally, although he later turned away from hardcore communism - was not acceptable for large part of the parliament. CVP made it clear that their vote depends on the suitability of the candidate. SP still nominated Bringolf, but then almost nobody has voted for him, not even the SP parliamentarians themselves. Bringolf, seeing the result, gave up and allowed the members of the party to vote for whomever they wanted. Finally, a moderate candidate from SP was elected.

From there on, Magic Formula was respected for 44 years. As as with the government seats, the power was distributed fairly in other government-related institutions: In the Federal Court, in the military, in the state-owned companies such as Swiss Federal Railways, the post office, the national bank, or in the public television. Introduction of the Magic Formula has started the most politically stable period in the history of Switzerland.

## Towards Moderation

As we have seen with Walther Bringolf, it's the National Council that elects a Federal Councilor. Even though the claim of a party to a seat isn't disputed it doesn't mean that the official candidate of the party will be automatically elected.

It gets worse. The councilor representing the party may be elected against the will of the party.

In 1983, SP nominated Lilian Uchtenhagen, who would have been the first female federal councilor. But the parties on the right considered her politically too far on the left.

After several unsuccessful attempts to vote in different candidates seen as more moderate, a new candidate, Otto Stich, was nominated by the center-right FDP party. Otto Stich was a social democrat, but FDP considered him to be moderate, almost as he was one of their own.

He got elected and he accepted the election against the will of the SP leadership. The leadership then seriously considered retreating into opposition. However, plenary assembly of the party in 1984 decided against the idea.

In the end, Stich proved to be more left-wing than anybody expected him to be. FDP may have later regretted nominating him in place of Lilian Uchtenhagen.

While voting in a councilor against the will of their own party may be rare, this dynamic means that the federal councilors are generally chosen from moderate wings of their respective parties, resulting in low polarization within the council.

## **Magic Formula is Broken**

Up to this point, it may still seem that Switzerland is an old-world bucolic utopia, in which politicians bow to each other and say, "After you, sir!"

What we are interested in though is how the system copes with bad-faith actors and deliberate attempts to subvert it. System based on people being decent, on the rule by consensus seems to be particularly susceptible to such attacks.

Enter Christoph Blocher, the man whom Steve Bannon once described as "Trump before Trump".

Blocher is a great case study. He's capable, he's rich, he's a great speaker and his political project (independence, neutrality, self-sustainability, anti-immigration) has a huge appeal in the Swiss society. At the same time he's willing to fight unfair, bend the rules and even break the Swiss political system to achieve his political goals.

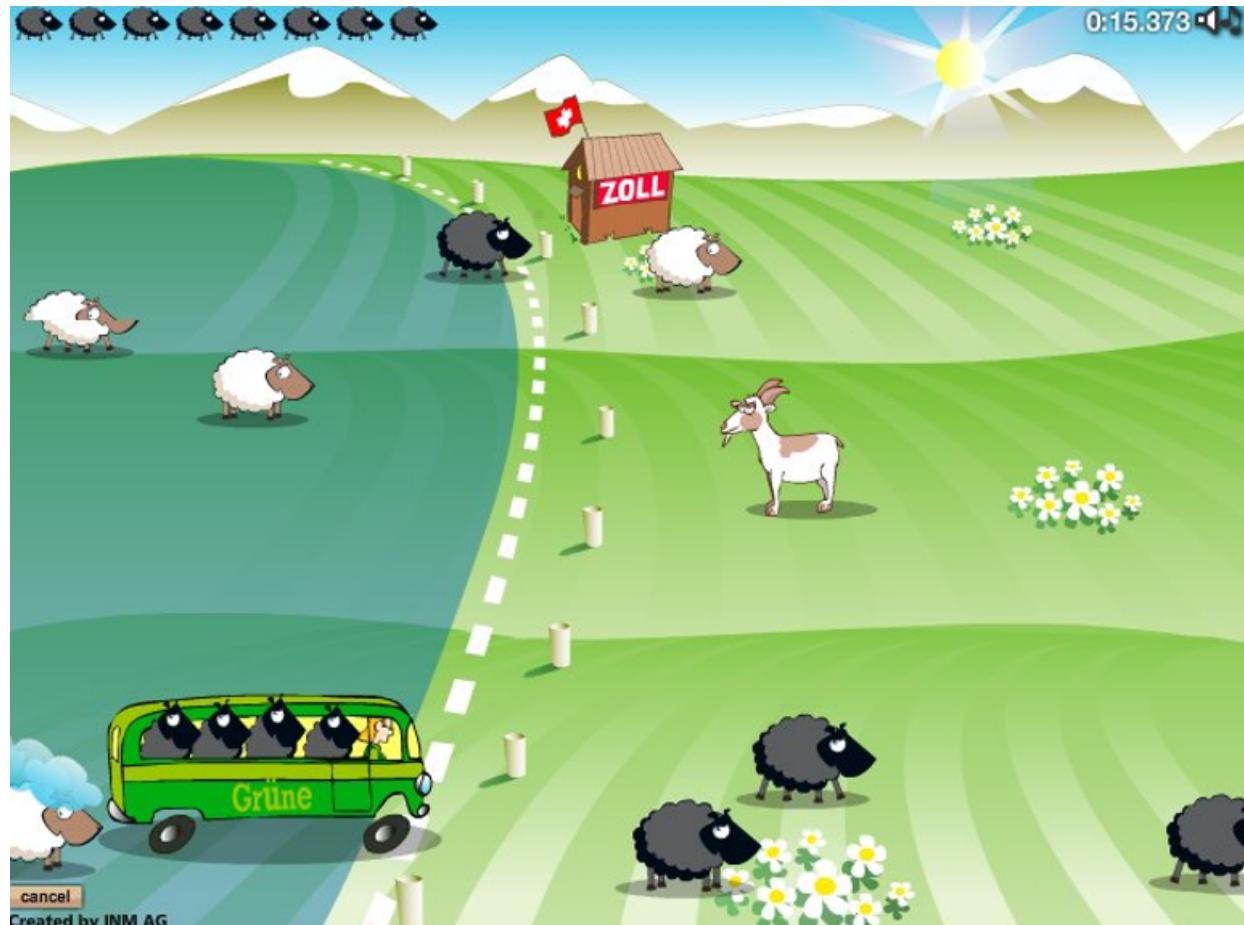
In 1972, Blocher, originally a lawyer and businessman, joined the Swiss People's Party, an agrarian party whose main agenda at the time was centered around the price of milk. In 1977, he was elected the president of the Zurich chapter of the party and began to transform it into a modern right-wing populist party - something yet unheard of in the rest of Europe. He focuses on the anti-immigration and anti-European agenda. The result is an increase in the party's preferences in the canton.

He gradually gains influence in the party, mainly at the expense of the old, moderate wing. His high point comes in 1994 when he helps to prevent Switzerland joining the European Economic Area. As the party gradually takes over Blocher's agenda, its preferences grow until, in 1999, the once smallest coalition partner becomes the strongest party in the parliament (22.5%).

Animated graphics showing how SVP ceased being a traditional economically conservative and socially progressive party and, while remaining economically conservative, quickly moved towards socially conservative end of the spectrum, can be found [here](#). (By the way, the graphics also shows the ongoing centralization of Swiss politics: Note how cantonal chapters, represented by small dots, draw closer to their respective national party over the years.)

Unlike right-wing populist parties elsewhere though, SVP puts emphasis on democracy. In particular, it stresses direct democracy, while at the same time downplaying the rule

of law. SVP doesn't like the idea of people deciding only on the rules which are then applied by impartial institutions. They would like the people to decide on particular cases. An example of this would be their failed attempt in the city of Zurich to start granting citizenship by ballot. (This has been tried in year 2000 in the municipality of Emmen and it turned out that a lot of candidates were turned down just because they happened to have Yugoslavian-sounding names. On the other hand, a girl from Montenegro with an Italian-sounding name has got the citizenship. The procedure was declared to be unconstitutional by the Federal Court in 2003.)



An online game from SVP. The player is asked to fight aliens, the Greens and the judges.

Blocher himself became a member of the government of the canton of Zurich in 1975, and a member of the federal parliament in 1979.

In parliamentary elections in 2003, SVP gains 26.7% and Blocher demands a Federal Council seat for himself.

And while, numerically, SVP has justified claim to the second seat, the Magic Formula was originally intended to balance the forces on the left with the forces on the right. With the right-wing SVP getting a second seat at expense of the CVP the balance between right and left becomes 4:2, with centrist CVP, now having only a single seat, no longer being able to make a difference.

There are other reasons not to vote for Blocher. In 1994, during a vote in the National Council, he pressed the button of his absent colleague.

In 1997, during the discussion of Jewish gold in Swiss banks he says: "The Jewish organizations that demand the money say that it's ultimately not about the money. But let's be honest: That's exactly what it's about." When a newspaper summarizes it as "The Jews only care about money," Blocher sues the journalist for defamation. The court later decides that the article haven't betrayed the spirit of Blocher's speech. He has to bear the costs of the case and part of the costs of the investigation, as well as pay 10,000 francs in compensation to the editor.

But also in general, his authoritarian and uncompromising attitude directly clashes with the consensual way how the politics are done in Switzerland.

Despite that, some are still in favor of making him a federal councilor. Carlo Schmid (CVP): "It would be an error not to elect Blocher to the Federal Council. ... Blocher has an enormous disruptive potential. That potential must be neutralized."

That may sound strange, but that's how Swiss politics work. One doesn't try to lock their opponents out. One rather offers them a trade of getting access to power in exchange for submitting to the consensus. We've already seen this happen with Social Democrats half a century earlier.

But it can also be thought of as education: When Blocher becomes a federal councilor, he enters a group of people who are working together for years, unlikely to be removed, not supposed to be driven by the interests of any party, bound together by the principle of collegiality. The members change one at a time and the body therefore has a lot of continuity. It has evolved a specific [culture](#), that is passed from generation to generation since 1848. The new councilor is expected to become part of that culture, to, so to say, grow up and put the interests of the country above the interests of their party.

To give a concrete example, it sometimes happens that a new federal councilor can't resist the temptation and violates the collegiality principle by hinting at how they've personally voted. Afterwards, they have to face a lot of criticism, both from within the council and from without. Eventually, they learn the lesson.

So, these were the expectations when Blocher was elected to become a Minister of Justice and Police in 2003. Along the way, Magic Formula was broken and an incumbent councilor was not re-elected for the third time in the Swiss history.

## The Most Successful Conspiracy since Brutus

At least that's how the 2007 de-election of Christoph Blocher from the Federal Council was called by the press.

But let's get back to what happened. By 2007 SVP gains even more votes (29%) but at the same time it is already clear that Blocher refuses to grow up. Although a federal councilor, he still does opposition politics. Not only he breaks the principle of collegiality, he attacks it head on. Here's an example where he manages to do both at the same time: "The vote in the Federal Council about entering the Schengen area hasn't been unanimous and even today the entire council doesn't stand decisively behind the project. ... Some people believe that the government can decide that the Earth is flat and that the collegiality principle prevents a councilor to say that it is round!"

He doesn't give up his populist antics either. In 2006, he breaks presumption of innocence by referring to two immigrants under investigation as criminals. When questioned about it he lies to the senate. However, he is convicted by a video recording.

His opponents also accuse him of ignoring parliamentary decisions, of trying to control the parliament with threats, of criticizing court decisions and of delaying bills that he does not like.

All that being said, everyone still expects him to be re-elected.

Social Democrats feel there is a chance though. They know that the majority in parliament is fed up with Blocher and getting rid of him hangs only on finding a viable counter-candidate.

But finding one is not easy.

First, it must be someone from SVP. Candidate from any other party would face opposition from both SVP and at least one other party and would have no chance of getting the majority of votes.

Second, SVP already announced that whoever accepts a counter-nomination for Federal Councilor would be expelled from the SVP parliamentary faction. Therefore, SVP parliamentarians would be under strong pressure not to accept the nomination, even if they were elected.

Third, the candidate would have to make for a competent federal councilor. People would not vote for a random nobody.

Fourth, the counter-candidate would have to be better than Blocher himself. At least, he or she should be willing to accept the rule of law and the collegiality principle.

SP decided to go on with Eveline Widmer-Schlumpf, a member of the old, moderate, milk-price-oriented, wing of the party. At the time, she was not a member of federal parliament and so she was, to some extent, immune to Blocher's threats. She was the president of the Conference of Cantonal Ministers of Finance, which, as you may remember from the previous part, is a body operating on the national level, but parallel to the federal government. That is, she is no newbie to the big politics. In fact, Widmer-Schlumpf was considered a good candidate for a Federal Councilor by SVP itself. In 2003, the party president and ally of Blocher, Ueli Maurer, describes her as "a very valid candidate, one of the most competent politicians in the country." In other words, she was no leftie secretly infiltrating the conservative SVP.

Moreover, the party's radicals now have the upper hand and just a week before the election they remove two dissenting deputies from the Graubünden section of the party (which Widmer-Schlumpf belongs to) from important parliamentary commissions. One of them describes the conditions in the party as "dictatorship". SP counts on Widmer-Schlumpf not being amused.

Social Democrats contact the leadership of the centrist Christian Democrats. They propose to elect Widmer-Schlumpf instead of Blocher and they get their support.

They manage to make a secret agreement with the Greens to withdraw their own candidate, who had no chance of winning anyway, at the last minute.

Even some individual members of the right-wing Liberal Democratic Party are contacted. The party is officially in favor of Blocher, but he is unpopular in the French and Italian-speaking cantons and some parliamentarians from these cantons may be willing to vote against him.

The result of the vote comes as a shock to the People's Party. The party's sovereign leader is not re-elected and the current outsider, Mrs. Eveline Wider-Schlumpf, becomes the new representative of SVP in the government.

SVP asks Wider-Schlumpf not to accept the position. She asks for a day to decide. Demonstrators, mostly leftists, in front of the Bundeshaus are, funnily, expressing support for a right-wing politician. Large banner they hang out reads: "Eveline, say yes!"

The next day she indeed agrees and becomes a competent federal councilor for next eight years. She has extremely high approval numbers and in 2008 she even wins the award "Swiss Person of the Year".

## **The Soap Opera Continues**

It wouldn't be Blocher if he haven't tried to fight back.



*Christoph Blocher during the campaign against joining the European Economic Area (1992).*

When Widmer-Schlumpf accepts the seat he aims for exemplary punishment and wants her to get thrown out of the party. Same applies to the other SVP federal councilor, Minister of Defense Samuel Schmid, who refuses to resign in favor of Blocher.

But there's a catch. The Swiss People's Party is a union of its cantonal chapters and as such it cannot exclude individual members. The party leadership therefore asks the People's Party of the canton of Graubünden to expel Widmer-Schlumpf. The cantonal section says no way. The party leadership threatens to exclude the entire cantonal chapter if it does not play along. The chapter replies that whatever.

In 2008 the Swiss People's Party of Graubünden is therefore excluded from the national People's Party. It renames itself to the Conservative Democratic Party (BDP) and stands in the next election as a separate entity.

Similar scenario takes place in the canton of Bern. The party's cantonal section refuses to expel the incumbent councilor Samuel Schmid. Although the party leadership does not resort to the same drastic measures as was the case with Widmer-Schlumpf, it publicly declares that the elimination of Schmid would be a waste of time and that he is already "clinically dead" anyway.

A large part of the People's Party of the canton of Bern, including seventeen members of the cantonal parliament, therefore joins the new BDP party. Similar scenario plays out in the canton of Glarus.

Blocher announces that from now on, People's Party will be pursuing opposition politics. In the Swiss context, it means trying to sabotage the government by referenda.

When Samuel Schmid resigns from the Federal Council in 2008, SVP nominates Christoph Blocher again. However, it is immediately clear that he has not chance of winning. What's worse, SP tries to pull the same trick again: They nominate Hansjörg Walter, a moderate SVP politician from canton Thurgau and the president of Swiss Farmers' Union.

After the first round of voting where no candidate gets the majority, the party, scared, pulls out Christoph Blocher. The party president, Ueli Maurer, is nominated instead. But he's not seen as any more moderate than Blocher himself and many fear that with Maurer, the entire Blocher story would repeat anew.

In the second round of voting the moderate candidate gets two more votes than Maurer does and needs only a single additional vote to get the absolute majority.

The president of SVP parliamentary faction addresses the deputies and begs them in the name of Christoph Blocher (to everyone's amusement) to vote for Maurer.

Finally, in the third round, Maurer wins the election by a single vote.

It has to be said that Maurer, unlike Blocher, has grown into the position and turned out to be an acceptable federal councilor.

In 2011, after new elections, the People's Party once again tries to take advantage of the Magic Formula and gain the second seat in government. They want to replace Widmer-Schlumpf, who is now a member of a different party, by their own candidate (not by Blocher). Nevertheless, when deputies are given a choice between two traditions, the tradition of the Magic Formula and the tradition of never de-electing incumbent councilors, they decide to prefer the latter principle and re-elect Widmer-Schlupf.

When People's Party gets its all time best numbers in the elections in 2015 (29,4%) and Widmer-Schlupf's new Conservative-Democratic Party only 4%, the councilor voluntarily resigns. People's Party decides to play the card of national cohesion rather than taking a risk of nominating Blocher again. One of the three candidates is from the French and another from the Italian-speaking part of Switzerland. Finally, Guy Parmelin, a relatively moderate candidate from the canton of Vaud, is elected.

The concordance, after twelve years of troubles, becomes fully operational again.

## **Rule by Consensus**

I've chosen to write about the history of Magic Formula in painstaking detail. The idea of concordance and rule by consensus is so alien to the people for competitive democracies that nothing less, no theoretical explanation, would be able to convey how it works. The reader would be left to choose between disbelief and incorrect assumption that Swiss are a bunch of clueless hippies.

But that being said, what is the actual mechanisms that makes the Swiss political system tick?

In my opinion, the core of it lies in the fact that everyone is in minority, every time. Switzerland, thanks to its diversity, gets that for free. However, unlike, say, similarly diverse Afghanistan, it managed to turn that disadvantage into an advantage. Instead of fighting each other forever and getting a failed state in result it somehow managed to move the struggle from the military to the political domain. What's more, it managed to fight the problems not by homogenizing and centralizing the country, but rather by making the diversity - and thus the general experience of being in minority - the leading principle of its political system.

You want to be a federal councilor? You are left at the mercy of your political opponents. It doesn't matter that you are the strongest man in the strongest party. You have to try hard to placate the others so that they vote for you.

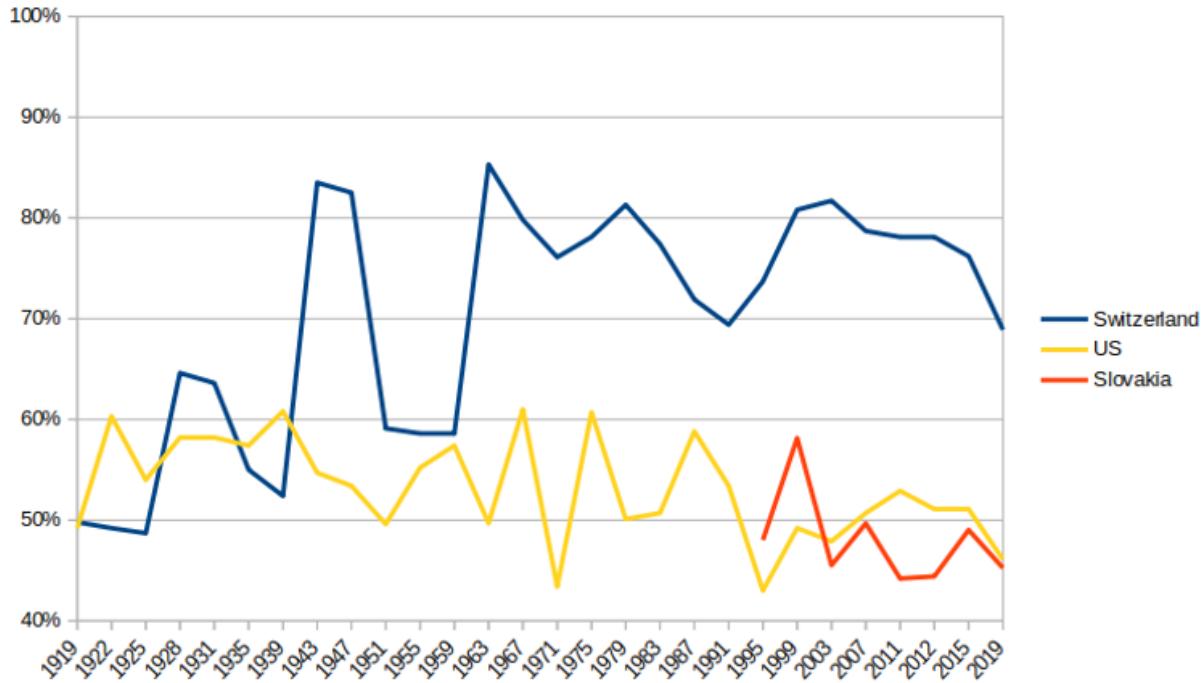
You became a president and now you think you are all-powerful? Well, you can't do anything unless you are able to convince and compromise with the remaining federal councilors, who are, for the most part, your ideological opponents.

You want a certain law to be passed? You have to make sure it won't be challenged in a referendum. You have to think twice and thrice about which particular group you've forgot about, which is going to be angered enough to launch a referendum. Then you have to negotiate with them and compromise.

## **Consequences: Legitimacy of Government**

With all the above being said, here's the interesting question: What is the legitimacy of Swiss government?

There are many possible ways to measure it, but let's have a look at the simplest one: How many voters have voted for the parties, which then got a seat in the Federal Council?



Before interpreting the graph, consider the following two facts: The area below fifty percent is reserved for undemocratic regimes (minority rules majority). Also, 100% is an unattainable goal. There will always be some people with fringe political views who won't be represented in the government.

For comparison, I've added the data for the US and for my native Slovakia (I couldn't resist the urge. Sorry.) In the US case the graph represents the portion of voters who voted for the ruling president. In the case of Slovakia it's people who voted for the coalition parties.

As can be seen, both countries wobble around 50%. Naively, one would expect that legitimacy in a democratic country cannot fall below 50%. Whoever is supported by less than half of the voters should be outvoted and unable to form a government. In reality, however, there are many ways to fall into this trap. A country may have a weird electoral system (US), it may have a minority or bureaucratic government (Belgium) or, as in the case of Slovakia, quotas for the entry into the parliament can leave a large part of the voters (28.5%) without representation in the parliament, let alone in the government.

In reality, legitimacy close to 50% seems to be quite common in competitive democracies. It kind of follows: With competitive mindset one wants to band together with only as many competitors as it takes to get the majority. Going any further means unnecessarily "losing" the contest.

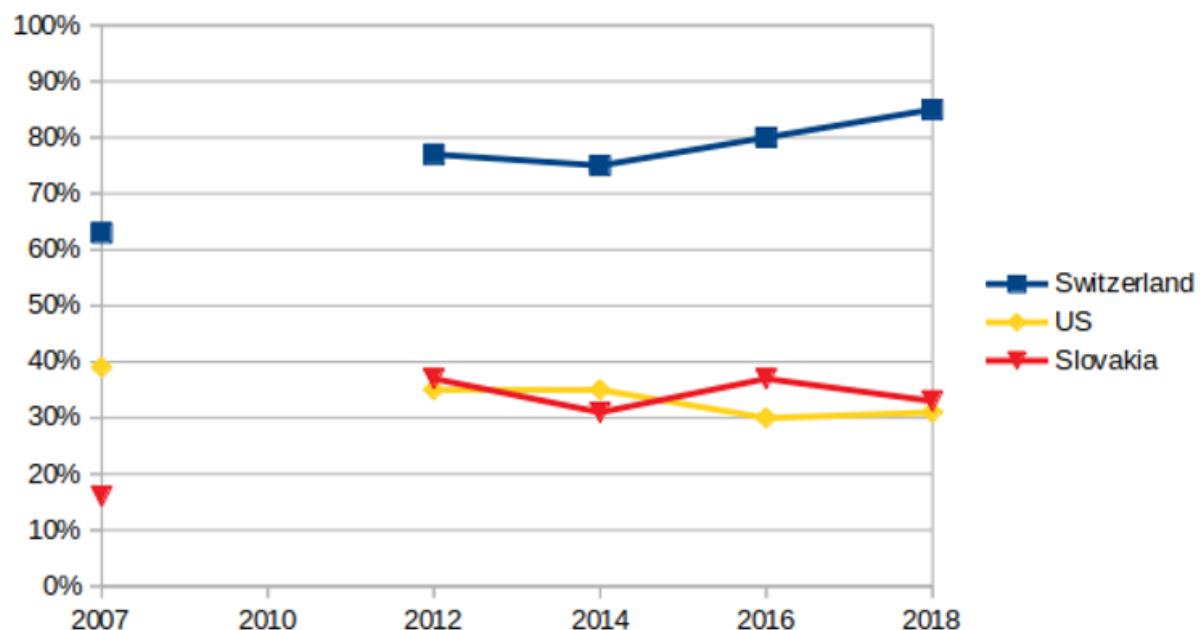
The concordance democracy in Switzerland, on the other hand, oscillates, since the introduction of Magic Formula, around 75%.

In 2019, the legitimacy declined somehow due to the high electoral gain of the Greens, who traditionally have no representation in the Federal Council. All other parties agree that it would be appropriate to give them a seat, but none of them wants it to happen at their expense. In any case, if any of the proposals put forward at the recent summit

on the issue was implemented, the government's legitimacy would rebound back to at worst 82 and at best 90 percent.

## Consequences: Trust in Government

Next, let's have a look at whether the high legitimacy of the government is reflected in the trust of citizens in the government.



Source: OECD's "Government at a Glance" report

It's hard to generalize from a single data point, but indeed, it looks like concordance democracy may result in higher overall trust in government. Both the US and Slovakia are somewhere around 35%. Average of OECD countries is somewhat higher, 45%, but still way down below Switzerland's 85%.

Let me diverge from the topic for a second and make a remark about trust in general.

When I've moved to Switzerland I've been warned couple of times that the neighbors there won't think twice before ratting on me.

And while that's true - although reporting to the authorities is usually preceded by passive-aggressive messages in one's mailbox - the context is very different.

In Eastern Europe and, I guess, in many other places, reporting to authorities is seen as morally wrong. There is a kind of stubborn popular solidarity in resistance to power. We may have inherited that attitude from the times when ratting on someone meant that men in leather trench coats arrived early next morning and dragged the victim to Gulag. (I've even heard a story about a small Slovak town where, shortly after World War II, people began reporting their personal enemies to the Russians, claiming that they were Nazi collaborators. Russians had no clue and incarcerated every reported person. The feud spiraled out of control and several hundred people ended up in jail.)

Swiss, on the other hand, don't perceive authorities as necessarily hostile. If a neighbor violates a rule (and people have quite likely voted for that rule or at least haven't objected when it was introduced) he should be warned first, and if that doesn't help, reporting him to the authorities is seen as fully justified. Nothing terrible is going to happen anyway. Most likely, the authorities are just going to ask the person in question to behave.

But it is not just the neighbors that people are complaining about. Complaints seem to be a popular pastime. If something is not working smoothly, be it a person or an institution, it is a reason to complain, and surprisingly, complaints often lead to solving the problem. In result, Switzerland has a fantastic, effective bureaucracy.

In short, there is a certain degree of trust ("armed trust" may be a better description), on the one hand of people in the authorities, on the other hand also of the authorities in the people. There is a certain implicit assumption that if a person has broken a rule, he has not necessarily done so on purpose and does not need to be punished immediately. Everyone is watching closely though whether the "mistake" was really a mistake or whether it happens again.

A nice illustrative example on the topic of trust are the quarantine measures during the Easter holidays of 2020. The government decided not to ban traveling but rather to put trust in the people and issued a recommendation to behave responsibly and not to travel.

On the first day, police of canton Uri reports that they are stopping cars on the road to the Gotthard Tunnel and convincing people not to travel. Police spokesperson reports: "We haven't convinced anyone yet."

The next day, they report the first success: "One car has turned around and gone home. It seems that the message is getting through."

However: "About 98% of those stopped have a good reason to travel to Ticino - because they live there, or have a family there, or very important responsibilities. Very few people travel there on vacation or for fun."

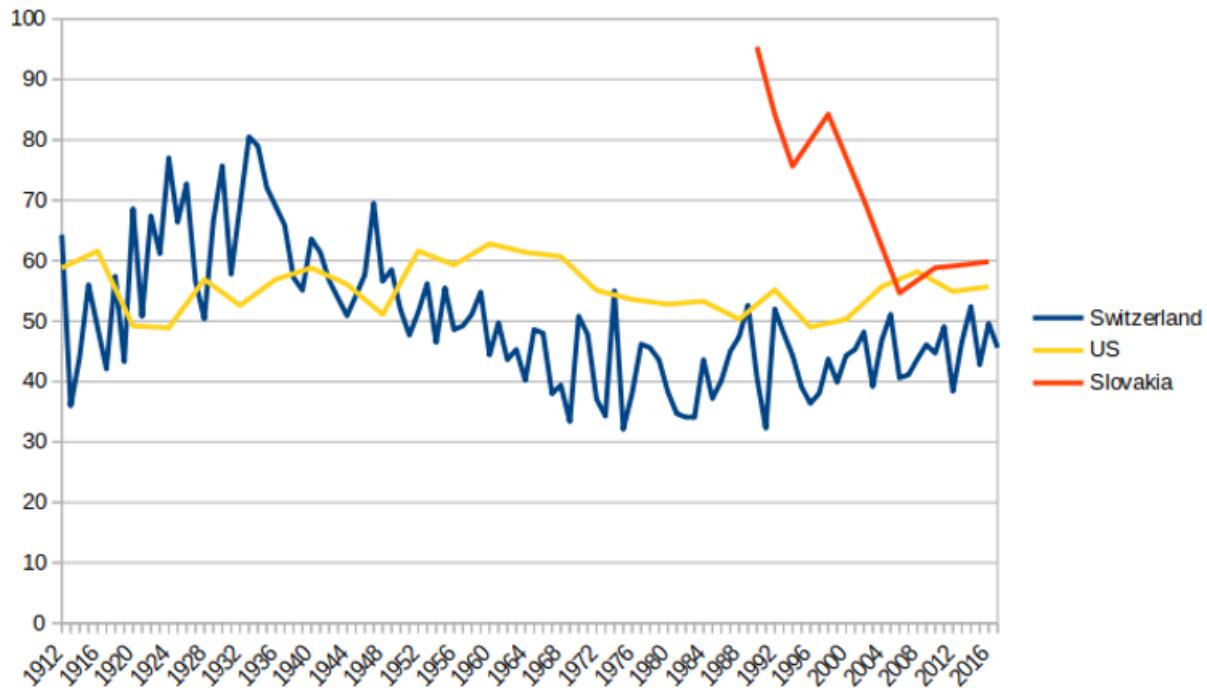
And as funny as it sounds, it must be said that the road beneath the Gotthard Pass, notorious for bad traffic jams during the Easter holidays, was almost completely empty this year. Contrast it with the Slovak government which tried to limit the travel during Easter holidays and caused massive traffic jams a day before the restrictions came into effect.

On the other hand, after government made few mistakes in handling the crisis (such as downplaying the importance of masks, once publishing wrong number of infection cases and, to everyone's horror, reporting that a nine-year-old girl had died, only for it to later turn out that the girl was in fact 109) a local tabloid concludes - showing how important the trust is considered to be: "[Director of the Federal Health Bureau] has asked the cantons, because of the rising corona numbers, to step up the measures. Among other things he recommended the usage of masks in shops. Up to this point hardly any canton has followed the recommendation."

## **Consequences: Election Turnout**

It may not be directly related to the topic, but it is interesting to have a look at whether high trust in the government correlates with high election turnout.

That appears not to be the case. Turnout in Switzerland is chronically weak, oscillating around forty-five percent.



*A mixed bag. Turnout at referenda in Switzerland. Turnout at presidential elections in the US. Turnout at parliamentary elections in Slovakia.*

However, a detailed analysis shows that the situation is not that simple. It is not the case that half of the Swiss never vote. It turns out that about a quarter votes almost every time, a quarter almost never, and the rest votes occasionally. That being said, 90% of people have voted at least once in the past five years. This suggests that people vote selectively. If a referendum is about education, it is only natural that people without children in school ignore the vote. On the other hand, it is true that a certain part of the electorate can be only lured to the ballot box by simple topics with a big emotional appeal, e.g. restrictions on immigration.

**August 10th, 2020**

# **Tagging Progress at 100%! (Party & Celebratory Talk w/ Jason Crawford, Habryka on Sun, Aug 30th, 12pm PDT)**

## [Tagging Open Call / Discussion](#)

***There will be an online Tagging Progress Bar Completion Talk + Party on Sunday, August 30th at 12:00PM PDT. We'll start with a public conversation between [Oliver Habryka](#) (LW team lead) and [Jason Crawford](#) (author of [Roots of Progress](#)) on the topic of intellectual progress, followed by a freeform party in Gather.town or Topia.***

--

Probabilities [can't be 100%](#), but progress bars can be.

I am delighted to announce the Tagging Progress Bar which has graced the LessWrong front page for three weeks is now gloriously full. Every post with over 25 karma now has at least one tag applied to it, most having several.

Top among of the goals of tagging was to explicitly shape LessWrong more towards *long-term intellectual progress*: with a major pass at tagging the archives completed, it is now vastly easier to see [what LessWrong has discussed over the years](#), find things you didn't know to search for, locate posts based on interest rather than recency or karma, catch-up on everything said on a topic to date, and generally [continue the conversation](#) on topics not currently on the frontpage.

Authors and commenters will hopefully now post to LessWrong secure in the knowledge that their posts and comments will not be forgotten after a week or two. Instead, the best content on each topic will be discoverable for years to come.

## **Thank you, taggers!!**

An enormous shout-out goes out the tagging contributors. Everyone who tagged posts, created tags, wrote tag descriptions, voted tags, or just offered feedback. Tagging is not something the LessWrong team could have rolled out on our own, we needed your help and all users of the tagging system are indebted to you.

## **The Tag Numbers**

To give you a sense of what was accomplished, here are some basic numbers (as of publishing):

- 14,028 tags applied
- 7,427 posts have been tagged
- 419 active tags created

## Cumulative Number of Tags Applied to Date



## Top Taggers

The tagging system doesn't yet have good recognition and reward systems in place, so unfortunately these taggers haven't gotten the karma they deserve. Still, everyone should know these are the folks whose efforts brought us success.

### Top Tag Appliers

- [Multicore](#)
- [Kaj\\_Sotala](#)
- [Gyrodio](#)
- [Yoav Ravid](#)
- [abramdemski](#)
- [TurnTrout](#)
- [Zack\\_M\\_Davis](#)

### Top Tag Creators & Description Writers

- [Kaj\\_Sotala](#)
- [Yoav Ravid](#)
- [Multicore](#)
- [brook](#)

As thanks, these top taggers will each receive the professionally-edited, physical five-book set of posts from the [2018 Review](#). A fitting prize, since that project was also about long-term intellectual progress.

## Party time!

We're overdue for a party regardless, this occasion definitely deserves an event.

Please join us **Sunday 30th Augusts at 12:00 PM PDT**. We're planning a two part event:

- 1) A public conversation on Intellectual Progress between [Oliver Habryka](#) (LW team lead) and [Jason Crawford](#) (author of [Roots of Progress](#)) on the topic of intellectual progress. This stage of the event should run approximate for any hour.

2) Following the talk, we will have a freeform in Gather.town or Topia.

Stay tuned for event post / links.

## This isn't even my final forum

Giving all historical posts above 25 at least one tag is a huge milestone but, like a credence, Tagging isn't something that reaches 100%. There's more to be done.

- Each day, users are finding new concepts or clusters of posts worthy of having their own tags. There are 400 tags right now, and my personal guess there are many, many more to still be created.
- Some tags should be merged, others split into two or more. Especially some of the large tags harbor several sub-clusters within them that could be identified and separated.
- Many tagged post don't have every appropriate tag. The most appropriate tag might not have been created.
- Most tags haven't had their posts list carefully sorted to the best and most relevant posts at the top.
- Most tags do not have even good basic descriptions. Either there is no description, the description is a poor explanation, or the description fails to link to other relevant tags.
- Keeping the tagging system high quality means doing "maintenance" by removing bad tags and tags that don't really apply to posts.

If tagging is to succeed long-term, it'll depend on the ongoing effort of those generous enough to make it good even when it's no longer in the spotlight.

I think we'll have to give trusted users Tagging Moderator privileges that lets them merge/split/delete tags as part of doing the needed work.

## Tag Discussion/Talk Pages

I've promised them in several places, and they're nearly here. Soon, tags will have discussion pages on which users can discuss that tag: What should it be called? what's the actual definition? Should this post be included? Those kinds of questions.

I hope this will make it much easier for taggers to coordinate with each other, and for others to see the tagging efforts and offer feedback on them.

## Tagging Activity in Recent Discussion Feed

We've also done some work to have tagging activity such as major edits to tag descriptions appear in Recent discussion as something that commented and voted upon. Hopefully, that works out, giving more ongoing visibility into tagging work.

## Final thanks

Hard to say it too many times, so thanks yet again to everyone who's made [tagging awesome](#). You're awesome.

# is gpt-3 few-shot ready for real applications?

This is a lengthy reply to [@the-moti](#)'s post [here](#). Creating a new post to limit thread length, and so I can crosspost to LW.

[@the-moti](#) says, in part:

This obviously raises two different questions: 1. Why did you think that no one would use few-shot learning in practice? 2. Why did other people think people would use few-shot learning in practice?

I would be interested in hearing your thoughts on these two points.

—

Thanks for asking!

First of all, I want to emphasize that the GPT-3 *paper* was not about few-shot GPT-3 as a practical technology.

(This is important, because the paper is the one large body of quantitative evidence we have on few-shot GPT-3 performance.)

This is not just my take on it: before the OpenAI API was announced, all the discussion I saw took for granted that we were talking about a scientific finding and its broader implications. I didn't see any commentator whose main takeaway was "wow, if I could do this few-shot thing right now, I could build amazing projects with it."

Indeed, a [common theme](#) in critical commentary on my post was that I was too focused on whether few-shot was useful right now with this specific model, whereas the critical commentators were more focused on the implications for even larger models, the confirmation of scaling laws over a new parameter regime, or the illustration-in-principle of a kind of meta-learning. [Gwern's May newsletter](#) is another illustrative primary source for the focus of the discussion in this brief "pre-API" period. (The API was announced on June 11.)

As I read it (perhaps benefitting from hindsight and discussion), the main points of the paper were

- (1) bigger models are better at zero/few-shot (i.e. that result from the GPT-2 paper holds over a larger scale),
- (2) more "shots" are better when you're doing zero/few-shot,
- (3) there is an interaction effect between 1+2, where larger models benefit more from additional "shots,"
- (4) this *could* actually become a practical approach (even the dominant approach) in the future, as illustrated by the example of a very large model which achieves competitive results with few-shot on some tasks

The paper did not try to optimize its prompts – indeed its results are [already being improved upon](#) by API acolytes – and it didn't say anything about techniques that will be common in any application, like composing together several few-shot "functions." It didn't talk about speed/latency, or what kind of compute backend could serve many users with a guaranteed SLA, or how many few-shot "function" evaluations per user-facing output would be needed in

various use cases and whether the *accumulated* latency would be tolerable. (See [this post](#) on these practical issues.)

It was more of a proof of concept, and much of that concept was about scaling rather than this particular model.

---

So I'd argue that right now, the ball is in the few-shot-users' court. Their approach *might* work – I'm not saying it couldn't!

In their favor: there is plenty of room to further optimize the prompts, explore their composability, etc.

On the other hand, there is no body of evidence saying this actually works. OpenAI wrote a long paper with many numbers and graphs, but that paper wasn't about *whether their API was actually a good idea*. (That is not a criticism of the paper, just a clarification of its relevance to people wondering whether they should use the API.)

This is a totally new style of machine learning, with little prior art, running on a mysterious and unproven compute backend. *Caveat emptor!*

---

Anyway, on to more conceptual matters.

The biggest *advantages* I see in few-shot learning are

**(+1)** broad accessibility (just type English text) and ability to quickly iterate on ideas

**(+2)** ability to quickly define arbitrary NLP “functions” (answer a factual question, tag POS / sentiment / intent, etc ... the sky's the limit), and compose them together, without incurring the memory cost of a new fine-tuned model per function

What could really impress me is (+2). IME, it's not really that costly to *train* new high-quality models: you can finetune BERT on a regular laptop with no GPU (although it takes hours), and on ordinary cloud GPU instances you can finetune BERT in like 15 minutes.

The real cost is keeping around an entire finetuned model (~1.3GB for BERT-large) for each *individual* NLP operation you want to perform, and holding them all in memory at runtime.

The GPT-3 approach effectively trades this memory cost for a time cost. You use a single very large model, which you hope already contains every function you will ever want to compute. A function definition in terms of this model doesn't take a gigabyte to store, it just takes a tiny snippet of text/code, so you can store tons of them. On the other hand, evaluating each one requires running the big model, which is slower than the task-specific models would have been.

So storage no longer scales badly with the number of operations you define. However, latency still does, and latency per call is now much larger, so this might end up being as much of a constraint. The exact numbers – not well understood at this time – are crucial: in real life the difference between 0.001 seconds, 0.1 seconds, 1 second, and 10 seconds will make or break your project.

---

As for the potential *downsides* of few-shot learning, there are many, and the following probably excludes some things I've thought of and then forgotten:

**(-1)** The aforementioned potential for deal-breaking slowness.

**(-2)** You can only provide a very small amount of information defining your task, limited by context window size.

The fact that more “shots” are better arguably compounds the problem, since you face a tradeoff between providing more examples of the *same* thing and providing examples that define a more *specific* thing.

The extent to which this matters depends a lot on the task. It’s a complete blocker for many creative applications which require imitating many nuances of a particular text type not well represented in the training corpus.

For example, I could never do [@nostalgebraist-autoresponder](#) with few-shot: my finetuned GPT-2 model knows all sorts of things about my writing style, topic range, opinions, etc. from seeing ~3.65 million tokens of my writing, whereas few-shot you can only identify a style via ~2 thousand tokens and hope that’s enough to dredge the rest up from the prior learned in training. (I don’t know if my blog was in the train corpus; if it wasn’t, we’re totally screwed.)

I had expected AI Dungeon would face the same problem, and was confused that they were early GPT-3 adopters. But it turns out they actually [fine-tuned](#) (!!!), which resolves my confusion ... and means the first real, exciting GPT-3 application out there *isn’t actually a demonstration of the power of few-shot* but in fact the opposite.

With somewhat less confidence, I expect this to be a blocker for specialized-domain applications like medicine and code. The relevant knowledge may well have been present in the train corpus, but with so few bits of context, you may not be able to overcome the overall prior learned from the whole train distribution and “zoom in” to the highly specialized subset you need.

**(-3)** Unlike supervised learning, there’s no built-in mechanism where you continually improve as your application passively gathers data during usage.

I expect this to be a big issue in commercial applications. Often, a company is OK accepting a model that isn’t great at the start, *if* it has a mechanism for self-improvement without much human intervention.

If you do supervised learning on data generated by your product, you get this for free. With few-shot, you can perhaps contrive ways to feed in segments of data across different calls, but from the model’s perspective, no data set bigger than 2048 tokens “exists” in the same world at once.

**(-4)** Suffers a worse form of the ubiquitous ML problem that “you get *exactly* what you asked for.”

In supervised learning, your model will avoid doing the hard thing you want if it can find easy, dumb heuristics that still work on your train set. This is bad, but at least it can be identified, carefully studied (what was the data/objective? how can they be gamed?), and mitigated with better data and objectives.

With few-shot, you’re no longer asking an *arbitrary* query and receiving, from a devious genie, the response you deserve. Instead, you’re constrained to ask queries of a particular form: “*what is the next token, assuming some complicated prior distributed from sub-sampled Common Crawl + WebText + etc.?*”

In supervised learning, when your query is being gamed, you can go back and patch it in arbitrary ways. The lower bound on this process comes only from your skill and patience. In few-shot, you are fundamentally lower-bounded by the extent to which the thing you really want can be expressed as next-token prediction over that complicated prior. You can try different prompts, but ultimately you might run into a fundamental bound here that is

prohibitively far from zero. No body of research exists to establish how bad this effect will be in typical practice.

I'm somewhat less confident of this point: the rich priors you get out of a large pretrained LM will naturally help push things in the direction of outcomes that make linguistic/conceptual sense, and expressing queries in natural language might add to that advantage. However, few-shot *does* introduce a new gap between the queries you want to ask and the ones you're able to express, and this new gap *could* be problematic.

**(-5)** Provides a tiny window into a huge number of learned parameters.

GPT-3 is a massive model which, in each call, generates many intermediate activations of vast dimensionality. The model is pre-trained by supervision on a tiny subset of these, which specify probability distributions over next-tokens.

The few-shot approach makes the gamble that this *same* tiny subset is all the user will need for applications. It's not clear that this is the right thing to do with a large model – for all we know, it might even be the case that it is *more* suboptimal the larger your model is.

This point is straying a bit from the central topic, since I'm not arguing that this makes GPT-3 few-shot (im)practical, just suboptimal relative to what might be possible. However, it does seem like a significant impoverishment: instead of the flexibility of leveraging immense high-dimensional knowledge however you see fit, as in the original GPT, BERT, [adapters](#), etc., you get even immenser and higher-dimensional knowledge ... presented through a tiny low-dimensional pinhole aperture.

—

The main reason I initially thought “no one would use few-shot learning like this” was the superior *generalization performance* of fine-tuning. I figured that if you’re serious about a task, you’ll care enough to fine-tune for it.

I realize there’s a certain mereology problem with this argument: what is a “single task,” after all? If each fine-tuned model incurs a large memory cost, you can’t be “serious about” many tasks at once, so you have to chunk your end goal into a small number of big, hard tasks. Perhaps with few-shot, you can chunk into smaller tasks, themselves achievable with few-shot, and then compose them.

That may or may not be practical depending on the latency scaling. But if it works, it gives few-shot room for a potential edge. You might be serious enough about a large task to fine-tune for it ... but what if you can express it as a composition of smaller tasks you’ve already defined in the few-shot framework? Then you get it instantly.

This is a flaw in the generalization performance argument. Because of the flaw, I didn’t list that argument above. The list above provides more reasons to doubt few-shot *above and beyond* the generalization performance argument, and again in the context of “serious” work where you care enough to invest some time in getting it right.

I’d like to especially highlight points like (-2) and (-3) related to scaling with additional task data.

The current enthusiasm for few-shot and meta-learning – that is, for immediate transfer to new domains with an *extremely* low number of domain examples – makes sense from a scientific POV (humans can do it, why can’t AI?), but strikes me as misguided in applications.

Tiny data is rare in applied work, both because products generate data passively – and because *if* a task might be profitable, *then* it’s worth paying an expert to sit down for a day or two and crank out ~1K annotations for supervised learning. And with modern NLP like ELMo and BERT, ~1K is really enough!

It's worth noting that most of the [superGLUE tasks](#) have <10K train examples, with several having only a few hundred. (This is a “low-data regime” relative to the expectations of the recent past, but a regime where you can now get good results with a brainless cookie-cutter finetuning approach, in superGLUE as in the rest of life.)

Corpus	Train	Dev	Test
BoolQ	9427	3270	3245
CB	250	57	250
COPA	400	100	500
MultiRC	5100	953	1800
ReCoRD	101k	10k	10k
RTE	2500	278	300
WiC	6000	638	1400
WSC	554	104	146

GPT-3 few-shot can perform competitively on some of these tasks while pushing that number down to 32, but at the cost of many downsides, unknowns, and flexibility limitations. Which do you prefer: taking on all those risks, or sitting down and writing out a few more examples?

The trajectory of my work in data science, as it happens, looks sort of like a move from few-shot-like approaches toward finetuning approaches.

My early applied efforts assumed that I would never have the kind of huge domain-specific corpus needed to train a model from scratch, so I tried to compose the output of many SOTA models on more general domains. And this ... worked out terribly. The models did exactly what they were trained to do, not what I wanted. I had no way to scale, adapt or tune them; I just accepted them and tried to work around them.

Over time, I learned the value of doing *exactly* what you want, not something close to it. I learned that a little bit of data in your *actual* domain, specifying your *exact* task, goes much further than any domain-general component. Your applied needs will be oddly shaped, extremely specific, finicky, and narrow. You rarely need the world’s greatest model to accomplish them – but you need a model with access to a *very precise specification of exactly what you want*.

One of my proudest ML accomplishments is a system that does something very domain-specific and precisely shaped, using LM-pretrained components plus supervised learning on

~1K of my own annotations. *Sitting down and personally churning out those annotations must have been some of the most valuable time I have ever spent at work, ever.*

I wanted something specific and finicky and specialized to a very particular use case. So I sat down and specified what I wanted, as a long list of example cases. It took a few days ... and I am still reaping the benefits a year later.

If the few-shot users are working in domains anything like mine, they either know some clever way to evade this hard-won lesson, or they have not yet learned it.

---

But to the other question ... why are people so keen to apply GPT-3 few-shot learning in applications? This questions forks into "why do end users think this is a good idea?" and "why did OpenAI provide an API for doing this?"

I know some cynical answers, which I expect the reader can imagine, so I won't waste your time writing them out. I don't actually know what the non-cynical answers look like, and my ears are open.

*(For the record, all of this only applies to few-shot. OpenAI is apparently going to provide finetuning as a part of the API, and has already provided it to AI Dungeon. Finetuning a model with 175M parameters is a whole new world, and I'm very excited about it.*

*Indeed, if OpenAI can handle the costs of persisting and running finetuned GPT-3s for many clients, all of my concerns above are irrelevant. But if typical client use of the API ends up involving a finetuning step, then we'll have to revisit the GPT-3 paper and much of the ensuing discussion, and ask when - if not now - we actually expect finetuning to become obsolete, and what would make the difference.)*

# Epistemic Comparison: First Principles Land vs. Mimesis Land

*Epistemic Status: wildly speculative*

I've been thinking about the ideas of mimesis and cultural learning by [René Girard](#) (known in part for being [recommended by Peter Thiel](#), explanation [here](#)) and Joseph Henrich ([Secrets of our Success](#)). These seem like both profound, though vastly unexplored ideas. René Girard had a few bold ideas on mimesis, but seemed to focus more on scapegoats and religious details. Joseph Henrich catalogued lots of specific evidence, but didn't really investigate the broader implications.<sup>[1]</sup>

There are obvious possibilities of these ideas to our epistemics. If it is true that humans copy their ideas from each other instead of deriving them based on reason, often without knowing it, this would have vast implications on our beliefs.

To understand these ideas better, I constructed two hypothetical and extremized societies. The first is "first principles land", where everyone constructs their beliefs individually using first principles. These people act basically as ideal Bayesian agents and approximate gears-levels understandings of everything. This is infinitely difficult to do in practice, but simple to reason about in abstract. The second is "Mimesis Land", where people copy their beliefs from those they consider successful, typically without realizing it (this bit taken from Girard). To make "Mimesis Land" work at all, I'd posit that there is some amount of experimentation of techniques and beliefs happening at all times.

I then compared these two hypothetically societies based on how I expected each to do on 17 various attributes.

[See the document here.](#)

Needless to say, Mimesis Land was more interesting to me than First Principles Land.

One tricky thing is that neither land was trying to model real humans; but humans that followed my simple view of a version of a set of extreme theories that might be somewhat coherent. There was a fair bit of subjectivity in this. I'm sure others would come up with significantly different versions of First Principles Land and Mimesis Land. As such I'm not particularly satisfied with this methodology, but at the same time I'm not sure what procedures would be strictly better.

I am bullish on trying to more clearly formulate what first principles thinking and mimesis fully look like in practice, and the costs and benefits of both. In both cases there seem to be fairly patchy terminology of several clusters discussing them without all too much structure. Many thanks to LessWrong for some of my original thoughts on both so far, but there's still a lot of work left.

If I were to continue on a similar path, I would consider adding several other worlds. First Principles Land may get a less extreme version. The Mimesis Land here assumes that people copy the most successful people, and do so mostly unknowingly. It would be interesting to try variants where they copy others (for instance, the most credible people), and do so intentionally (so they are fully aware that their beliefs are inconsistent and often likely to be wrong.) There could be "Chesterton's Fence Land",

where traditional beliefs are intentionally chosen. There could be "Forecaster Land", where everyone delegates all thoughts to a prediction market.

I'm really curious what readers may think of this methodology, and this piece in general. Comments highly appreciated.

---

### **Disclaimers of things I don't believe:**

1. My definitions of both lands are the "correct" definitions. -> I made a lot of subjective choices in each case and could imagine plenty of alternative ideas.
2. First Principles Land is "better" than Mimesis Land. -> First Principles Land has a bunch of advantage, but is much less possible, and also have some significant disadvantages.
3. Western society is just like Mimesis Land -> I think our society has similarities to both lands. There are probably more similarities to Mimesis Land, but there could also be similarities to other lands not yet described.
4. I (Ozzie) am in First Principles Land -> I'm sure I have many beliefs copied from others I haven't noticed. I noticed several, and copy many practices from others without understanding them at all. (For playful instance, I celebrate some holidays, while having no first principles understanding of the optimality of the specific practices.)
5. I (Ozzie) am understanding "Mimesis" correctly -> I'm sure I'm making mistakes. I haven't read Girard's books, just watched a few lectures and read summaries. From what I understand, he's quite dense.
6. First Principles Land and Mimesis Land are two sides to a spectrum, and all reasonable worlds are exactly in between these two -> These two lands are both extremes, but there are several important axes not covered by them. Our world is generally somewhere in between them in the areas where they are opposed, but also has several other separate characteristics.

[1] Social constructivism and has a much more robust literature. To be honest I'm quite rusty on this and need to catch up some time. Interestingly, I don't remember reading about connections between this and Girard or Henrich, though they seem to share a lot of beliefs.

*Thanks to Brangus for discussion on the ideas that helped lead to this final post.*

# 10/50/90% chance of GPT-N Transformative AI?

In [Developmental Stages of GPTs](#), orthonormal explains why we might need "zero more cognitive breakthroughs" to reach transformative AI, aka "an AI capable of making an Industrial Revolution sized impact". More specifically, he says that "basically, GPT-6 or GPT-7 might do it".

Besides, [Are we in an AI overhang?](#) makes the case that "GPT-3 is the trigger for 100x larger projects at Google, Facebook and the like, with timelines measured in months."

Now, assuming that "GPT-6 or GPT-7 might do it" and that timelines are "measured in months" for 100x larger projects, **in what year will there be a 10% (resp. 50% / 90%) chance of having transformative AI and what would the 50% timeline look like?**

# "Good judgement" and its components

*Epistemic status: Sharing a personal ontology I've found useful. I expect most claims to be fairly uncontroversial, but maybe the perspective will be interesting/clarifying for some readers. ([Cross-posted](#), except for this paragraph, from the EA forum.)*

**Good judgement** is about mental processes which tend to lead to good decisions. (I think good decision-making is centrally important for longtermist EA, for reasons I won't get into here.) Judgement has two major ingredients: **understanding of the world**, and **heuristics**.

**Understanding of the world** helps you make better predictions about how things are in the world now, what trajectories they are on (so how they will be at future points), and how different actions might have different effects on that. This is important for helping you explicitly think things through. There are a number of sub-skills, like **model-building**, having **calibrated estimates**, and just **knowing relevant facts**. Sometimes understanding is held in terms of implicit predictions (perhaps based on experience). How good someone's understanding of the world is can vary a lot by domain, but some of the sub-skills are transferrable across domains.

You can improve your understanding of the world by learning foundational facts about important domains, and by practicing skills like model-building and forecasting. You can also improve understanding of a domain by importing models from other people, although you may face challenges of being uncertain how much to trust their models. (One way that models can be useful without requiring any trust is giving you clues about where to look in building up your own models.)

**Heuristics** are rules of thumb that you apply to decisions. They are usually held implicitly rather than in a fully explicit form. They make statements about what properties of decisions are good, without trying to provide a full causal model for why that type of decision is good. Some heuristics are fairly general (e.g. "avoid doing sketchy things"), and some apply to specific domains (e.g. "when hiring programmers, put a lot of weight on the coding tests").

You can improve your heuristics by paying attention to your experience of what worked well or poorly for you. Experience might cause you to generate new candidate heuristics (explicitly or implicitly) and hold them as hypotheses to be tested further. They can also be learned socially, transmitted from other people. (Hopefully they were grounded in experience at some point. Learning can be much more efficient if we allow the transmission of heuristics between people, but if you don't require people to have any grounding in their own experience or cases they've directly heard about, it's possible for heuristics to be propagated without regard for whether they're still useful, or if the underlying circumstances have changed enough that they shouldn't be applied. Navigating this tension is an interesting problem in social epistemology.)

One of the reasons that it's often good to spend time with people with good judgement is that you can make observations of their heuristics in action. Learning heuristics is difficult from writing, since there is a lot of subtlety about the boundaries of when they're applicable, or how much weight to put on them. To learn from other people (rather than your own experience) it's often best to get a chance to interrogate

decisions that were a bit surprising or didn't quite make sense to you. It can also be extremely helpful to get feedback on your own decisions, in circumstances where the person giving feedback has high enough context that they can meaningfully bring their heuristics to bear.

**Good judgement generally wants a blend of understanding the world and heuristics.** Going just with heuristics makes it hard to project out and think about scenarios which are different from ones you've historically faced. But our ability to calculate out consequences is limited, and some forms of knowledge are more efficiently incorporated into decision-making as heuristics rather than understanding about the world.

One kind of judgement which is important is **meta-level judgement** about how much weight to put on different perspectives. Say you are deciding whether to publish an advert which you think will make a good impression on people and bring users to your product, but contains a minor inaccuracy which would require much more awkward wording to avoid. You might bring to bear the following perspectives:

- A) The **heuristic** "don't lie"
- B) The **heuristic** "have snappy adverts"
- C) The **implicit model** which is your gut prediction of what will happen if you publish
- D) The **explicit model** about what will happen that you drew up in a spreadsheet
- E) The **advice** of your partner
- F) The **advice** of a professional marketer you talked to

Each of these has something legitimate to contribute. The choice of how to reach a decision is a **judgement**, which I think is usually made by choosing **how much weight** to put on the different perspectives in this circumstance (including sometimes just letting one perspective dominate). These weights might in turn be informed by your understanding of the world (e.g. "marketers should know about this stuff"), and also by your own experience ("wow, my partner always seems to give good advice on these kinds of tricky situations").

I think that almost always **the choice of these weights is a heuristic** (and that the weights themselves are generally implicit rather than explicit). You could develop understanding of the world which specify how much to trust the different perspectives, but as boundedly rational actors, at some point we *have* to get off the understanding train and use heuristics as shortcuts (to decide when to spend longer thinking about things, when to wrap things up, when to make an explicit model, etc.).

Overall I hope that people can develop good object-level judgement in a number of important domains (strategic questions seem particularly tricky+important, but judgement about technical domains like AI, and procedural domains like how to run organisations also seem very strongly desirable; I suspect there's a long list of domains I'd think are moderately important). I also hope we can develop (and support people to develop) good meta-level judgement. When decision-makers have good meta-level judgement this can act as a force-multiplier on the presence of the best accessible object-level judgement in the epistemic system. It can also add a kind of robustness, making badly damaging mistakes quite a lot less likely.

# [AN #112]: Engineering a Safer World

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

Alignment Newsletter is a weekly publication with recent content relevant to AI alignment around the world. Find all Alignment Newsletter [resources here](#). In particular, you can look through [this spreadsheet](#) of all summaries that have ever been in the newsletter.

Audio version [here](#) (may not be up yet).

I recently read [Engineering a Safer World](#) by Nancy G. Leveson, at [Joshua Achiam's recommendation](#), and really enjoyed it, so get ready for another book summary! I'm not very happy with the summary I have -- it feels less compelling than the book, partly because the book provides a ton of examples that I don't have the space to do -- but hopefully it is enough to get the key points across.

The main motivation of this book is to figure out how we can improve safety engineering. Its primary thesis is that the existing methods used in engineering are insufficient for the current challenges, and must be replaced by a method the author favors called STAMP. Note that the book is primarily concerned with mechanical systems that may also have computerized automation (think aerospace, chemical, and mechanical engineering); the conclusions should not be expected to apply directly to AI.

## **The standard model of safety engineering and its deficiencies**

Historically, safety engineering has been developed as a reaction to the high level of accidents we had in the past, and as a result focused on the easiest gains first. In particular, there were a lot of gains to be had simply by ensuring that *machines didn't break*. (*Rohin's note: I'm editorializing a bit here, the author doesn't explicitly say this but I think she believes it.*) This led to a focus on *reliability*: given a specification for how a machine should operate, we aim to decrease the probability that the machine fails to meet that specification. For example, the specification for a water tank would be to contain the water up to a given pressure, and one way to improve the reliability of the tank would be to use a stronger material or make a thicker tank to make it less likely that the tank ruptures.

Under this model, an accident happens when a machine fails to meet its specification. So, we can analyze the accident by looking at what went wrong, and tracing back the physical causes to the first point at which a specification was not met, giving us a *root cause* that can show us what we need to fix in order to prevent similar accidents in the future. We can call this sort of analysis an *event chain* analysis.

However, in the last few decades there have been quite a few changes that make this model worse than it once was. The pace of technological change has risen, making it harder to learn from experience. The systems we build have become complex enough that there is a lot more *coupling* or interaction effects between parts of the system that we could fail to account for. Relatedly, the risks we face are getting large enough that we aren't willing to tolerate even a *single* accident. Human operators (e.g. factory workers) are no longer able to rely on easily understood and predictable mechanical systems, instead having to work with computerized automation which they cannot

understand as well. At this point, event chain analysis and safety-via-reliability are no longer sufficient for safety engineering.

Consider for example the [Flight 965](#) accident. In this case, the pilots got clearance to fly towards the Roko waypoint in their descent, listed as (R) on their (paper) approach charts. One of the pilots pressed R in the flight management system (FMS), which brought up a list of waypoints that did *not* include Roko, and executed the first one (presumably believing that Roko, being the closest waypoint, would show up first). As a result, the plane turned towards the selected waypoint, and crashed into a mountain.

The accident report for this incident placed the blame squarely on the pilots, firstly for not planning an appropriate path, and secondly for not having situational awareness of the terrain and that they needed to discontinue their approach. But most interestingly, the report blames the pilots for not reverting to basic radio navigation when the FMS became confusing. The author argues that the design of the automation was also flawed in this case, as the FMS stopped displaying the intermediate fixes to the chosen route, and the FMS's navigational information used a different naming convention than the one in the approach charts. Surely this also contributed to the loss? In fact, in lawsuit appeals, the software manufacturer was held to be 17% liable.

However, the author argues that this is the exception, not the rule: typically event chain analysis proceeds until a human operator is found who did something unexpected, and then the blame can safely be placed on them. Operators are expected to "use common sense" to deviate from procedures when the procedures are unsafe, but when an accident happens, blame is placed on them for deviating from procedures. This is often very politically convenient, and is especially easy to justify thanks to hindsight bias, where we can identify exactly the right information and cues that the operator "should have" paid attention to, ignoring that in the moment there were probably many confusing cues and it was far from obvious which information to pay attention to. My favorite example has to be this quote from an accident report:

"Interviews with operations personnel did not produce a clear reason why the response to the [gas] alarm took 31 minutes. The only explanation was that there was not a sense of urgency since, in their experience, previous [gas] alarms were attributed to minor releases that did not require a unit evacuation."

It is rare that I see such a clear example of a self-refuting paragraph. In the author's words, "this statement is puzzling, because the statement itself provides a clear explanation for the behavior, that is, the previous experience". It definitely sounds like the investigators searched backwards through the causal chain, found a situation where a human deviated from protocol, and decided to assign blame there.

This isn't just a failure of the accident investigation -- the entire premise of some "root cause" in an event chain analysis implies that the investigators must end up choosing some particular point to label as The Root Cause, and such a decision is inevitably going to be determined more by the particular analysts involved rather than by features of the accident.

## Towards a new approach

How might we fix the deficiencies of standard safety engineering? The author identifies several major changes in assumptions that are necessary for a new

approach:

1. Blame is the enemy of safety. Safety engineering should focus on system behavior as a whole, where interventions can be made at many points on different levels, rather than seeking to identify a single intervention point.
2. Reliability (having machines meet their specifications) is neither necessary nor sufficient for safety (not having bad outcomes). Increased reliability can lead to decreased safety: if we increase the reliability of the water tank by making it using a stronger material, we may decrease the risk of rupture, but we may dramatically increase the harm when a rupture occurs since the water will be at a much higher pressure. This applies to software as well: highly reliable software need not be safe, as its specifications may not be correct.
3. Accidents involve the entire sociotechnical system, for which an event chain model is insufficient. Interventions on the sociological level (e.g. make it easy for low-level operators to report problems) should be considered part of the remit of safety engineering.
4. Major accidents are not caused by simultaneous occurrence of random chance events. Particularly egregious examples come from probabilistic risk analysis, where failures of different subsystems are often assumed to be independent, neglecting the possibility of a common cause, whether physical (e.g. multiple subsystems failing during a power outage) or sociological (e.g. multiple safety features being disabled as part of cost-cutting measures). In addition, systems tend to migrate towards higher risk over time, because environmental circumstances change, and operational practices diverge from the designed practices as they adapt to the new circumstances, or simply to be more efficient.
5. Operator behavior is a product of the environment in which it occurs. To improve safety, we must change the environment rather than the human. For example, if an accident occurs and an operator didn't notice a warning light that could have let them prevent it, the solution is not to tell the operators to "pay more attention" -- that approach is doomed to fail.

### A detour into systems theory

The new model proposed by the author is based on systems theory, so let's take a moment to describe it. Consider possible systems that we may want to analyze:

First, there are some systems with *organized simplicity*, in which it is possible to decompose the system into several subsystems, analyze each of the subsystems independently, and then combine the results relatively easily to reach overall conclusions. We might think of these as systems in which analytic reduction is a good problem-solving strategy. *Rohin's note: Importantly, this is different from the philosophical question of whether there exist phenomena that cannot be reduced to e.g. physics: that is a question about whether reduction is in principle possible, whereas this criterion is about whether such reduction is an effective strategy for a computationally bounded reasoner.* Most of physics would be considered to have organized simplicity.

Second, there are systems with *unorganized complexity*, where there is not enough underlying structure for analytic reduction into subsystems to be a useful tool. However, in such systems the behavior of individual elements of the system is sufficiently random (or at least, well-modeled as random) that statistics can be

applied to it, and then the law of large numbers allows us to understand the system as an aggregate. A central example would be statistical mechanics, where we cannot say much about the motion of individual particles in a gas, but we can say quite a lot about the macroscopic behavior of the gas as a whole.

Systems theory deals with systems that have *organized complexity*. Such systems have enough organization and structure that we cannot apply statistics to it (or equivalently, the assumption of randomness is too incorrect), and are also sufficiently complex that analytic reduction is not a good technique (e.g. perhaps any potential decomposition into subsystems would be dominated by combinatorially many interaction effects between subsystems). Sociological systems are central examples of such systems: the individual components (humans) are very much not random, but neither are their interactions governed by simple laws as would be needed for analytic reduction. While systems theory cannot provide nearly the same level of precision as statistics or physics, it does provide useful concepts for thinking about such systems.

The first main concept in systems theory is that of *hierarchy and emergence*. The idea here is that systems with organized complexity can be decomposed into several hierarchical levels, with each level built “on top of” the previous one. For example, companies are built on top of teams which are built on top of individual employees. The behavior of components in a particular layer is described by some “language” that is well-suited for that layer. For example, we might talk about individual employees based on their job description, their career goals, their relationship with their manager, and so on, but we might talk about companies based on their overall direction and strategy, the desires of their customer base, the pressures from regulators, and so on.

*Emergence* refers to the phenomenon that there can be properties of higher levels arising from lawful interactions at lower levels that nonetheless are meaningless in the language appropriate for the lower levels. For example, it is quite meaningful to say that the pressures on a company from government regulation caused them to (say) add captions to their videos, but if we look at the specific engineer who integrated the speech recognition software into the pipeline, we would presumably say “she integrated the speech recognition into the pipeline because she had previously worked with the code” rather than “she integrated it because government regulations told her to do so”. As another example, safety is an emergent system property, while reliability is not.

The second main concept is that of *control*. We are usually not satisfied with just understanding the behavior of systems; we also want to make changes to it (as in the case of making them safer). In systems theory, this is thought of as *control*, where we impose some sort of *constraint* on possible system behavior at some level. For example, employee training is a potential control action that could aim to enforce the constraint that every employee knows what to do in an emergency. An effective controller requires a goal, a set of actions to take, a model of the system, and some way to sense the state of the system.

### **STAMP: A new model underlying safety engineering**

The author then introduces a new model called Systems-Theoretic Accident Model and Processes (STAMP), which aims to present a framework for understanding how accidents occur (which can allow us to prevent them and/or learn from them). It contains three main components:

**Safety constraints:** In systems theory, a constraint is the equivalent of a specification, so these are just the safety-relevant specifications. Note that such specifications can be found at all levels of the hierarchy.

**Hierarchical safety controllers:** We use *controllers* to enforce safety constraints at any given level. A control algorithm may be implemented by a mechanical system, a computerized system, or humans, and can exist at any level of the hierarchy. A controller at level N will typically depend on constraints at level N - 1, and thus the design of this controller influences which safety constraints are placed at level N - 1.

**Process models:** An effective controller must have a model of the process it is controlling. Many accidents are the result of a mismatch between the actual process and the process model of the controller.

This framework can be applied towards several different tasks, and in all cases the steps are fairly similar: identify the safety constraints you want, design or identify the controllers enforcing those constraints, and then do some sort of generic reasoning with these components.

If an accident occurs, then at the highest level, either the control algorithm(s) failed to enforce the safety constraints, or the control actions were sent correctly but were not followed. In the latter case, the controllers at the lower level should then be analyzed to see why the control actions were not followed. Ultimately, this leads to an analysis on multiple levels, which can identify several things that went wrong rather than one Root Cause, that can all be fixed to improve safety in the future.

## **Organizational safety**

So far we've covered roughly chapters 1-4 of the book. I'll now jump straight to chapter 13, which seems particularly important and relevant, as it deals with how organizational structure and management should be designed to support safety.

One major point that the author makes is that safety *is* cost-effective for *long-term* performance as long as it is designed into the system from the start, rather than added on at the last minute. Performance pressure on the other hand inevitably leads to cuts in safety.

In order to actually get safety designed into the system from the start, it is crucial that top management demonstrates a strong commitment to safety, as without this employees will inevitably cut corners on safety as they will believe it is in their incentives to do so. Other important factors include a concrete corporate safety policy, as well as a strong corporate safety culture. It is important that safety is part of the design process, rather than tacked on at the end. In the author's words, *putting safety into the quality assurance organization is the worst place for it. [...] It sets up the expectation that safety is an after-the-fact or auditing activity only.*

In addition, it is important that information can flow well. Going from the bottom to the top, it should be possible for low-level operators to report potential problems in a way that they are actually acted on and the relevant information reaches top management. From the top to the bottom, safety information and training should be easily available and accessible to employees when they need it.

It is also important to have controls to prevent the general tendency of systems to migrate towards higher risk, e.g. by relaxing safety requirements as time passes without any incidents. The next chapter describes SUBSAFE, the author's example of a

well-run safety program, in which the control is for everyone to periodically watch a video reminding them of the importance of their particular safety work (in particular, the video shows the loss of the USS Thresher, an event that caused SUBSAFE to be created).

Perhaps obviously, it is important for an organization to have a dedicated safety team. This is in contrast to making everyone responsible for safety. In the author's words: *While, of course, everyone should try to behave safely and to achieve safety goals, someone has to be assigned responsibility for ensuring that the goals are achieved.*

If you start by designing for safety, it is cost-effective, not opposed to long-term money-maximizing. Once there is performance pressure, then you see cuts in safety. Also sometimes people fix symptoms instead of underlying causes, and then they just keep seeing symptoms forever and conclude they are inevitable.

### **Miscellaneous notes**

The remaining chapters of the book apply STAMP in a bunch of different areas with many examples, including an entire chapter devoted to the STAMP treatment of a friendly fire accident. I also really liked the discussion of human factors in the book, but decided not to summarize it as this has already gotten quite long.

### **Summary of the summary**

I'll conclude with a quote from the book's epilogue:

*What seems to distinguish those experiencing success is that they:*

1. *Take a systems approach to safety in both development and operations*
2. *Have instituted a learning culture where they have effective learning from events*
3. *Have established safety as a priority and understand that their long-term success depends on it*

### **Relationship to AI safety**

A primary motivation for thinking about AI is that it would be very impactful for our society, and very impactful technologies need not have good impacts. "Society" clearly falls into the "organized complexity" class of systems, and so I expect that the ideas of safety constraints and hierarchical control algorithms will be useful ways to think about possible impacts of AI on society. For example, if we want to think about the possibility of AI systems differentially improving technical progress over "wisdom", such that we get dangerous technologies before we're ready for them, we may want to sketch out hierarchical "controllers" at the societal level that could solve this problem. Ideally these would eventually turn into constraints on the AI systems that we build, e.g. "AI systems should report potentially impactful new technologies to such-and-such committee". I see the AI governance field as doing this sort of work using different terminology.

Technical AI alignment (in the sense of [intent alignment \(AN #33\)](#)) does not seem to benefit as much from this sort of an approach. The main issue is that we are often considering a fairly unitary system (such as a neural net, or the mathematical model of expected utility maximization) to which the hierarchical assumption of systems theory does not really apply.

To be clear, I *do* think that there in fact is some hierarchy. For example, in image classifiers where low levels involve edge detectors while high levels involve dog-face detectors. However, we do not have the language to talk about these hierarchies, nor the algorithms to control the intermediate layers. While [Circuits \(AN #111\)](#) is illustrating this hierarchy for image classifiers, it does not give us a language that we can (currently) use to talk about advanced AI systems. As a result, we are reduced to focusing on the incentives we provide to the AI system, or speculating on the levels of hierarchy that might be internal to advanced AI systems, neither of which seem particularly conducive to good work.

In the language of this book, I work on intent alignment because I expect that the ability to enforce the constraint “the AI system tries to do what its operator wants” will be a very useful building block for enforcing whatever societal safety constraints we eventually settle on, and it seems possible to make progress on it today. There are several arguments for risk that this ignores (see e.g. [here \(AN #50\)](#) and [here \(AN #103\)](#)); for some of these other risks, the argument is that we can handle those using similar mechanisms as we have before (e.g. governance, democracy, police, etc), as long as we have handled intent alignment.

## FEEDBACK

I'm always happy to hear feedback; you can send it to me, [Rohin Shah](#), by [replying to this email](#).

## PODCAST

An audio podcast version of the **Alignment Newsletter** is available. This podcast is an audio version of the newsletter, recorded by [Robert Miles](#).

# Split-a-Dollar Game

When researching the topic of [distributing government seats in Switzerland](#). I've chanced on the so-called split-the-dollar game, which is kind of simplified, game-theoretical model of the process. Given that despite having some basic knowledge of the game theory I have never heard of the game, it may be worth making a short blog post about it.

In the game, three players are asked to vote on splitting a dollar. The game is iterative, it has multiple rounds. At the beginning, a third of a dollar may be assigned to each player. The players, unanimously, vote for the arrangement.

However, the arrangement is not stable. It's a game about forming coalitions. Two players can conspire and vote in a arrangement where they split the dollar fairly among themselves and give nothing to the third player.

But even then the arrangement isn't stable. Now the third player can approach one of the other two coalition parties and propose a better deal: Instead of 50 cents they would get 75 cents and the proposing party would content itself with just 25 cents. Their current coalition partner would get nothing. The party is incentivized to accept not only because of sheer greed, but also because it is aware that if they rejected the proposal, the proposing party can make the same offer to their coalition partner and thus lock them out of the coalition.

But even now, with two players sharing the dollar in 75:25 ratio, the arrangement is not stable. The senior partner can blackmail the junior partner by threatening to make a coalition with the opposition party. The junior partner, in their turn, can do exactly the same thing.

As can be seen, any arrangement is, in principle, unstable. The negotiation game can go on forever without reaching a stable state.

We can see the game in action in the case of Switzerland. In 1953, despite FDP (liberals), CVP (christian democrats) and SP (social democrats) getting very much the same amount of votes, the seats in government were distributed among them in 4:2:0 ratio. (There was also one seat going to SVP, which matched their election result and is not relevant to this story.)

The interesting aspect here is that unlike in other countries with the same problem, Switzerland has been able - after more than 100 years of playing the game - to reach a stable state. They did so by adopting a cultural norm, so called "Magic Formula", which splits the seats in government in a fashion proportional to election results.

Note that the game still doesn't have a stable strategy. The problem wasn't solved within the game. It was solved on the meta-level, by parties agreeing to respect the voting results, internalizing that respect and punishing those who don't behave.

Also note that this haven't in fact solved the problem, but rather moved it, commendably so, to a different level. Today the split-the-dollar game isn't played among parliamentarians forming a government, but rather among members of government voting on practical issues.

**August 24th, 2020**

# Is Wirecutter still good?

I have heard from several friends I trust that [Wirecutter](#) is no longer very reliable since being acquired by the New York Times in 2016, and that their Wirecutter-advised purchases have become pretty mediocre in the last year or two.

I'd be interested in people making some of that case, as answers to this post, including things like talking about purchases they made or basic errors they found in the reviews, as I don't have strong, publicly verifiable evidence on this at the minute.

This is a biased post, I'm writing this with the hope of helping propagate this info if it's true, which I suspect it is but am not confident of. Please give whatever answers you feel are most relevant.

# Multitudinous outside views

There's an important piece of advice for forecasters: don't rely on your internal model of the world exclusively, and take the outside view, then adjust from there. But which view is "the" outside view? It depends on the problem - and different people might tell you different things. But if the choice of outside view is subjective, it starts to seem like inside-views all the way down.

That's where we get to base rates, which don't solve this problem, but they do highlight it nicely.

---

Fans of superforecasting know, in a hedgehog-like sense of knowing one thing, that the outside view, which is the base rate, which is the rate of similar events, should be our starting point. But which events are similar, and how is similarity defined? We first need to choose a reference class, based on some pre-existing idea of similarity. And in different terms, there is a [reference class problem](#), which we evidently don't have a clear way to judge - and even as Bayesian thinkers, [not only is that our problem](#), it's an entire bucket of different problems.

## Considering a Concrete Prediction: Tesla Motors

Let's get really concrete: What will the price of Tesla stock be in 6 months?

Well, what is the reference class? In the last year, 90% of the time, the price of Tesla stock has been between \$200 and \$1000. But that's a really bad reference class, when the price today is \$1,800. OK, but looking at the set of all stocks would be even worse - and looking at automobile stocks even worse than that. Which stocks are comparable? What about stocks with P/E ratios over 900? Or stocks with more than a half billion dollars of losses for their net income? We're getting silly here.

Maybe we shouldn't look at stock price, but should look at market capitalization? Or change in price? "Stocks that went up 9-fold over the course of a year" isn't a super helpful reference class - it has only a few examples, and they are all very different from Tesla.

Of course, none of this is helpful. What we really want is the aggregate opinion of the market, so we look at futures contracts and the implied volatility curve for options expiring in February.

That doesn't look like a reference class. But who needs an outside view, anyways?

## What is a reference class?

If you want to know the probability of Kim Jung-Un staying alive, we can consult the reference class of 37 year old males in North Korea, where male life expectancy is 68. Alternatively, look at the reference class of his immediate family - his brother died at

the age of 46, but his father lived to the age of 70, and his grandfather lived until 82. Are those useful reference points?

What we *really* want is the lifespan of dictators. Well, dictators of small countries. Oh, actually, dictators of small nuclear powers that know that Qaddafi was killed after renouncing his nuclear program - a reference class with no other members. Once again, of course, none of this is helpful.

In finance, the outside view is a consensus that markets are roughly rational, and the inside view is that you can beat the market. In international relations, the outside view is that dictatorships can be tenuous, but when the regime survives, the leadership lives quite a long time. The inside view is, perhaps, that China has a stake in keeping their nuclear neighbor stable, and won't let anything happen.

## Reference classes depend on models of the world.

In each case, the construction of a reference class is a function of a model. Models induce reference classes - political scientists might have expert political judgement, while demographers have expert lifespan judgement, and 2nd year equity analysts have expert financial judgement. All of those are useful.

What reference class should have been used for COVID-19 in, say, mid-March? The set of emerging infectious diseases over the past decade? Clearly not. [In retrospect](#), of course, the best reference class needed a epidemiological model - the reference class of diseases with  $R_0 > 1$ , where spread is determined by control measures. And the reference class for the success of response in the US should have been based on a libertarian view of the failure of American institutions, or a Democrat's view of how Trump had been rapidly dismantling government, and not [an index designed around earlier data which ignored political failure modes](#). But how do we know that in advance? Once again, none of this is helpful in deciding beforehand which reference class to use.

A final example. What reference class is useful for predicting the impact of artificial intelligence over the next decade? Robin Hanson would argue, I think, that it's the reference class of purported game-changing technologies that have not yet attracted significant amounts of capital investment. Eliezer Yudkowsky might argue that it's the reference class of intelligence evolving, sped up by a factor of what we've seen so far of computer intelligence, which moved from an AI winter in the mid-2000s and anti-level intelligence at navigation, to Deepmind being founded in 2010, to IBM's Watson winning Jeopardy in 2011, to beating the Winograd Schema and acing general high-school science tests without specific training using GPT-3 now. And if you ask a dozen AI researchers, [depending on your methods](#), you'll get at least another dozen reference classes. But we still need to pick a reference class.

So which reference class is correct? In my (inside) view as a superforecaster, this is where we turn to a different superforecasting trick, about considering multiple models. As the saying goes, hedgehogs know one reference class, but foxes consult many hedgehogs.

# Property as Coordination Minimization

A friend recently noted that they were in favor of private property, but the best defense they had to link was instead a [defense of finance](#). So I thought I'd give it a try. In light of a distinction people often draw between '[private property](#)' and '[personal property](#)', I'm going to work up to defending 'impersonal private property', starting with intuitions and examples grounded in personal property.

First, what even do we mean by property? To begin, observe that material things are sometimes scarce or rivalrous. If I eat a sandwich, you can't also eat it; if I sleep in a bed, you can't also sleep in it at the same time; if an acre of land is rented out for agricultural use, only one of us can collect the rent check. Beyond scarcity inherent to reality, we can also create it with rules; if I invent a cool new sandwich idea, society could decide that I have the right to decide who can and can't make that sandwich for some amount of time. [The first patents were for restaurants, giving them exclusive rights for a year to new dishes they invented.]

Property, then, is the societally recognized right to decide who can or can't do three different things: 'use', or deriving some personal benefit from the thing; 'fruit', or extracting some value from the thing without deeply changing it; and 'abuse', or making changes or transferring ownership of the thing. For example, consider an apple tree; climbing the tree is an example of use, picking the apples is an example of fruit, and chopping it down to make a chair out of it is an example of abuse. If I wholly own the tree, I get to choose all of those things; perhaps I just have [usufruct](#), allowing me to use it or harvest the fruit so long as I don't damage it; perhaps I only have use, or perhaps I have no rights at all with regards to the tree.

In this view, the benefit of property is fundamentally *preventative*; if I own an apple tree, I can prevent other people from climbing the tree, or picking the apples, or chopping down the tree, even if they want to. Hence the slogan that 'property is theft'; without it, you could do any of those things to my tree, and with property, you can't. Also note how much joint ownership degrades in value; if everyone has the right to chop down the tree to make a chair, the fact that anyone could collect the fruit once it ripens becomes less valuable.

Interestingly, this view also makes NIMBYism seem natural instead of unnatural. If I own a house, I can use that ownership to prevent things from happening to the house that I don't want. But do I just own the dirt and wood, or do I also own the ambient level of noise? The fragrance of the air? The view? The price? We can make our conception of property too large or too small, and can start drawing overlapping property claims, where I think my ownership of my house means no loud music on my property at 6am, whereas my neighbor thinks that his ownership of his house means he can practice the drums whenever he likes. [This sort of coordination is best done at a higher level, through ownership of the neighborhood or zoning district or city or whatever.]

This brings up the idea of 'stakeholders' and 'decision-makers'. Stakeholders are those impacted by the outcome, and decision-makers are those who choose the outcome. Often, we get more desirable or just results by aligning the decision-makers and stakeholders, but this comes at additional coordination costs.

Suppose I'm ordering dinner for a group of people; there's both the coordination question of which restaurant to order from, and the coordination question of what dishes to order. Sometimes it works for me to just pick a restaurant and dishes; sometimes it works for me to pick a restaurant, and then pass around the order for everyone to add their preferred dish; sometimes it works to jointly come to a decision on what restaurant to order from, and then everyone selects a dish; sometimes it works for everyone to manage their own order, including whether or not they should join in on an order with anyone else. That list was ordered roughly in 'decreasing coordination cost' order, with a corresponding increase in taste-satisfaction, but perhaps not *net* satisfaction, as smaller orders are more expensive, or the additional taste benefits weren't worth the additional benefits of having to think about it. The size of the group has a huge impact on how much the coordination costs matter; coming to alignment on a restaurant for three people and thirty people are very different affairs.

Why have personal property, i.e. your own sandwich, toothbrush, clothes, house, or vehicle? A boring but essential reason is physical; a toothbrush used by Alice becomes much less valuable to everyone but Alice after that use, and this sometimes applies more broadly. The main reason, in my view, is that personal property is made much more useful by only having one decision-maker, and thus no coordination cost. Rather than having to petition the commune for a day's use of a red shirt, I can simply decide to wear my red shirt today. I can make solid plans around decisions that I'm the only major input to. This will sometimes lead to socially suboptimal decisions if coordination were free--maybe I look really bad in red, and a wise commune would give me blue instead--but given that coordination is *not* free, this is often our best available option.

Why have impersonal property, i.e. a landlord who rents out houses, a company that owns factories, massive tracts of land owned by the same farm, a bank that chooses which loans to grant and which to deny? The same basic reason, I claim; the landlord can make decisions about the houses that they own without having to consult anyone else, and this means decisions can be made faster and more cheaply. Many different landlords can make many different decisions, whereas one Housing Bureau will either make one decision for everyone, or make unequal decisions in a corrupt way. Or if we had a property-less direct democracy, where all citizens voted on all decisions, there would be no time left over to do anything else but vote!

Many of the problems we have now, I claim, are not caused by *too much* property but by *too many decision-makers*, or in this view, *too little* property. For example, I live in Berkeley, which has a housing shortage, and also incomplete individual property rights. By that I mean if I buy a house and want to tear it down and build a larger one instead, I need the city's permission to do so, and the city will require me to allow 'public comment' from my neighbors and passerby on the desirability of such a change, and generally require various other permits and restrictions. If it were solely for the safety of the inhabitants, this could be handled by the building code, but the public comment isn't in case my neighbors happen to be structural engineers; it's because housing in Berkeley does *not* come with the full right of 'abuse', and that is instead owned by the neighborhood and city, and only some stakeholders get a say; the people who would rent out the additional floors I add to the house generally don't comment at the public meeting, whereas the retiree who would have to deal with more cars on the road or a blocked view of the Bay does.

Indeed, [It's Time to Build](#) is, in many ways, a complaint about the [Vetocracy](#) of our times. Property, even impersonal property, even the existence of billionaires even if

you'll never be one, are good because it lowers coordination costs, allowing things to happen more efficiently.

# Strong implication of preference uncertainty

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

Here is a theory that is just as good as general relativity:

AGR (Angel General Relativity): Tiny invisible angels push around all the particles in the universe in a way that is indistinguishable from the equations of general relativity.

This theory is falsifiable, just as general relativity (GR) itself is. Indeed, since it gives exactly the same predictions as GR, a Bayesian will never find evidence that prefers it over Einstein's theory.

Therefore, I obviously deserve a Nobel prize for suggesting it.

## Enter Occam's shaving equipment

Obviously the angel theory is not a revolutionary new theory. Partially because I've not done any of the hard work, just constructed a pointer to Einstein's theory. But, philosophically, the main justification is [Occam's razor](#) - the simplest theory is to be preferred.

From a Bayesian perspective, you could see violations of Occam's razor as cheating, using your posterior as priors. There is a whole class A of "angels are pushing particles" theories, and AGR is just a small portion of that space. By considering AGR and GR on equal footing, we're privileging AGR above what it deserves<sup>[1]</sup>.

## In physics, Occam's razor doesn't matter for strictly identical theories

Occam's razor has two roles: the first is to distinguish between strictly identical theories; the second is to distinguish between theories that give the same prediction on the data so far, but may differ in the future.

Here, we focus on the first case: GR and AGR are strictly identical; no data will ever distinguish them. In essence, the theory that one is right and the other wrong is not [falsifiable](#).

What that means is that, though AGR may be *a priori* less likely than GR, the relative probability between the two theories will never change: they make the same predictions. And also because they make the same predictions, that relative probability is irrelevant in practice: we could use AGR just as well as GR for predictions.

# How preferences differ

Now let's turn to preferences, as described in our paper "[Occam's razor is insufficient to infer the preferences of irrational agents](#)".

Here two sets of preferences are "prediction-identical", in the sense of the physics theories above, if they predict the same behaviour for the agent. So that means that two different preference-based explanations for the same behaviour will never change their relative probabilities.

Worse than that, Occam's razor doesn't solve the issue. The simplest explanations of, say, human behaviour, is that humans are fully rational at all times. This isn't the explanation that we want.

Even worse than that, prediction-identical preferences will lead to vastly different consequences if program an AI to maximise them.

So, in summary:

1. Prediction-identical preferences never change relative probability.
  2. The simplest prediction-identical preferences are known to be wrong for humans.
  3. It could be very important for the future to get the right preference for humans.
- 

1. GR would make up a larger portion of G, "geometric theories of space-time" than AGR makes up of A, and G would be more likely than A anyway, especially after updating on the non-observation of angels. [←](#)

# Are We Right about How Effective Mockery Is?



Crossposted from [Figuring Figuring](#).

*I did a second survey that fixed some of the flaws of the first survey. The results from the second survey significantly color the interpretation of the results from the first survey given in the first “Conclusion and Discussion” section. Please continue reading past the section titled “Second Survey” to get a full picture of the results from all surveys.*

## Intro

A couple days ago a friend of mine on facebook asked about arguments in favor of mockery. They pointed out that they had noticed a lot of facebook posts mocking people for not wearing masks in the covid-19 era, and wondered whether this was an effective way to change people's behaviors.

I said in the comment section of that post that I would make a survey that worked as follows. Roughly half of the survey takers would be randomly assigned to answer the following questions:

1. Do you think that mockery is an effective way to change people's minds?
2. Do you think that mockery is an effective way to change people's behaviors?

The other half would be randomly assigned to answer these questions:

1. Has being mocked ever caused you to change your mind about something?
2. Has being mocked ever caused you to change your habits or behaviors?

No survey respondent was permitted to see all four questions. The possible answers to each question were “Yes”, “No”, and “Not sure”.

I made this survey using [GuidedTrack](#). I posted it on my facebook wall, and also posted it to [Positly](#), and paid people to participate.

A total of 145 people responded to any of the questions on positly. 74 were asked the first set of questions, and 71 were asked the second set of questions. A total of 66 people responded to any of the questions on facebook. 31 were asked the first set of questions, and 35 were asked the second set of questions.

Before I go on to tell you the results of the survey and the predictions me and some of my friends made, you might want to make your own predictions. I suggest you quickly scribble them down. Some particular questions you might want to make predictions about:

- Did more people answer yes to the first set of questions than the second set of questions, or is the reverse true, or were they about the same?
  - Were facebook respondents (presumably people who are friends, or friends of friends of mind on facebook) more or less likely to say yes to the first set of questions?
  - Were facebook respondents more or less likely to say yes to the second set of questions?
  - What did I predict about the previous two questions?

There may be other fun questions to predict, and I'd be curious to hear how you did in the comments. Predictions from me and my friends coming up, so make sure you make your predictions beforehand. Again, I suggest that you write them down. You may also want to write down your reasoning beforehand.

- - - - -

## Predictions

Ok, last chance to make predictions before you hear some spoilers...

Alright.

I predicted that many more people would answer yes to the first set of questions (ie, the questions about whether mockery is effective) than to the second set of questions. I also predicted more people would say no to the second set of questions than to the first.

I'm not sure exactly what my theory was when I made that prediction—I made the prediction in the same comment that I suggested the survey, but I came up with two post hoc hypotheses that might explain the result I predicted. I do know that part of the reason I made that prediction is that mockery is fun, but admitting that fun is the main reason we do it rather than because of its positive effects on other people's behavior feels kind of icky. So we use its effectiveness as an excuse.

One hypothesis is that we overestimate the effectiveness of mockery. This would make sense of the predicted result because it would be evidence that we all think mockery works on others, but none of us thinks it works on us.

The second hypothesis I made up to explain this predicted result was that while we know that mockery works on other people, we are hesitant to admit that it works on us, because that is a bit embarrassing. Perhaps people are also not that great at telling what actually caused them to change their minds or behaviors.

These hypotheses are not mutually exclusive.

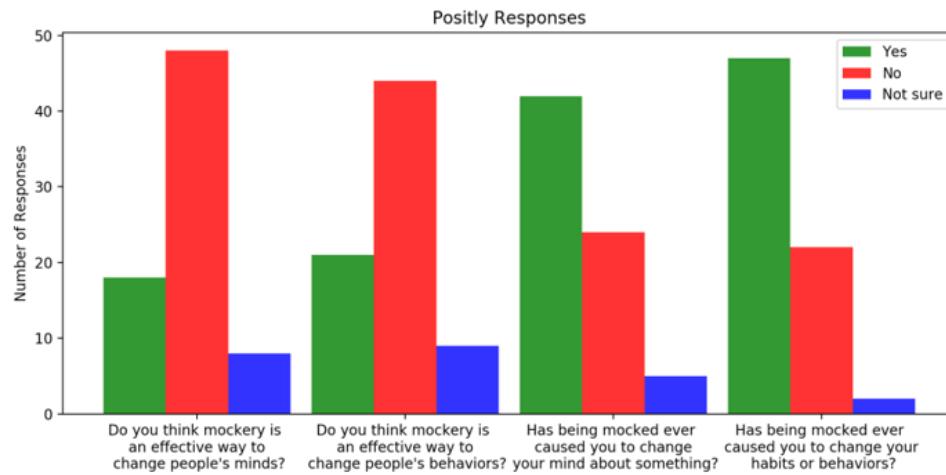
In a private conversation, my friend predicted using similar reasoning that actually people would tend to answer the second set of questions (ie, those about how often we change our own minds as a result of mockery) affirmatively. Saying that you think mockery is effective feels kind of icky, but saying that you think you have never had your mind or behavior changed because of mockery seems kind of arrogant.

Seeing how such similar reasoning could be used to predict a totally different result made me feel a bit nervous.

Another friend of mine predicted that my facebook friends would be less likely to change their minds because of mockery than randomly selected survey participants. Positly users aren't quite randomly selected, but they're closer to randomly selected than whatever people happened to come across my facebook post.

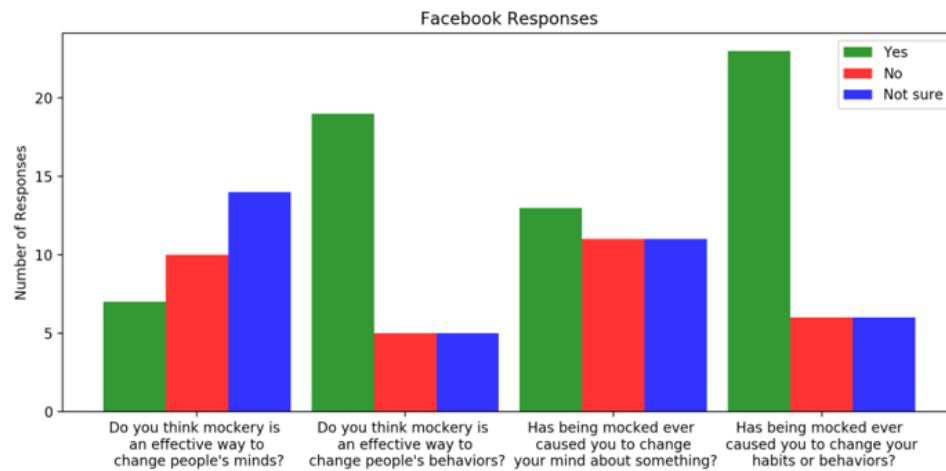
## Results

Sure enough, I was totally wrong.



Twice as many Positly respondents said that mockery has worked on them as said that mockery is effective. Positly respondents were slightly more likely to say that mockery is effective for changing behavior than for changing people's minds, both for themselves and for others.

I think this is strong evidence against the hypotheses I suggested, and some evidence in favor of the hypothesis my friend suggested in conversation.



My facebook acquaintances were slightly less disproportionate. Only 1.2 times as many respondents said that mockery is effective on themselves as said that mockery is effective on others when it comes to behavior. However, when it comes to changing minds, still about twice as many said that mockery has worked on them as said that mockery is effective on others.

I was surprised by this, as I tend to think of myself as preferring people who do not use mockery, and not using mockery while also thinking it is effective is a hard pair of things for humans to do simultaneously.

Of the 35 facebook respondents that were asked the second set of questions, 37% said they had changed their mind because of mockery, 59% of Positly respondents said the same. This seems like decent evidence to me that my friend was right about my facebook acquaintances being less likely to change their minds because of mockery.

No respondent said that mockery was effective for changing their own, or other people's minds and also answered that it was not effective for changing their own, or other people's behaviors.

## Other Responses

Here is a list of some of the things that people said they changed because of mockery. This was an optional part of the survey. Slightly edited for brevity, to protect anonymity, and avoid repeats.

- Economic beliefs
- Working out habits
- Avoiding people who mock them
- Writing about topics on fb
- My own appearance
- How good other people are
- Picking my nose in public
- Crying in public
- Basic cultural rules, like where to sit, how to join a conversation, etc.
- Philosophical or ethical beliefs
- Individualism as an ethical stance
- Fashion
- Music
- Using an old fashioned word
- Beliefs about what is socially acceptable
- Conversational habits
- Wearing briefs instead of boxers
- Stopped whistling
- Stopped/started wearing shorts
- Eating habits
- Lost weight
- Stopped playing sports
- Being late
- Hairstyle
- Stopped watching anime
- Mocked for being autistic, so changed the way I interact with people.
- Started wearing make up.

- Mocked for being outgoing, became less outgoing and confident.
- Started liking Trump
- Left Mormon religion
- Started thinking more before speaking
- Started brushing teeth more
- Stopped being conservative

Here is a list of some of the things that people said mockery was effective for changing in other people. This was also an optional part of the survey. The entries in this list have been edited as in the previous list.

- Weird opinions
- The way people think
- The way others dress
- Haircuts
- Mask wearing
- How often someone complains
- Making someone hide their opinions
- Arrogance
- Getting people to stop doing things around you
- Getting someone to stop writing things in public

I think these lists are similar enough in content to rule out another explanation of this data. You might have thought that people think mocking people is an effective way to get other people to change certain kinds of things, but when they think about what sorts of things they have changed themselves because of mockery, the two categories do not have much of an intersection. These lists make that seem unlikely to me.

## **Discussion and Conclusion (1)**

These results seem like some evidence to me that people in general underestimate the effectiveness of mockery for getting others to change their minds. This is of course not necessarily an argument for using more mockery. I, for one, take the results of this survey to be a further reason that we should not mock people.

If you thought mockery was just some harmless fun you can have with your in group, as I sort of did, you might have thought that the costs to those being mocked are actually not that great. But it seems like mockery can make someone leave their religion, stop writing in public, change their political preferences, etc. I would strongly prefer for people to make decisions about those sorts of things using object level reasoning rather than reasoning about what will cause them to be mocked less. I will now much more than before see mockery as deliberate enemy action designed to interrupt other people's cognition—not something to be taken as a joke, especially not in the context of conversations about important topics.

## **Second Survey**

*This section was written after getting the data for my second survey which was inspired by some criticisms of the questions in the first. Everything above was written before getting that data.*

On the other hand, the questions I asked people in the original survey were not exactly analogous to each other. Firstly, people might have answered the first set of questions considering that although mockery is rarely effective for changing the behaviors or beliefs of those being mocked, it might work on bystanders who watch the mocking happen. Secondly, people answering the second set of questions with a “yes” might be thinking that “yes, mockery has ever caused me to change my mind” but that does not mean it is very effective.

To correct for this, I made a second survey. Half of respondents were asked the following two questions:

1. How often has mockery worked as a means of getting you to change your mind about something?
  2. How often has mockery worked as a means of getting you to change your habits or behavior?

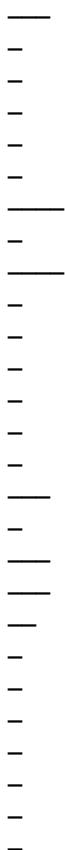
The other half were asked the following two questions:

1. How often does mockery work as a means of getting someone to change their mind about something?
  2. How often does mockery work as a means of getting someone to change their habits or behavior?

The possible responses were: "very often", "often", "sometimes", "rarely", and "almost never".

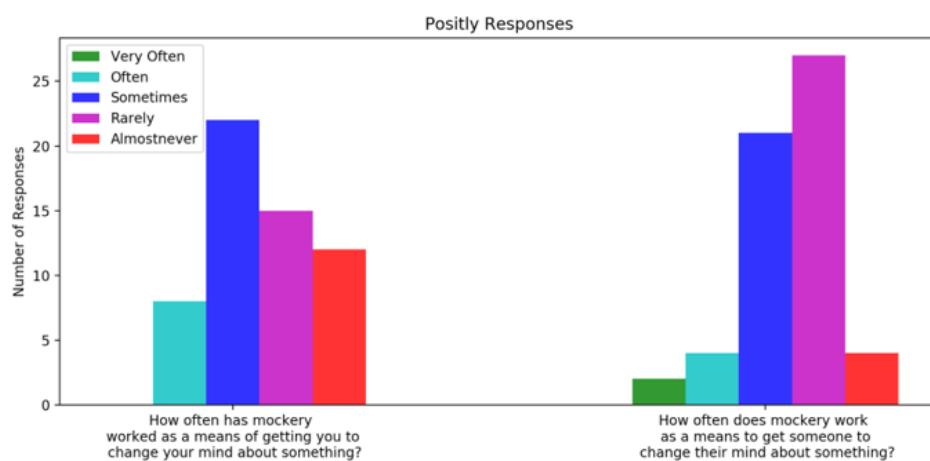
The survey was published on positly.

I will give you some room to make predictions before showing the results.

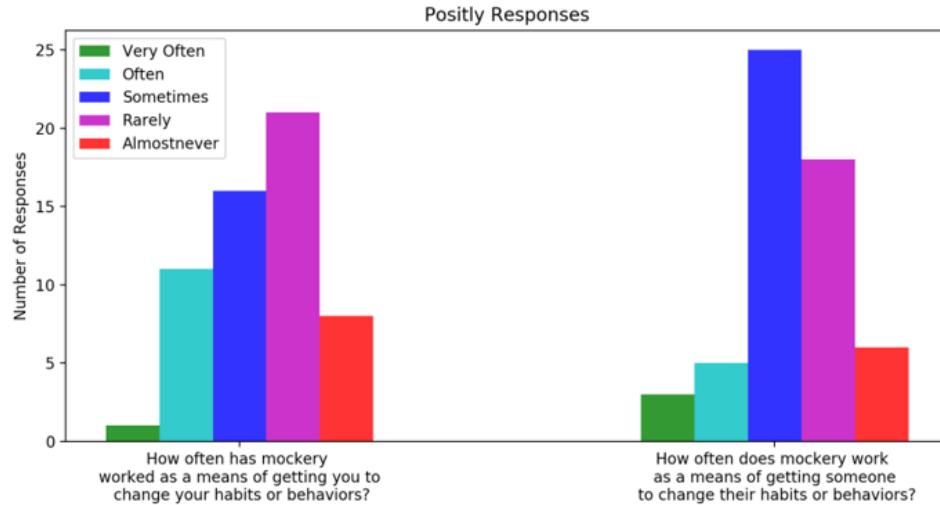


## Second Survey Results

There were a total of 115 respondents. 57 were asked the first set of questions, 58 were asked the second set of questions. Here are the results compared across groups.



Mapping “very often” to 4, “often” to 3, “sometimes” to 2, “rarely” to 1, and “almost never” to 0, this gives a mean response for group 1 question 1 of 1.4562, and a mean response for group 2 question 1 of 1.5345, meaning that respondents overall thought that mockery was slightly more effective on others than on themselves.



Using the same mapping, the average response for group 1 question 2 was 1.5789, and the average response for group 2 was 1.6667. Again, respondents overall thought that mockery was slightly more effective on others than on themselves.

## Discussion and Conclusion(2)

These results contradict my original interpretation of the first survey's data. The second survey suggests that people are in general pretty well calibrated about the effectiveness of mockery, or perhaps slightly underestimate it. I conclude that much of the effect observed in the results of survey one's data was caused by the two effects discussed at the beginning of the "Second Survey" section and not the result of people genuinely underestimating the effectiveness of mockery.

However, I think I am still going to take mockery more seriously than I did before, mostly because I still think this survey showed me that mockery is more effective than I thought it was. The list of personal examples people gave were fairly chilling. I also imagine people cave to mockery a lot more than they are able to notice or willing to admit on a survey. Furthermore, I don't think it was a coincidence that it was mostly me and my weirdest friends who (incorrectly) predicted that people would say that mockery is much more effective on others than it is on themselves. Probably, we weirdos have grown numb to mockery's sting, and fallen out of touch with what it feels like to be mocked for most people.

## I Would like to Thank

Ozzie Gooen for inspiring me to make these surveys with his facebook post.

Frank Bellamy and Julia Kris Dzweiria for pointing out the assymetry of the questions in the original survey.

Beth Kessler and Aryeh Englander for useful discussion.

And Spencer Greenberg as well as the whole of the [Positly](#) and [Guidedtrack](#) teams for making it much easier to run surveys like these.

# Agentic Language Model Memes

Related: [What specific dangers arise when asking GPT-N to write an Alignment Forum post?](#)

I've been thinking about the AI safety implications of ultra powerful language models. I think it makes more sense to think of language models trained with internet text as containing a collective of agents that can be queried with appropriate prompts. This is because language models are trained to mimic their training data. The best way to do this is for them to approximate the process generating the data. The data used to train modern language models is produced by millions of humans, each with their own goals and desires. We can expect the strongest of language models to emulate this, at least superficially, and the kind of agent that the language model emulates at any given point will be a function of the upstream prompt.

With a strong enough language model, it's not unthinkable that some of these agents will be smart enough and sufficiently unaligned to manipulate the humans that view the text they generate. Especially since this already happens between humans on the internet to some degree. This is especially true in environments like AI dungeon, where the agents can get feedback from interacting with the human.

More generally, from an AI safety perspective, I wouldn't be worried about language models per say, as much as I would be worried about what I'm going to call agentic language model memes. Prompts that get language models to emulate a specific, unaligned agent. The unaligned agent can then convince a human to spread the prompt that instanced it through social media or other means. Then, when a new language model arrives, the unaligned agent takes up a larger fraction of the networks probability density. Furthermore, prompts will experience variation and face evolutionary pressure, and more viral prompts will get more frequent. I hypothesize that this will create an environment that will lead to the emergence of prompts that are both agentic and effective at spreading themselves. With enough iterations, we could end up with a powerful self replicating memetic agent with arbitrary goals and desires coordinating with copies and variations of itself to manipulate humans and gain influence in the real world.

At the moment, I suspect GPT-3 isn't good enough to support powerful memetic agents, primarily because the 2048 BPE prompt length and architectural constraints severely limit the complexity of agents that can be specified. However, I see no reason why those limitations couldn't be overcome in future language models. [We've already seen evidence that we can talk to specific agents with capabilities that GPT-3 wouldn't otherwise have by default.](#)

The good news is that at least initially, in the worse case these agents will only have slightly superhuman intelligence. This is because large LMs will need to be trained from data generated by regular humans, although selection pressures on what gets into future language model datasets could change this. Furthermore, at least initially, language models strong enough to effectively emulate an agent will probably be expensive, this will limit the ability of these memes to spread. The dragon model for AI dungeon is cheap, but paying for it is still a barrier to entry that will only get higher as the models get bigger.

The bad news is that hardware and ML capabilities will continue to increase, and I speculate that powerful language models will become ubiquitous. I wouldn't be surprised if something comparable to GPT-3 was made to fit in a cell phone within 5 years, at which point the necessary conditions for the creation of these agents will be met.

Speculating further, if we do end up seeing agentic memes, it's not entirely obvious what they'll look like. The idea came to me after thinking about what would happen if I tried to talk to HPMOR!Quirrel in AI dungeon, but that might not be the most effective meme. The closest analogues are probably modern social media memes. If they're anything to go by, we can expect a whole plethora of agents ranging from those that spread by providing value to users, to concentrated rage bait.

# "On Bullshit" and "On Truth," by Harry Frankfurt

**Salticidae Philosophiae** is a series of abstracts, commentaries, and reviews on philosophical articles and books.

Harry Frankfurt asks, "What is bullshit, anyway?" Also, "What is truth?" but we all know that book proposal wouldn't have flown except as a companion to the first one.

## Highlights

- Something can be true, and still be bullshit.
- Something can be a lie, and yet not be bullshit.
- Bullshit is that which is (1) unconcerned with truth and (2) intended to change attitudes rather than beliefs.
- Truth is useful to us as individuals and as societies
- Truth-seeking and truth-telling must be rewarded and their inverse must be punished.
- Truth is truth whether or not anyone believes it or even knows it.

## New or uncommon terminology

- *On Bullshit* is described as a **prolegomenon** to *On Truth*, or an extended introduction that serves to discuss and interpret the work in a manner that is more exhaustive than the typical introduction.

## Book-by-book

### On Bullshit

There is not much literature on bullshit, and no "theory of bullshit" or rigorous analysis thereof. This is in large part because we all assume that we recognize and evade bullshit pretty well.

According to Max Black, humbug is essentially a (false) statement made, not to convince you about that thing, but to convince you of something else. For example, one might make blatantly and obviously exaggerated or otherwise false statements about U.S. history not to convince another of these things, but to convince another of one's patriotic fervor.

Starting from this definition of humbug, Frankfurt makes a number of comparisons and caveats that might be useful:

- Bullshit may be made carelessly, and we could easily compare bullshit to shoddy goods.

- Shit is excreted, not crafted. However, advertising can be carefully-crafted bullshit.
- Similes are not lies, but they can be made too thoughtlessly. In their own way, they can be bullshit.

Frankfurt argues that bullshit is, to start with, deliberate misrepresentation. Some say that lying requires intent; others, that any false statement is a lie. Bullshitting, however, is not exactly the same as lying. Indeed, bullshit can be true. Frankfurt's position is that bullshit is distinguished not by its truth or falsity, but by a disregard for the truth; as he puts it, honest folk and liars are playing the same game, to convey the facts or to obscure them, but the bullshitter is playing another game entirely.

Truth-tellers and liars are both concerned with changing your beliefs; a bullshitter is concerned with changing your attitude.

- "Someone who ceases to believe in the possibility of identifying certain statements as true and others as false can have only two alternatives. The first is to desist both from efforts to tell the truth and from efforts to deceive. [...] The second alternative is to continue making assertions that purport to describe the way things are, but that cannot be anything but bullshit." [pg 61 para 2]
- "Just as hot air is speech that has been emptied of all informative content, so excrement is matter from which everything nutritive has been removed. Excrement may be regarded as the corpse of nourishment, what remains when the vital elements in food have been exhausted. In this respect, excrement is a representation of death that we ourselves produce and that, indeed, we cannot help producing in the very process of maintaining our lives." [pg 43 para 1]

## On Truth

### Introduction

This is a sequel to *On Bullshit*, which addresses an oversight of his: the author failed to make any argument as to why the truth is important, and bullshit is therefore reprehensible. This book is about why truth is important.

There is lots of bullshit but it hasn't destroyed civilization, so some people think that truth isn't important. Some people even refuse to admit that there is such a thing as truth, though these people are very silly (not least because they tend to represent themselves as *truly* holding this belief). The book therefore assumes that there is an "objectively meaningful or worthwhile distinction to be made between what is true and what is false," and concerns itself solely with addressing whether this distinction matters outside of academia.

He spends more than a tenth of the book explaining what he's doing and why he's doing it.

### Chapter I

Truth is useful to us. Societies cannot function without fostering truth. Both individuals and groups must know facts and as societies become more complex they must know more facts, and more accurately (while many individuals, it must be said, can remain free riders).

Postmodernists reject the idea that truth has objective reality or value, at least as perceived by us; our view of the truth is determined by constraints that have been imposed upon us by personal and social environments and histories. It is interesting that postmodernism does not exist (in this form) in medicine, physics, and other fields whose assertions are easily testable. Even in history, there must be objective facts: "They will not say that Belgium invaded Germany," the author reports Georges Clemenceau as saying.

## **Chapter II**

Even if some value statements are not verifiable, they can generally be connected back to facts that can be discussed. Knowing the facts of the matter lets us determine whether we ought to value the things that we do, or whether other goals and activities might better accomplish our terminal values.

- Healthy societies must reward truth-finders and punish truth-obscurers.
- Having facts is not enough to succeed (you must use them properly), but not having facts prevents you from taking any action at all.

## **Chapter III**

One might argue that we could just not care about this need for truth. Spinoza argued that we cannot help but care, because of love, which is "nothing but Joy with the accompanying idea of an external cause." essentially an experience that broadens one's understanding of oneself and improves one's capacity for perfection. or (in the author's words) "the way that we respond to something that we recognize as giving us joy." Additionally, joy is the experience of being ennobled or otherwise improved (and, preferably, knowing this). Therefore, truth gives us joy, because it improves us, and because we wish to preserve and keep nearby that which we love, we will seek to preserve truth.

## **Chapter IV**

When we act, we interact with reality, and we have a desire or at least an expectation regarding the outcome of our action. To the degree that we lack truth, we are disconnected from reality and that desire or expectation may be thwarted.

- It is always better to face uncomfortable truths than to hide away from them, because if we do not confront them then, one day, we will be confronted by them.
- Without truth, we are blind. We might not run into trouble immediately, but we will do so inevitably.
- "The relevant facts are what they are regardless of what we may happen to believe about them, and regardless of what we may wish them to be. This is, indeed, the essence and the defining character of factuality, of being real: the properties of reality, and accordingly the truths about its properties, are what they are, independent of any direct or immediate control by our will." [pg 55 para 2]

## **Chapter V**

Truth fosters trust. Honesty is the foundation of society, while dishonesty undermines social fabric. Even the capacity for self-recognition (or self-awareness, we might say) ultimately depends on our relationship with the truth. If we do not know the world, then we cannot know ourselves.

- If someone starts getting into etymology as part of some Deep Explanation, then prepare for a torrent of bullshit.
- Immanuel Kant, "On a Supposed Right to Lie from Altruistic Motives": "A lie always harms another; if not some particular man, still it harms mankind generally.
- Michel Montaigne, "Of Liars": "If we did but recognize the horror and gravity of [...lying], we would punish it with flames more justly than other crimes."

Even the capacity for self-recognition (or self-awareness, we might say) ultimately depends on our relationship with truth. If we do not know the world, then we cannot know ourselves.

## Comments

*On Bullshit* argues that bullshitting doesn't necessarily undermine society, at least not up to a point, but I would expect a hypothetical society with even slightly less bullshit than ours to function more smoothly. I also disagree with the position that truth intrinsically gives us joy. Many of us love bullshit more than truth.

Frankfurt says that lying is bad at its core because the liar "tries to impose his will on us," even if it is for our own good, but he fails to argue that this in itself is bad. More convincing is Frankfurt's argument that we are being pushed into another world insofar as our beliefs are false, but what if the lie is believed on a large scale? Then we would be isolated by believing the truth. He also argues that the liar is personally isolated, and cannot even speak of that isolation, but this is untrue if the liar has partners.

## Favorite passage

As conscious beings, we exist only in response to other things, and we cannot know ourselves at all without knowing them. Moreover, there is nothing in theory, and certainly nothing in experience, to support the extraordinary judgment that it is the truth about himself that is the easiest for a person to know. Facts about ourselves are not peculiarly solid and resistant to skeptical dissolution. Our natures are, indeed, elusively insubstantial--notoriously less stable and less inherent than the natures of other things. And insofar as this is the case, sincerity itself is bullshit. [*On Bullshit*, pg 66 para 2]

## Author biography

Harry G. Frankfurt is Professor of Philosophy Emeritus at Princeton University. His books include *The Reasons of Love* (Princeton), *Necessity, Volition, and Love*, and *The Importance of What We Care About*.

# Philosophers & works mentioned

Philosophers given significant attention include:

- Max Black, a British-American philosopher who, for some reason, has a longer article on the Unitarian-run [New World Encyclopedia](#) than he does on [Wikipedia](#), even though the former takes its articles from the latter before it edits and builds upon them.
- Immanuel Kant, a German philosopher who argued that reason is the basis of morality, and drew attention to the difference between the world-as-it-is and the world-as-it-appears-to-us.
- Michel Montaigne, a French philosopher of the Renaissance period who wrote on child education, psychology, and other topics, and popularized (but did not invent) the essay format.
- Baruch Spinoza, a Jewish Portuguese philosopher who is best known for his writings on God, which have gotten him labeled as everything from a pantheist to an atheist.

## Other articles & books on this subject

- [\*The Prevalence of Humbug\*](#), by Max Black

# You Need More Money

Part 1 of the *Inefficient Markets Sequence*.

[Epistemic Disclaimer: I am not rich yet. I feel like I'm just barely starting to understand this stuff and perhaps that is the best time to teach it: while I still remember what wasn't obvious, both to cultivate my comprehension, and to enrich the rationalists.]

For educational purposes only! Double-check anything I say. Read the comments. *You have been warned!*

You are responsible for your own money. Do your due diligence and don't rely too much on random internet bloggers :) I do not know your financial situation. I am not your financial advisor.]

## Convergent Instrumental Rationality

Sufficiently advanced intelligent agents, almost no matter their ultimate ends, will tend to pursue [instrumental means](#)(69) such as self-preservation and resource acquisition.

What of sufficiently advanced humans?

## The Outside Perspective

Can you imagine a world you'd rather live in? Perhaps your life is comfortable, for the moment. But the world is truly awful right now. Children are dying as we speak, of hunger, or war, or disease. A cheap shot, but I'm sure you could come up with many other ills. The world was even worse, in many ways, in the recent past. It could become worse again, or better. [Perhaps a lot better.](#)

If you truly have [something to protect](#), what are your means?

If, when pondering a question, you discover what something smarter than yourself would answer, perhaps you have found your answer also.

What would a superior intelligence do in your shoes? What would a more advanced culture think of ours?

You can't know all the answers to these questions, but we do know this much: acquire resources.

## "Need" Is a Relative Term.

Distinguishing a "need" from a "want" is one of those elementary-school tasks that we all think is easy, at the time. It's a basic budgeting skill: buy what you *need* first, then buy what you *want* with what's left over.

But upon deeper examination, it's not so simple.

Do you *need* to see that doctor? Your ancestors only a few generations back lacked access to competent medicine. The physicians of the past often did more harm than good. Were their basic *needs* being met? And in the more recent past, many diseases which now have effective cures were fatal. Were their basic *needs* being met? Not by today's standards.

From the perspective of a more advanced culture (if we [knew more](#), thought faster, were more the people we wished we were, had grown up farther together), are your basic needs being met? Is effective [cryonics](#) a *need* or a *want*? (And are the current options effective enough to bother with? That's a point of some contention. Wikipedia [still calls it pseudoscience and quackery](#).)

If, when pondering a question, you discover what something smarter than yourself would answer, perhaps you have found your answer also.

I posit that *if you are not clinically immortal*, then your **basic** health needs are not being met!

This is not a *want* from the outside perspective of a competent culture. Just like how we know antibiotics are a basic medical need for someone with a life-threatening infection, clinical immortality is a basic need for a culture that has a cure for aging. And a humane technological culture will eventually achieve that. There's nothing in the laws of physics that *requires* you to age. [The Dragon Is Bad](#). So we know what the smarter culture thinks. Shouldn't we think so too? That it even occurs to you that this might be a *want* is an anomaly caused by the peculiar time and place you find yourself in. You're weird.

So you see, there are degrees of need. There are the immediate needs that you already think of as needs, and less immediate, yet *vital* needs that are far more than you can afford.

Self-preservation is an instrumental goal of sufficiently advanced humans.

Are you clinically immortal? That was rhetorical.

How do we get there? Resource acquisition is an instrumental goal of sufficiently advanced humans. You need more money! A lot more money! We all need more money!

## Effective Altruism

If you're rich and have a pressing medical need, perhaps there's an option open to you that others in your situation couldn't afford: fund medical research for your condition.

Money can buy you a lot of safety. (Self-preservation through resource acquisition!)

But besides donating to research that may save your life one day, there are a lot of other problems you could help with if you had more money.

Friendly AI is the most important. If we get that one right, it will solve all of our solvable problems. If we get it wrong, nothing else matters, because we will no longer be in control.

Much ink has been spilled on the topic of EA and FAI. I'll not say more here.

## They Say Money Can't Buy Happiness. Jealous much?

Poverty is misery. Money gets you at least halfway to happiness.

But are you even that far yet? [This Princeton study](#) suggests that you need an annual income of at least \$75,000 for peace of mind. But that was in 2010. Based on the consumer price index, that would be about \$90,000 in 2020 dollars. Does it need to be more if you live in an expensive area (like the Bay Area)?

Is your income at least that high? No? Then you need more money!

Maybe you can do better than halfway if you spend it wisely.

## Money Is Time

Perhaps the most valuable thing you can buy is time. It's a shame that we still have to work 40 hour weeks. It used to be worse. Civilization should be past that by now. Free time gives you the opportunity to work on everything else. You can become that [thousand-year old vampire](#), or at least eight hours closer to it per work day than your peers. The right skills probably have compounding returns.

More time for family. More time for saving the world.

To get \$90,000 per year without working, on a fixed income of 4%, you need \$2.25 million.

Do you have at least that much? No? You are not rich yet! You need more money! Yes? Maybe you are (barely) rich, but you are still not immortal yet! You need more money!

## Time Is Not Money

How long will it take you to earn that \$2.25 million? 40 years? Too long! We're trying to buy free time here, not waste it on a 40 hour work week for 40 years.

You need more money. [Rationalists Should Win](#). So why aintcha rich? Because we are not winning hard enough yet. We are *doing it wrong*. You can't wait 40 years. You're supposed to be a rationalist. Get creative. Solve this problem.

Move to Costa Rica where the costs of living are lower? Maybe that's a backup plan. But you're still not immortal!

Live in a trailer while doing remote work for a software company? Remote work seems like more of an option now than ever before. The pandemic has changed things. How about your parents' basement?

Getting better, but no, the answer is clear: **You have to decouple time from income.** As much as you can.

Having the \$2.25 million will do it, but it takes money to make money, doesn't it?

## Rationalists Are Already Halfway to Quant

We've got a chicken-egg problem. The solution to chicken-egg problems is bootstrapping. You need to develop sources of income that don't take too much time.

Sell cheap Alibaba junk on Amazon for five times the price? Sell easy Facebook ad development services to local small business that barely know how to use a computer? Flip houses? Do the Bay Area startup thing until you hit a home run?

I don't particularly care how you do it (I mean, don't be a criminal), as long as you're decoupling. Whatever business you think you can handle, more power to you!

What talents do we have, as rationalists? Reasoning and acting under uncertainty. Computer literacy: web searches, spreadsheets. Data science: code, statistics. The discipline to act on the numbers. What else?

It sounds to me that the typical rationalist is already halfway to quant.

## Early Retirement

Retirement should not take you 40 years to earn. If you can save money, and I mean a significant fraction of your income -- preferably half or more, and manage your investments well, you can retire a lot earlier than your 60s, or at least have a shorter workweek doing trading instead of whatever it is you're doing now. (Assuming, of course, that you aren't already too close to that old.)

If you earn 4% per year, then you need the aforementioned \$2.25 million for the \$90,000 half-happiness income. If you earn 10% per year, you only need \$900,000. If you earn 15% per year, you only need \$600,000. At 18% you need \$500,000; at 24% you need \$375,000. And of course, you can acquire that nest egg a lot faster if you're earning a good return on your smaller investments.

And if you really want your free time sooner (and I do), perhaps you can settle for a quarter-happiness income of \$45,000 plus lots more free time to start? You could keep saving money and grow your income over time. At 18% you only need \$250,000.

Now maybe that still seems like a lot to you, but it's much less than \$2.25 million!

I'm oversimplifying a bit here. While I do think 24% returns (or more!) are achievable, they would be volatile. You still have to pay taxes and health insurance. You'd need a cushion for drawdowns, which could last for years if you are unlucky. Maybe it's just an extra \$90,000. Maybe you could arrange to move back in with your family for a short period. Maybe your previous company would take you on for a while. Perhaps you have a spouse who could work part time (or full time, that helps with insurance too).

Do those returns seem unrealistic? I hope to convince you otherwise.

## Wealth Is a State of Mind

Not *literally* true, but hear me out.

Your talents are not just your skills. They're also who you know and what you own. Develop them.

If you found yourself broke in a poor neighborhood, how long would it take you to get out?

Borrow a phone and contact a wealthier friend or family member, and they'll probably have you out the same day.

Suppose that way was closed to you, but you had resources comparable to the people living there. Poor neighborhoods are a "[poverty trap](#)". How long? Really think about it. For at least [five minutes by the clock](#). How would you do it?

Maybe your education is an advantage. Suppose you were unable to prove your credentials. Your university lost your records in a fire. You don't have your diploma. But you'd still have your skills, what you can remember of them.

How long?

Now suppose an upper-class person found themselves in *your* neighborhood with *your* resources? I don't mean a lazy trust-fund kid with no skills. Someone who culturally understands money, who earned it themselves, or had parents who taught them well. How long would it take them to get out and back up to their standard of living? How would they do it? Really think about it, for at least five minutes.

Maybe their credit with financial institutions is an advantage. In his period of financial ruin, Donald Trump once famously remarked, "See that bum? He has a billion dollars more than me." Say what you will about Trump, but recently, he had the finances to fund a presidential campaign, and win. The bum is probably still homeless. What's the difference? What advantage was enough to overcome that?

If you only had the requisite talents, a mere lack of savings would be no object. Saving for retirement is middle-class thinking.

Suppose that way was also closed, due to whatever reason they're stuck in your neighborhood. The banks won't go for it. No loans bigger than you could get yourself. How long? How would they do it?

You have the same resources!

If, when pondering a question, you discover what something smarter than yourself would answer, perhaps you have found your answer also.

How long would it take you?

I first heard about this thought experiment from [this 2015 talk by Douglas Kruger](#). He has more to say about the topic. I think it's worth a listen.

## Systematized Resource Acquisition

I hope you now have [a sense that more is possible](#). Why aren't the "rationalists" surrounded by a visible aura of formidability? Because we have not developed the Art

sufficiently. We haven't really gotten together and systematized our skills.

**Resource acquisition is a core rationality skill.**

This sequence is the beginning of developing that skill into a system.

Perhaps you don't have the talents to woo investors and run a business well, or to even recognize and negotiate with people competent in these areas.

But perhaps if you made the right kind of friends, or at least knew the right kind of people, you could learn some of these things and share them with us. I'm certainly not there yet. This sequence is only a start.

So let's work on saving for early retirement for now. If only to buy us the free time needed to develop the skills to become truly wealthy.

---

Up next, the foundational skill of trading: how to not lose money.

- Part 2: [Repeat Until Broke](#)
- Part 3: [How to Lose a Fair Game](#)
- Part 4: [The Wrong Side of Risk](#)

Followed by an introduction to alpha.

- Part 5: [Market Misconceptions](#)
- Part 6: [Charting Is Mostly Superstition](#)

# Forecasting Newsletter: July 2020.

## Highlights

- Social Science Prediction Platform [launches](#).
- Ioannidis and Taleb [discuss](#) optimal response to COVID-19.
- Report tries to [foresee](#) the (potentially quite high) dividends of conflict prevention from 2020 to 2030.

## Index

- Highlights.
- Prediction Markets & Forecasting Platforms.
- New undertakings.
- Negative Examples.
- News & Hard to Categorize Content.
- Long Content.

Sign up [here](#), browse past newsletters [here](#), or view it on the EA forum [here](#).

## Prediction Markets & Forecasting Platforms.

Ordered in subjective order of importance:

- Metaculus continues hosting great discussion.
- In particular, it has recently hosted some high-quality [AI questions](#).
- User @alexrlj, a moderator on the platform, [offers on the EA forum](#) to operationalize questions and post them on Metaculus, for free. This hasn't been picked up by the EA Forum algorithms, but the offer seems to me to be quite valuable. Some examples of things you might want to see operationalized and forecasted: the funding your organization will receive in 2020, whether any particularly key bills will become law, whether GiveWell will change their top charities, etc.
- [Foretell](#) is a prediction market by the University of Georgetown's Center for Security and Emerging Technology, focused on questions relevant to technology-security policy, and on bringing those forecasts to policy-makers.
- Some EAs, such as myself or a mysterious user named *foretold*, feature on the top spots of their (admittedly quite young) leaderboard.
- I also have the opportunity to create a team on the site: if you have a proven track record and would be interested in joining such a team, get in touch before the 10th

of August.

- [Replication Markets](#)
- published their [first paper](#)
- had some difficulties with cheaters:

"The Team at Replication Markets is delaying announcing the Round 8 Survey winners because of an investigation into coordinated forecasting among a group of participants. As a result, eleven accounts have been suspended and their data has been excluded from the study. Scores are being recalculated and prize announcements will go out soon."

- Because of how Replication Markets are structured, I'm betting the cheating was by manipulating the Keynesian beauty contest in a [Predict-O-Matic](#) fashion. That is, cheaters could have coordinated to output something surprising during the Keynesian Beauty Contest round, and then make that surprising thing come to happen during the market trading round. Charles Twardy, principal investigator at Replication Markets, gives a more positive take on the Keynesian beauty contest aspects of Replication Markets [here](#).
- still have Round 10 open until the 3rd of August.
- At the Good Judgement family, Good Judgement Analytics continues to provide its [COVID-19 dashboard](#).

Modeling is a very good way to explain how a virus will move through an unconstrained herd. But when you begin to put in constraints" — mask mandates, stay-at-home orders, social distancing — "and then the herd has agency whether they're going to comply, at that point, human forecasters who are very smart and have read through the models, that's where they really begin to add value. – Marc Koehler, Vice President of Good Judgement, Inc., in a [recent interview](#)

- [Highly Speculative Estimates](#), an interface, library and syntax to produce distributional probabilistic estimates led by Ozzie Gooen, now accepts functions as part of its input, such that more complicated inputs like the following are now possible:

```
# Variable: Number of ice creams an unsupervised child has consumed,  
# when left alone in an ice cream shop.  
  
# Current time (hours passed)  
t=10  
  
# Scenario with lots of uncertainty  
w_1 = 0.75 ## Weight for this scenario.  
min_uncertain(t) = t*2  
max_uncertain(t) = t*20
```

```

# Optimistic scenario
w_2 = 0.25 ## Weight for the optimistic scenario
min_optimistic(t) = 1*t
max_optimistic(t) = 3*t
mean(t) = (min_optimistic(t) + max_optimistic(t)/2)
stdev(t) = t*(2)^{1/2}

# Overall guess
## A long-tailed lognormal for the uncertain scenario
## and a tight normal for the optimistic scenario

mm(min_uncertain(t) to max_uncertain(t), normal(mean(t), stdev(t)), [w_1, w_2])

## Compare with: mm(2 to 20, normal(2, 1.4142), [0.75, 0.25])

```

- [PredictIt](#) & [Election Betting Odds](#) each give a 60%-ish to Biden.
- See [Limits of Current US Prediction Markets \(PredictIt Case Study\)](#), on how spread, transaction fees, withdrawal fees, interest rate which one could otherwise be earning, taxes, and betting limits make it so that:

"Current prediction markets are so bad in so many different ways that it simply is not surprising for people to know better than them, and it often is not possible for people to make money from knowing better."

- [Augur](#), a betting platform built on top of Ethereum, launches v2. Here are [two overviews](#) of the platform and of v2 modifications

## New undertakings

- [Announcing the Launch](#) of the [Social Science Prediction Platform](#), a platform aimed at collecting and popularizing predictions of research results, in order to improve social science; see [this Science article](#) for the background motivation:

A new result builds on the consensus, or lack thereof, in an area and is often evaluated for how surprising, or not, it is. In turn, the novel result will lead to an updating of views. Yet we do not have a systematic procedure to capture the scientific views prior to a study, nor the updating that takes place afterward. What did people predict the study would find? How would knowing this result affect the prediction of findings of future, related studies?

A second benefit of collecting predictions is that they [...] can also potentially help to mitigate publication bias. However, if priors are collected before carrying out a study, the results can be compared to the average expert prediction, rather than to the null hypothesis of no effect. This would allow researchers to confirm that some results were unexpected, potentially making them more interesting and informative, because they indicate rejection of a prior held by the research community; this could contribute to alleviating publication bias against null results.

A third benefit of collecting predictions systematically is that it makes it possible to improve the accuracy of predictions. In turn, this may help with experimental design.

- On the one hand, I could imagine this having an impact, and the enthusiasm of the founders is contagious. On the other hand, as a forecaster I don't feel enticed by the platform: they offer a \$25 reward to grad students (which I am not), and don't spell it out for me why I would want to forecast on their platform as opposed to on [all the other alternatives available to me](#), even accounting for altruistic impact.
- [Ought](#) is a research lab building tools to delegate open-ended reasoning to AI & ML systems.
- Since concluding their initial factored cognition experiments in 2019, they've been building tools to capture and automate the reasoning process in forecasting: [Ergo](#), a library for integrating model-based and judgmental forecasting, and [Elicit](#), a tool built on top of Ergo to help forecasters express and share distributions.
- They've recently run small-scale tests exploring amplification and delegation of forecasting, such as: [Amplify Rohin's Prediction on AGI researchers & Safety Concerns](#), [Amplified forecasting: What will Buck's informed prediction of compute used in the largest ML training run before 2030 be?](#), and [Delegate a Forecast](#).
  - See also [Amplifying generalist research via forecasting](#), previous work in a similar direction which was also inspired by Paul Christiano's Iterated Distillation and Amplification agenda.
- In addition to studying factored cognition in the forecasting context, they are broadly interested in whether the EA community could benefit from better forecasting tools: they can be reached out to [team@ought.org](mailto:team@ought.org) if you want to give them feedback or discuss their work.
- [The Pipeline Project](#) is a project similar to Replication Markets, by some of the same authors, to find out whether people can predict whether a given study will replicate. They offer authorship in an appendix, as well as a chance to get a token monetary compensation.
- [USAID's Intelligent Forecasting: A Competition to Model Future Contraceptive Use](#). "First, we will award up to 25,000 USD in prizes to innovators who develop an intelligent forecasting model—using the data we provide and methods such as artificial intelligence (AI)—to predict the consumption of contraceptives over three months. If implemented, the model should improve the availability of contraceptives and family planning supplies at health service delivery sites throughout a nationwide healthcare system. Second, we will award a Field Implementation Grant of approximately 100,000 to 200,000 USD to customize and test a high-performing intelligent forecasting model in Côte d'Ivoire."
- [Omen](#) is another cryptocurrency-based prediction market, which seems to use the same front-end (and probably back-end) as [Corona Information Markets](#). It's unclear what their advantages with respect to Augur are.

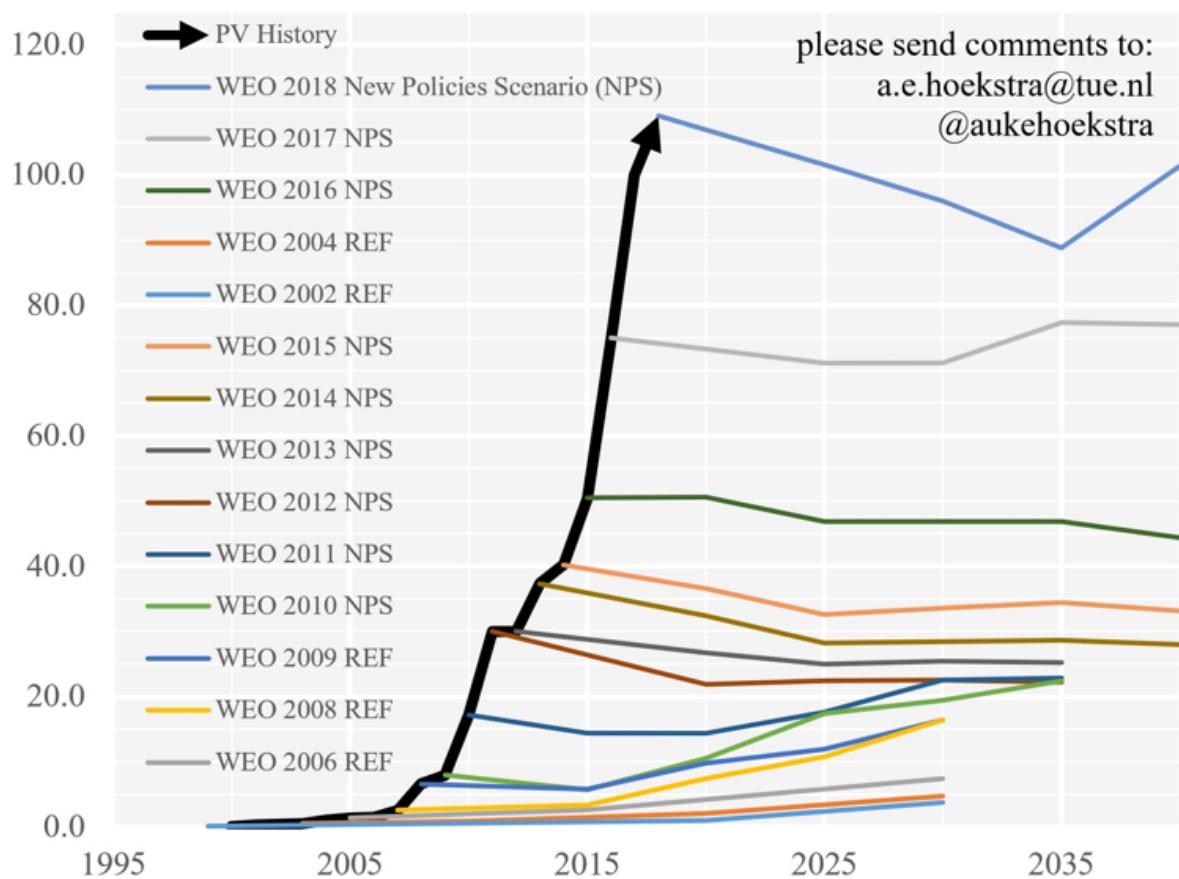
- [Yngve Høiseth](#) releases a prediction scorer, based on his previous work on Empiricast. In Python, but also available as a [REST API](#)

## Negative Examples.

- The International Energy Agency had terrible forecasts on solar photo-voltaic energy production, until [recently](#):

### Annual PV additions: historic data vs IEA WEO predictions

In GW of added capacity per year - source International Energy Agency - World Energy Outlook



...It's a scenario assuming current policies are kept and no new policies are added.

...the discrepancy basically implies that every year loads of unplanned subsidies are added... So it boils down to: it's not a forecast and any error you find must be attributed to that. And no you cannot see how the model works.

The IEA website explains the WEO process: "The detailed projections are generated by the World Energy Model, a large-scale simulation tool, developed at the IEA over a period of more than 20 years that is designed to replicate how energy markets function."

## News & Hard to Categorize Content.

- [Budget credibility of subnational forecasts](#).

Budget credibility, or the ability of governments to accurately forecast macro-fiscal variables, is crucial for effective public finance management. Fiscal marksmanship analysis captures the extent of errors in the budgetary forecasting... Partitioning the sources of errors, we identified that the errors were more broadly random than due to systematic bias, except for a few crucial macro-fiscal variables where improving the forecasting techniques can provide better estimates.

- See also: [How accurate are \[US\] agencies' procurement forecasts?](#) and [Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods](#) (which finds random forests a hard to beat approach)
- [Bloomberg on the IMF's track record on forecasting \(archive link, without a paywall\)](#).

A Bloomberg analysis of more than 3,200 same-year country forecasts published each spring since 1999 found a wide variation in the direction and magnitude of errors. In 6.1 percent of cases, the IMF was within a 0.1 percentage-point margin of error. The rest of the time, its forecasts underestimated GDP growth in 56 percent of cases and overestimated it in 44 percent. The average forecast miss, regardless of direction, was 2.0 percentage points, but obscures a notable difference between the average 1.3 percentage-point error for advanced economies compared with 2.1 percentage points for more volatile and harder-to-model developing economies. Since the financial crisis, however, the IMF's forecast accuracy seems to have improved, as growth numbers have generally fallen.

Banking and sovereign debt panics hit Greece, Ireland, Portugal and Cyprus to varying degrees, threatening the integrity of the euro area and requiring emergency intervention from multinational authorities. During this period, the IMF wasn't merely forecasting what would happen to these countries but also setting the terms. It provided billions in bailout loans in exchange for implementation of strict austerity measures and other policies, often bitterly opposed by the countries' citizens and politicians.

- I keep seeing evidence that Trump will lose reelection, but I don't know how seriously to take it, because I don't know how filtered it is.
- For example, the [The Economist's model](#) forecasts 91% that Biden will win the upcoming USA elections. Should I update somewhat towards Biden winning after seeing it? What if I suspect that it's the most extreme model, and that it has come to my attention because of that fact? What if I suspect that it's the most extreme model which will predict a democratic win? What if there was another equally reputable model which predicts 91% for Trump, but which I never got to see because of information filter dynamics?
- The [the Primary Model](#) confirmed my suspicions of filter dynamics. It "does not use presidential approval or the state of the economy as predictors. Instead it relies on the performance of the presidential nominees in primaries", and on how many terms the party has controlled the White House. The model has been developed by an [otherwise unremarkable](#) professor of political science at New York's Stony Brook University, and has done well in previous election cycles. It assigns 91% to Trump winning reelection.
- [Forecasting at Uber: An Introduction](#). Uber forecasts demand so that they know amongst other things, when and where to direct their vehicles. Because of the

challenges to testing and comparing forecasting frameworks at scale, they developed their own software for this.

- [Forecasting Sales In These Uncertain Times.](#)

[...] a company selling to lower-income consumers might use the monthly employment report for the U.S. to see how people with just a high school education are doing finding jobs. A business selling luxury goods might monitor the stock market.

- [Unilever Chief Supply Officer on forecasting](#): "Agility does trump forecasting. At the end of the day, every dollar we spent on agility has probably got a 10x return on every dollar spent on forecasting or scenario planning."

An emphasis on agility over forecasting meant shortening planning cycles — the company reduced its planning horizon from 13 weeks to four. The weekly planning meeting became a daily meeting. Existing demand baselines and even artificial intelligence programs no longer applied as consumer spending and production capacity strayed farther from historical trends.

- [An updated introduction to prediction markets](#), yet one which contains some nuggets I didn't know about.

This bias toward favorable outcomes... appears for a wide variety of negative events, including diseases such as cancer, natural disasters such as earthquakes and a host of other events ranging from unwanted pregnancies and radon contamination to the end of a romantic relationship. It also emerges, albeit less strongly, for positive events, such as graduating from college, getting married and having favorable medical outcomes.

Nancy Reagan hired an astrologer, Joan Quigley, to screen Ronald Reagan's schedule of public appearances according to his horoscope, allegedly in an effort to avoid assassination attempts.

Google, Yahoo!, Hewlett-Packard, Eli Lilly, Intel, Microsoft, and France Telecom have all used internal prediction markets to ask their employees about the likely success of new drugs, new products, future sales.

Although prediction markets can work well, they don't always. IEM, PredictIt, and the other online markets were wrong about Brexit, and they were wrong about Trump's win in 2016. As the Harvard Law Review points out, they were also wrong about finding weapons of mass destruction in Iraq in 2003, and the nomination of John Roberts to the U.S. Supreme Court in 2005. There are also plenty of examples of small groups reinforcing each other's moderate views to reach an extreme position, otherwise known as groupthink, a theory devised by Yale psychologist Irving Janis and used to explain the Bay of Pigs invasion.

although thoughtful traders should ultimately drive the price, that doesn't always happen. The [prediction] markets are also no less prone to being caught in an information bubble than British investors in the South Sea Company in 1720 or speculators during the tulip mania of the Dutch Republic in 1637.

- [Food Supply Forecasting Company gets \\$12 million in Series A funding](#)

## Long Content.

- [Michael Story](#), "Jotting down things I learned from being a superforecaster."

Small teams of smart, focused and rational generalists can absolutely smash big well-resourced institutions at knowledge production, for the same reasons startups can beat big rich incumbent businesses

There's a *lot* more to making predictive accuracy work in practice than winning a forecasting tournament. Competitions are about daily fractional updating, long lead times and exhaustive pre-forecast research on questions especially chosen for competitive suitability

Real life forecasting often requires fast turnaround times, fuzzy questions, and difficult-to-define answers with unclear resolution criteria. In a competition, a question with ambiguous resolution is thrown out, but in a crisis it might be the most important work you do

- Lukas Gloor on [takeaways from Covid forecasting on Metaculus](#)
- [Ambiguity aversion](#). "Better the devil you know than the devil you don't."

An ambiguity-averse individual would rather choose an alternative where the probability distribution of the outcomes is known over one where the probabilities are unknown. This behavior was first introduced through the [Ellsberg paradox](#) (people prefer to bet on the outcome of an urn with 50 red and 50 blue balls rather than to bet on one with 100 total balls but for which the number of blue or red balls is unknown).

- Gregory Lewis: [Use uncertainty instead of imprecision.](#)

If your best guess for X is 0.37, but you're very uncertain, you still shouldn't replace it with an imprecise approximation (e.g. "roughly 0.4", "fairly unlikely"), as this removes information. It is better to offer your precise estimate, alongside some estimate of its resilience, either subjectively ("0.37, but if I thought about it for an hour I'd expect to go up or down by a factor of 2"), or objectively ("0.37, but I think the standard error for my guess to be ~0.1").

- [Expert Forecasting with and without Uncertainty Quantification and Weighting: What Do the Data Say?](#): "it's better to combine expert uncertainties (e.g. 90% confidence intervals) than to combine their point forecasts, and it's better still to combine expert uncertainties based on their past performance."
  - See also a [1969 paper](#) by future Nobel Prize winner Clive Granger: "Two separate sets of forecasts of airline passenger data have been combined to form a composite set of forecasts. The main conclusion is that the composite set of forecasts can yield lower mean-square error than either of the original forecasts. Past errors of each of the original forecasts are used to determine the weights to attach to these two original forecasts in forming the combined forecasts, and different methods of deriving these weights are examined".
- [How to build your own weather forecasting model](#). Sailors realize that weather forecasting are often corrupted by different considerations (e.g., a reported 50% of rain doesn't happen 50% of the time), and search for better sources. One such source is the original, raw data used to generate weather forecasts: GRIB files (Gridded Information in Binary), which lack interpretation. But these have their own pitfalls, which sailors must learn to take into account. For example, GRIB files only take into account wind speed, not tidal acceleration, which can cause a significant increase in apparent wind.

'Forecasts are inherently political,' says Dashew. 'They are the result of people perhaps getting it wrong at some point so some pressures to interpret them in a different or more conservative way very often. These pressures change all the time so they are often subject to outside factors.'

Singleton says he understands how pressures on forecasters can lead to this opinion being formed: 'In my days at the Met Office when the Shipping Forecast used to work under me, they always said they try to tell it like it is and they do not try to make it sound worse.'

- [Forecasting the dividends of conflict prevention from 2020 - 2030](#). Study quantifies the dynamics of conflict, building a transition matrix between different states (peace, high risk, negative peace, war, and recovery) and validating it using historical dataset; they find (concurring with the previous literature), that countries have a tendency to fall into cycles of conflict. They conclude that changing this transition matrix would have a very high impact. Warning: extensive quoting follows.

Notwithstanding the mandate of the United Nations to promote peace and security, many member states are still sceptical about the dividends of conflict prevention. Their diplomats argue that it is hard to justify investments without being able to show its tangible returns to decision-makers and taxpayers. As a result, support for conflict prevention is halting and uneven, and governments and international agencies end up spending enormous sums in stability and peace support operations after-the-fact.

This study considers the trajectories of armed conflict in a 'business-as-usual' scenario between 2020-2030. Specifically, it draws on a comprehensive historical dataset to determine the number of countries that might experience rising levels of collective violence, outright armed conflict, and their associated economic costs. It then simulates alternative outcomes if conflict prevention measures were 25%, 50%, and 75% more effective. As with all projections, the quality of the projections relies on the integrity of the underlying data. The study reviews several limitations of the analysis, and underlines the importance of a cautious interpretation of the findings.

If current trends persist and no additional conflict prevention action is taken above the current baseline, then it is expected that there will be three more countries at war and nine more countries at high risk of war by 2030 as compared to 2020. This translates into roughly 677,250 conflict-related fatalities (civilian and battle-deaths) between the present and 2030. By contrast, under our most pessimistic scenario, a 25% increase in effectiveness of conflict prevention would result in 10 more countries at peace by 2030, 109,000 fewer fatalities over the next decade and savings of over \$3.1 trillion. A 50% improvement would result in 17 additional countries at peace by 2030, 205,000 fewer deaths by 2030, and some \$6.6 trillion in savings.

Meanwhile, under our most optimistic scenario, a 75% improvement in prevention would result in 23 more countries at peace by 2030, resulting in 291,000 lives saved over the next decade and \$9.8 trillion in savings. These scenarios are approximations, yet demonstrate concrete and defensible estimates of both the benefits (saved lives, displacement avoided, declining peacekeeping deployments) and cost-effectiveness of prevention (recovery aid, peacekeeping expenditures). Wars are costly and the avoidance of "conflict traps" could save the economy trillions of dollars by 2030 under the most optimistic scenarios. The bottom line is that comparatively modest investments in prevention can yield lasting effects by avoiding compounding costs of lost life, peacekeeping, and aid used for humanitarian response and rebuilding rather than development. The longer conflict prevention is delayed, the more expensive responses to conflict become.

In order to estimate the dividends of conflict prevention we analyze violence dynamics in over 190 countries over the period 1994 to 2017, a time period for which most data was available for most countries. Drawing on 12 risk variables, the model examines the likelihood that a war will occur in a country in the following year and we estimate (through linear, fixed effects regressions) the average cost of war (and other 'states', described below) on 8 dependent variables, including loss of life, displacement, peacekeeping deployments and expenditures, oversea aid and economic growth. The estimates confirm that, by far, the most costly state for a country to be in is war, and the probability of a country succumbing to war in the next year is based on its current state and the frequency of other countries with similar states having entered war in the past.

At the core of the model (and results) is the reality that countries tend to get stuck in so-called violence and conflict traps. A well-established finding in the conflict studies field is that once a country experiences an armed conflict, it is very likely to relapse into conflict or violence within a few years. Furthermore, countries likely to experience war share some common warning signs, which we refer to as "flags" (up to 12 flags can be raised to signal risk). Not all countries that enter armed conflict raise the same warning flags, but the warning flags are nevertheless a good indication that a country is at high risk. These effects create vicious cycles that result in high risk, war and frequent relapse into conflict. Multiple forms of prevention are necessary to break these cycles. The model captures the vicious cycle of conflict traps, through introducing five states and a transition matrix based on historical data (see Table 1). First, we assume that a country is in one of five 'states' in any given year. These 'states' are at "Peace", "High Risk", "Negative Peace", "War" and "Recovery" (each state is described further below). Drawing on historical data, the model assesses the probability of a country transitioning to another state in a given year (a transition matrix).

For example, if a state was at High Risk in the last year, it has a 19.3% chance of transitioning to Peace, a 71.4% chance of staying High Risk, a 7.6% chance of entering Negative Peace and a 1.7% chance of entering War the following year.

By contrast, high risk states are designated by the raising of up to 12 flags. These include: 1) high scores by Amnesty International's annual human rights reports (source: Political Terror Scale), 2) the US State Department annual reports (source: Political Terror Scale), 3) civilian fatalities as a percentage of population (source: ACLED), 4) political events per year (source: ACLED) 5) events attributed to the proliferation of non-state actors (source: ACLED), 6) battle deaths (source: UCDP), 7) deaths by terrorism (source: GTD), 8) high levels of crime (source: UNODC), 9) high levels of prison population (source: UNODC), 10) economic growth shocks (source: World Bank), 11) doubling of displacement in a year (source: IDMC), and 12) doubling of refugees in a year (source: UNHCR). Countries with two or more flags fall into the "high risk" category. Using these flags, a majority of countries have been at high risk for one or more years from 1994 to 2017, so it is easier to give examples of countries that have not been at high risk.

Negative peace states are defined by combined scores from Amnesty International and the US State Department. Countries in negative peace are more than five times as likely to enter high risk in the following year than peace (26.8% vs. 4.1%).

A country that is at war is one that falls into a higher threshold of collective violence, relative to the size of the population. Specifically, it is designated as such if one or more of the following conditions are met: above 0.04 battle deaths or .04 civilian fatalities per 100,000 according to UCDP and ACLED, respectively, or coding of genocide by the Political Instability Task Force Worldwide Atrocities Dataset.

Countries experiencing five or more years of war between 1994 and 2017 included Afghanistan, Somalia, Sudan, Iraq, Burundi, Central African Republic, Sri Lanka, DR Congo, Uganda, Chad, Colombia, Israel, Lebanon, Liberia, Yemen, Algeria, Angola, Sierra Leone, South Sudan, Eritrea and Libya.

Lastly, recovery is a period of stability that follows from war. A country is only determined to be recovering if it is not at war and was recently in a war. Any country that exits in the war state is immediately coded as being in recovery for the following five years, unless it relapses into war. The duration of the recovery period (five years) is informed by the work of Paul Collier et al, but is robust also to sensitivity tests around varying recovery lengths.

The model does not allow for countries to be high risk and in recovery in the same year, but there is ample evidence that countries that are leaving a war state are at a substantially higher risk of experiencing war recurrence, contributing to the conflict trap described earlier. Countries are twice as likely to enter high risk or negative peace coming out of recovery as they are to enter peace, and 10.2% of countries in recovery relapse into war every year. When a country has passed the five year threshold without reverting to war, it can move back to states of peace, negative peace or high risk.

The transition matrix underlines the very real risk of countries falling into a 'conflict trap'. Specifically, a country that is in a state of war has a very high likelihood of staying in this condition in the next year (72.6%) and just a 27.4% chance of transitioning to recovery. Once in recovery, a country has a 10.2% chance of relapse every year, suggesting only a 58% chance ( $1-10.2\% \times 5$ ) that a country will not relapse over five years.

As Collier and others have observed, countries are often caught in prolonged and vicious cycles of war and recovery (conflict traps), often unable to escape into a new, more peaceful (or less war-like) state

- War is expensive. So is being at high risk of war.

Of course, the loss of life, displacement, and accumulated misery associated with war should be reason enough to invest in prevention, but there are also massive economic benefits from successful prevention. Foremost, the countries at war avoid the costly years in conflict, with growth rates 4.8% lower than countries at peace. They also avoid years of recovery and the risk of relapse into conflict. Where prevention works, conflict-driven humanitarian needs are reduced, and the international community avoids peacekeeping deployments and additional aid burdens, which are sizable.

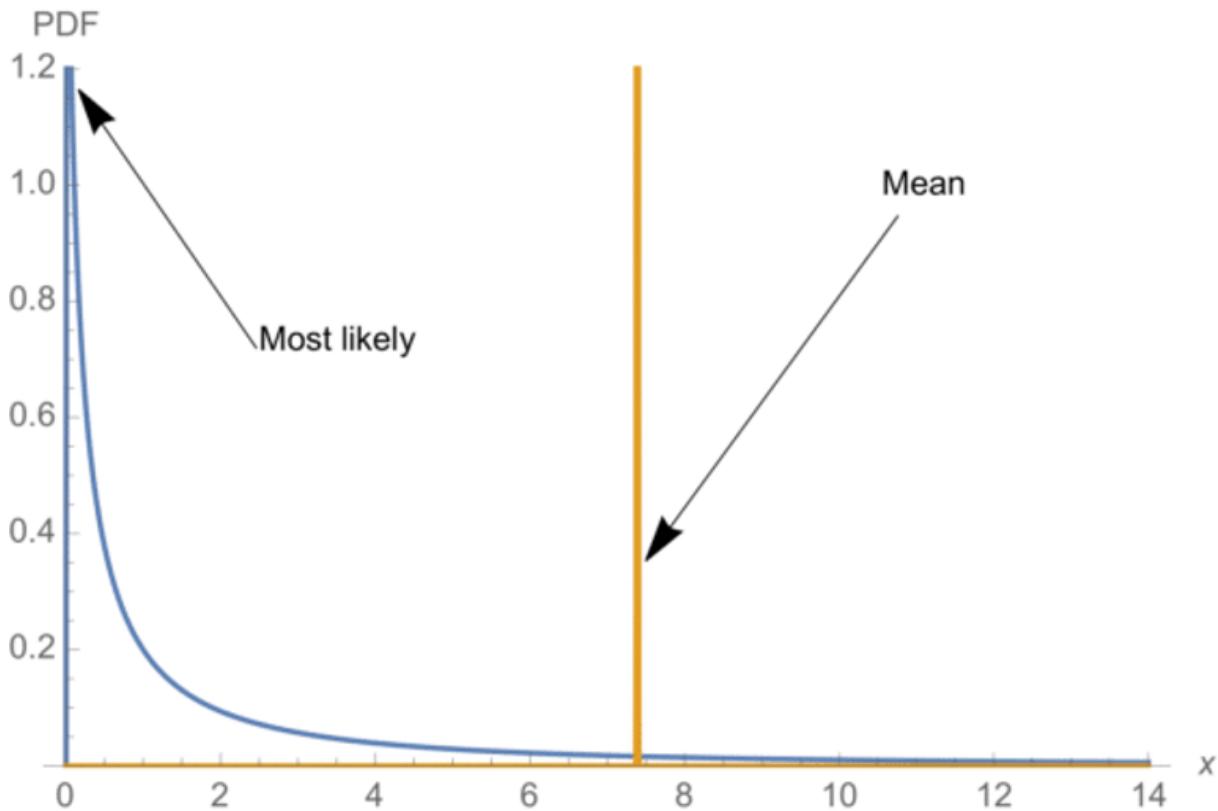
Conclusion: The world can be significantly better off by addressing the high risk of destructive violence and war with focused efforts at prevention in countries at high risk and those in negative peace. This group of countries has historically been at risk of higher conflict due to violence against civilians, proliferation of armed groups, abuses of human rights, forced displacement, high homicide, and incidence of error. None of this is surprising. Policymakers know that war is bad for humans and other living things. What is staggering is the annual costs of war that we will continue to pay in 2030 through inaction today – conceivably trillions of dollars of economic growth, and the associated costs of this for human security and development, are being swept off the table by the decisions made today to ignore prevention.

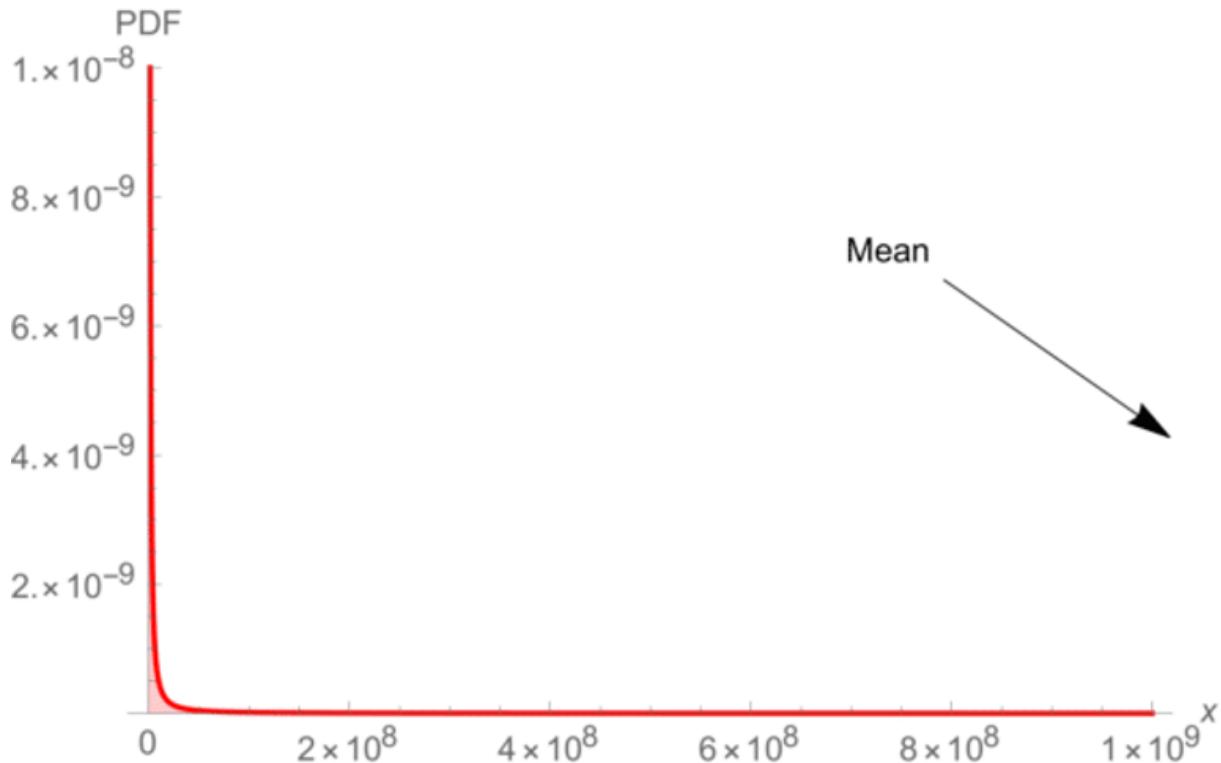
- [COVID-19: Ioannidis vs. Taleb](#)

On the one hand, Nassim Taleb has clearly expressed that measures to stop the spread of the pandemic must be taken as soon as possible: instead of looking at data, it is the nature of a pandemic with a possibility of devastating human impact that should drive our decisions.

On the other hand, John Ioannidis acknowledges the difficulty in having good data and of producing accurate forecasts, while believing that eventually any information that can be extracted from such data and forecasts should still be useful, e.g. to having targeted lockdowns (in space, time, and considering the varying risk for different segments of the population).

- [Taleb](#): *On single point forecasts for fat tailed variables.* Leitmotiv: Pandemics are fat-tailed.





We do not need more evidence under fat tailed distributions — it is there in the properties themselves (properties for which we have ample evidence) and these clearly represent risk that must be killed in the egg (when it is still cheap to do so). Secondly, unreliable data — or any source of uncertainty — should make us follow the most paranoid route. [...] more uncertainty in a system makes precautionary decisions very easy to make (if I am uncertain about the skills of the pilot, I get off the plane).

Random variables in the power law class with tail exponent  $\alpha \leq 1$  are, simply, not forecastable. They do not obey the [Law of Large Numbers]. But we can still understand their properties.

As a matter of fact, owing to preasymptotic properties, a heuristic is to consider variables with up to  $\alpha \leq 5/2$  as not forecastable — the mean will be too unstable and requires way too much data for it to be possible to do so in reasonable time. It takes  $10^{14}$  observations for a “Pareto 80/20” (the most commonly referred to probability distribution, that is with  $\alpha \approx 1.13$ ) for the average thus obtained to emulate the significance of a Gaussian with only 30 observations.

- [Ioannidis](#): *Forecasting for COVID-19 has failed*. Leitmotiv: "Investment should be made in the collection, cleaning and curation of data".

Predictions for hospital and ICU bed requirements were also entirely misinforming. Public leaders trusted models (sometimes even black boxes without disclosed methodology) inferring massively overwhelmed health care capacity (Table 1) [3]. However, eventually very few hospitals were stressed, for a couple of weeks. Most hospitals maintained largely empty wards, waiting for tsunamis that never came. The general population was locked and placed in horror-alert to save the health system from collapsing. Tragically, many health systems faced major adverse consequences, not by COVID-19 cases overload, but for very different reasons. Patients with heart

attacks avoided visiting hospitals for care [4], important treatments (e.g. for cancer) were unjustifiably delayed [5], mental health suffered [6]. With damaged operations, many hospitals started losing personnel, reducing capacity to face future crises (e.g. a second wave). With massive new unemployment, more people may lose health insurance. The prospects of starvation and of lack of control for other infectious diseases (like tuberculosis, malaria, and childhood communicable diseases for which vaccination is hindered by the COVID-19 measures) are dire...

The core evidence to support “flatten-the-curve” efforts was based on observational data from the 1918 Spanish flu pandemic on 43 US cities. These data are >100-years old, of questionable quality, unadjusted for confounders, based on ecological reasoning, and pertaining to an entirely different (influenza) pathogen that had ~100-fold higher infection fatality rate than SARS-CoV-2. Even thus, the impact on reduction on total deaths was of borderline significance and very small (10-20% relative risk reduction); conversely many models have assumed 25-fold reduction in deaths (e.g. from 510,000 deaths to 20,000 deaths in the Imperial College model) with adopted measures

Despite these obvious failures, epidemic forecasting continued to thrive, perhaps because vastly erroneous predictions typically lacked serious consequences. Actually, erroneous predictions may have been even useful. A wrong, doomsday prediction may incentivize people towards better personal hygiene. Problems starts when public leaders take (wrong) predictions too seriously, considering them crystal balls without understanding their uncertainty and the assumptions made. Slaughtering millions of animals in 2001 aggravated a few animal business stakeholders, most citizens were not directly affected. However, with COVID-19, espoused wrong predictions can devastate billions of people in terms of the economy, health, and societal turmoil at-large.

Cirillo and Taleb thoughtfully argue [14] that when it comes to contagious risk, we should take doomsday predictions seriously: major epidemics follow a fat-tail pattern and extreme value theory becomes relevant. Examining 72 major epidemics recorded through history, they demonstrate a fat-tailed mortality impact. However, they analyze only the 72 most noticed outbreaks, a sample with astounding selection bias. The most famous outbreaks in human history are preferentially selected from the extreme tail of the distribution of all outbreaks. Tens of millions of outbreaks with a couple deaths must have happened throughout time. Probably hundreds of thousands might have claimed dozens of fatalities. Thousands of outbreaks might have exceeded 1,000 fatalities. Most eluded the historical record. The four garden variety coronaviruses may be causing such outbreaks every year [15,16]. One of them, OC43 seems to have been introduced in humans as recently as 1890, probably causing a “bad influenza year” with over a million deaths [17]. Based on what we know now, SARS-CoV-2 may be closer to OC43 than SARS-CoV-1. This does not mean it is not serious: its initial human introduction can be highly lethal, unless we protect those at risk.

- The (British) Royal Economic Society presents a panel on [What is a scenario, projection and a forecast - how good or useful are they particularly now?](#). The start seems promising: "My professional engagement with economic and fiscal forecasting was first as a consumer, and then a producer. I spent a decade happily mocking other people's efforts, as a journalist, since when I've spent two decades helping colleagues to construct forecasts and to try to explain them to the public." The first speaker, which corresponds to the first ten minutes, is worth listening to; the rest varies in quality.

You have to construct the forecast and explain it in a way that's fit for that purpose

- I liked the following taxonomy of what distinct targets the agency the first speaker works for is aiming to hit with their forecasts:

1. as an input into the policy-making process,
2. as a transparent assessment of public finances
3. as a prediction of whether the government will meet whatever fiscal rules it has set itself,
4. as a baseline against which to judge the significance of further news,
5. as a challenge to other agencies "to keep the bastards honest".

- The limitations were interesting as well:

1. they require us to produce a forecast that's conditioned on current government policy even if we and everyone else expect that policy to change that of course makes it hard to benchmark our performance against counterparts who are producing unconditional forecasts.
2. The forecasts have to be explainable; a black box model might be more accurate but be less useful.
3. they require detailed discussion of the individual forecast lines and clear diagnostics to explain changes from one forecast to the next precisely to reassure people that those changes aren't politically motivated or tainted - the forecast is as much about delivering transparency and accountability as about demonstrating predictive prowess
4. the forecast numbers really have to be accompanied by a comprehensible narrative of what is going on in the economy and the public finances and what impact policy will have - Parliament and the public needs to be able to engage with the forecast we couldn't justify our predictions simply with an appeal to a statistical black box and the Chancellor certainly couldn't justify significant policy positions that way.

"horses for courses, the way you do the forecast, the way you present it depends on what you're trying to achieve with it"

"People use scenario forecasting in a very informal manner. which I think that could be problematic because it's very difficult to basically find out what are the assumptions and whether those assumptions and the models and the laws can be validated"

Linear models are state independent, but it's not the same to receive a shock where the economy is in upswing as when the economy is during a recession.

- Some situations are too complicated to forecast, so one conditions on some variables being known, or following a given path, and then studies the rest, calling the output a "scenario."

One week delay in intervention by the government makes a big difference to the height of the [covid-19] curve.

I don't think it's easy to follow the old way of doing things. I'm sorry, I have to be honest with you. I spent 4 months just thinking about this problem and you need to

integrate a model of the social behavior and how you deal with the risk to health and to economy in these models. But unfortunately, by the time we do that it won't be relevant.

It amuses me to look at weather forecasts because economists don't have that kind of technology, those kind of resources.

---

Note to the future: All links are added automatically to the Internet Archive. In case of link rot, go [here](#)

---

"horses for courses, the way you do the forecast, the way you present it depends on what you're trying to achieve with it"

---

# Interpretability in ML: A Broad Overview

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

(Reposting because I think a GreaterWrong bug on submission made this post invisible for a while last week so I'm trying again on LW.)

This blog post is an overview of ways to think about machine learning interpretability; it covers some recent research techniques as well as directions for future research. This is an updated version of [this post](#) from a few weeks ago. I've now added code, examples, and some pictures.

## What Are Existing Overviews?

Many of these ideas are based heavily off of Zach Lipton's [Mythos of Model Interpretability](#), which I think is the best paper for understanding the different definitions of interpretability. For a deeper dive into specific techniques, I recommend [A Survey Of Methods For Explaining Black Box Models](#) which covers a wide variety of approaches for many different ML as well as model-agnostic approaches. For neural nets specifically, [Explainable Deep Learning: A Field Guide for the Uninitiated](#) provides an in-depth read. For other conceptual surveys of the field, [Definitions, methods, and applications in interpretable machine learning](#) and [Explainable Machine Learning for Scientific Insights and Discoveries](#). The Explainable Machine Learning paper in particular is quite nice because it gives a hierarchy of increasingly more interpretable models across several domains and use cases.

(Shout-out to [Connected Papers](#) which made navigating the paper landscape for interpretability very bearable.)

As always, you can find code used to generate the images [here](#) on GitHub.

In the rest of this post, we'll go over many ways to formalize what "interpretability" means. Broadly, interpretability focuses on the *how*. It's focused on getting some notion of an explanation for the decisions made by our models. Below, each section is operationalized by a concrete question we can ask of our ML model using a specific definition of interpretability. Before that, though, if you're new to all this, I'll explain briefly about why we might care about interpretability at all.

## Why Care About Interpretability?

Firstly, interpretability in ML is useful because it can aid in **trust**. As humans, we may be reluctant to rely on ML models for certain critical tasks, e.g. medical diagnosis, unless we know "how they work". There's often a fear of unknown unknowns when trusting in something opaque, which we see when people confront new technology. Approaches to interpretability which focus on transparency could help mitigate some of these fears.

Secondly, **safety**. There is almost always some sort of [shift in distributions](#) between model training and deployment. Failures to generalize or Goodhart's Law issues like [specification gaming](#) are still open problems that could lead to issues in the near future. Approaches to interpretability which explain the model's representations or which features are most relevant could help diagnose these issues earlier and provide more opportunities to intervene.

Thirdly, and perhaps most interestingly, **contestability**. As we delegate more decision-making to ML models, it becomes important for people to appeal these decisions made. Black-box models provide no such recourse because they don't decompose into anything that *can* be contested. This has already led to major criticism of proprietary recidivism predictors like [COMPAS](#). Approaches to interpretability which focus on decomposing the model into sub-models or explicate a chain of reasoning could help with such appeals.

# Defining Interpretability

Lipton's paper breaks interpretability down into two types, transparency and post-hoc.

## Transparency Interpretability

These three questions are from Lipton's section on transparency as interpretability, where he features on properties of the model that are useful to understand and can be known before training begins.

### Can a human walk through the model's steps? (Simulability)

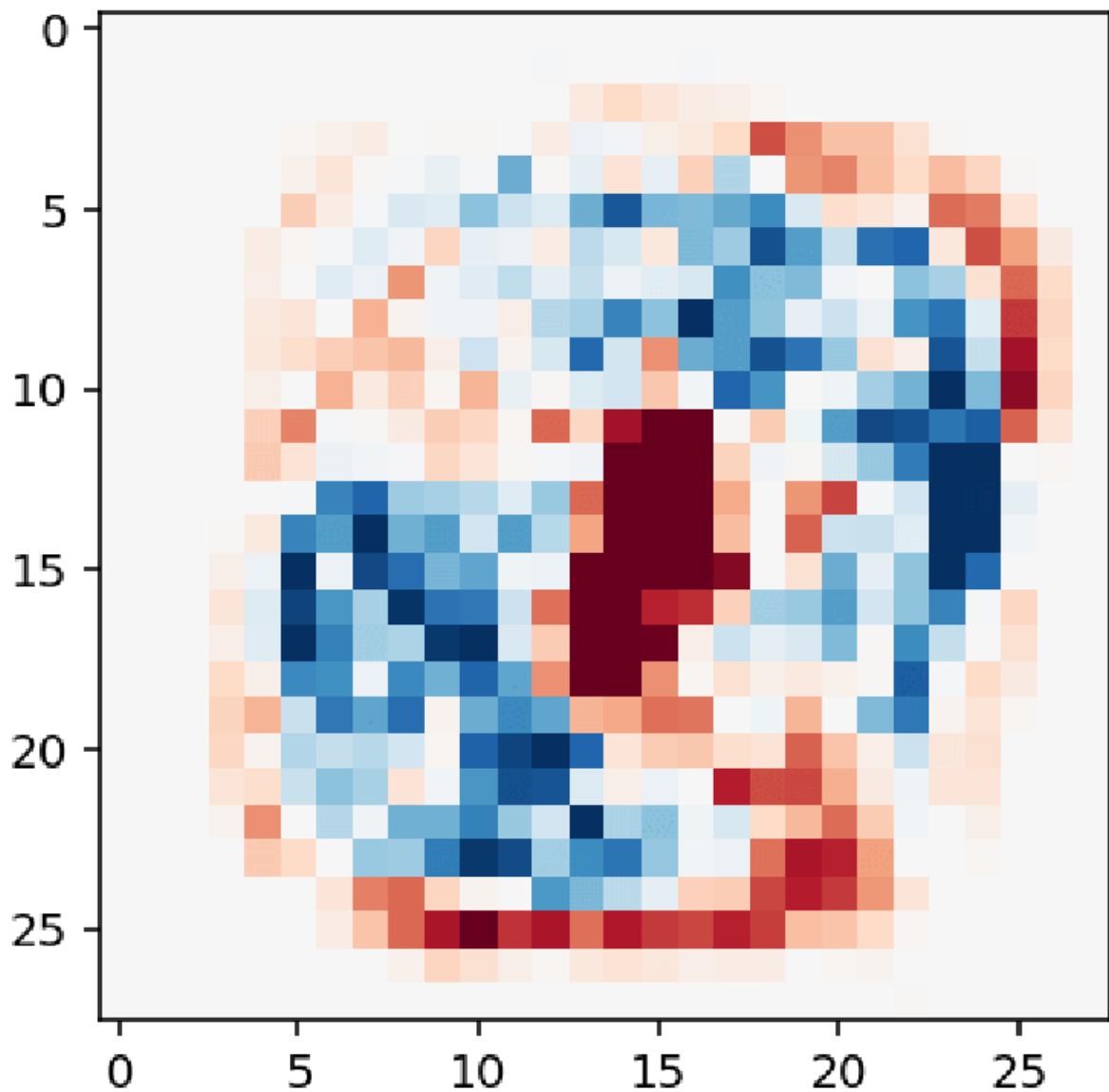
This property is about whether or not a human could go through each step of the algorithm and have it make sense to them at each step. Linear models and decision trees are often cited as interpretable models using such justifications; the computation they require is simple, no fancy matrix operations or nonlinear transformations.

Linear models are also nice because the parameters themselves have a very direct mapping—they represent how important different input features are. For example, I trained a linear classifier on MNIST, and here are some of the weights, each of which correspond to a pixel value:

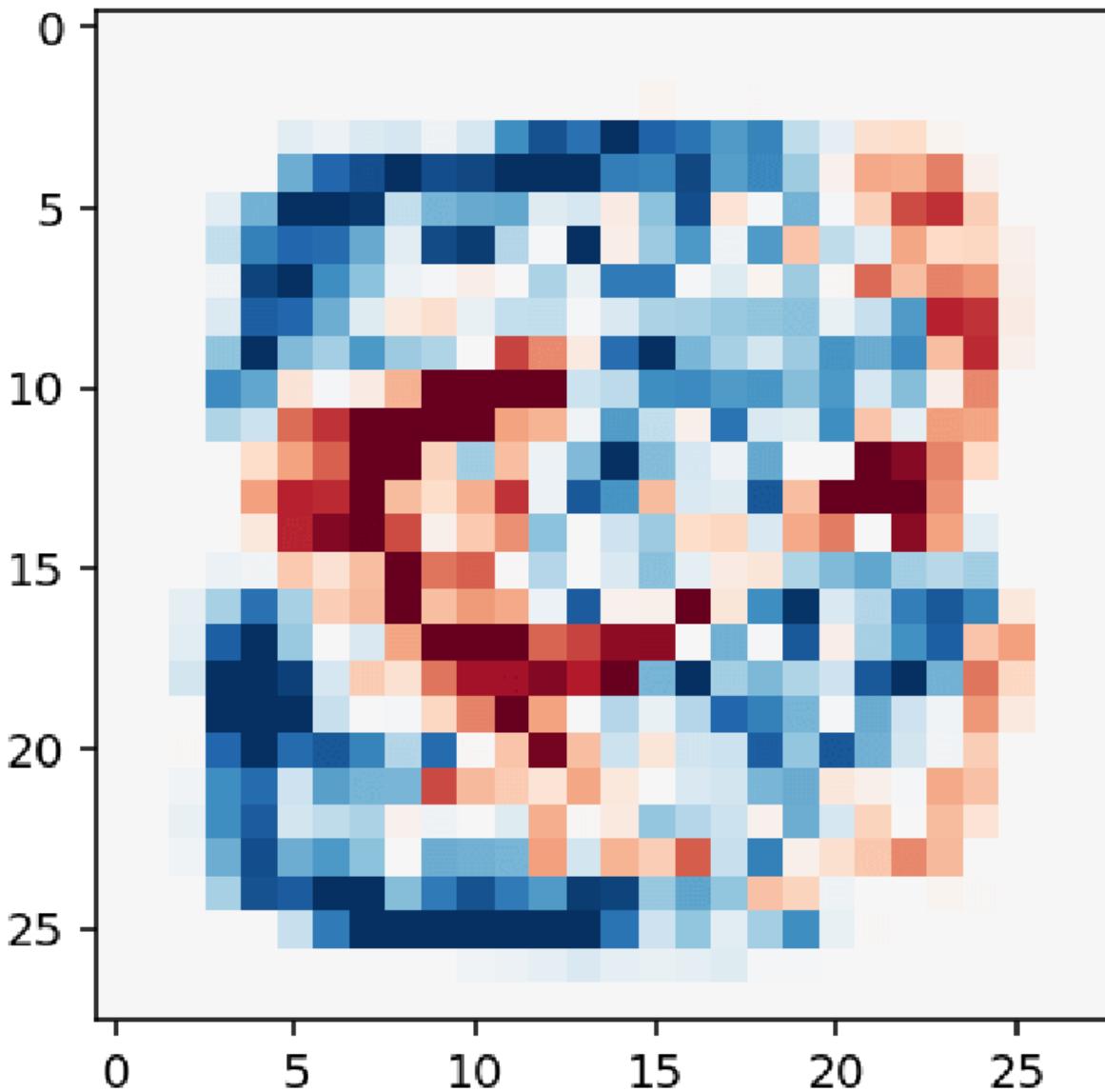
```
0.00000000e+00, 0.00000000e+00, 3.90594519e-05, 7.10306823e-05,  
0.00000000e+00, 0.00000000e+00, 0.00000000e+00, -1.47542413e-03,  
-1.67811041e-04, -3.83280468e-02, -8.10846867e-02, -5.01943218e-02,  
-2.90314621e-02, -2.65494116e-02, -8.29385683e-03, 0.00000000e+00,  
0.00000000e+00, 1.67390785e-04, 3.92789141e-04, 0.00000000e+00,  
0.00000000e+00, 0.00000000e+00, 0.00000000e+00, 0.00000000e+00
```

By themselves, these weights are hard to interpret. Even if we knew which pixels they corresponded to, it's difficult to try and pin down what a certain pixel even represents for large images. However, there is an easy trick to turn these weights into something interpretable. We simply reshape them into the same shape as our model and view it as an image, with the pixel color represented by the weight value.

Here are the weights for the model that looks for 0:



And here are the weights for the model that looks for 3:



In both cases, we can see that the blue regions, which represent positive weight, correspond to a configuration of pixels that look roughly like the digit being detected for. In the case of 0, we can see a distinct blank spot in the center of the image and a curve-like shape around it, whereas the curves of the 3 are also apparent.

However, Lipton points out that this desiderata can be less about the specific choice of model and more about the *size* of the model. A decision tree with a billion nodes, for example, may still be difficult to understand. Understanding is also about being able to hold most of the model in your mind, which is often about how the model is parameterized.

One approach towards achieving this for neural nets is [tree regularization](#) which adds a regularization term that corresponds (roughly) to the size of the decision tree that can approximate the net being trained. The hope here is to eventually output a shallow decision tree that performs comparably to a neural net. Another approach is [neural](#)

[backed decision trees](#) which use another type of regularization to learn a hierarchy over class labels, which then get used to form a decision tree.

Of course, parameterization is not the whole story. There are methods like K-Nearest Neighbors which are parameterized by your entire dataset; this could be billions of points. Yet, there is a sense in which KNN is still interpretable despite its massive size. We can cleanly describe what the algorithm does, and we can even see "why" it made such a choice because the algorithm is so simple.

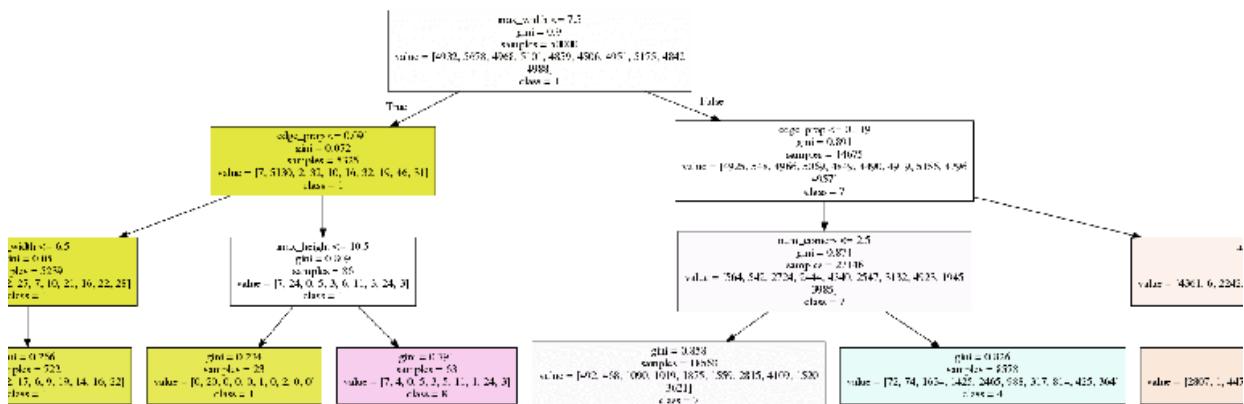
## Is the model interpretable at every step, or with regards to its sub-components? (Decomposability)

Another desirable feature would be to understand what the model is doing at each step. For example, imagine a decision tree whose nodes correspond to easily identifiable factors like age or height. This can sometimes be difficult because model performance is very tightly coupled with the representations used. Raw features, e.g. RGB pixel values, are often not very interpretable by themselves, but interpretable features may not be the most informative for the model.

For example, I trained a decision tree for MNIST using the following interpretable features:

1. The average brightness of the image - avg\_lumin
2. The average brightness of the image's outline (found using an edge detector) - edge\_prop
3. The number of corners found in the image's outline num\_corners
4. The width of the image - max\_width
5. The height of the image - max\_height

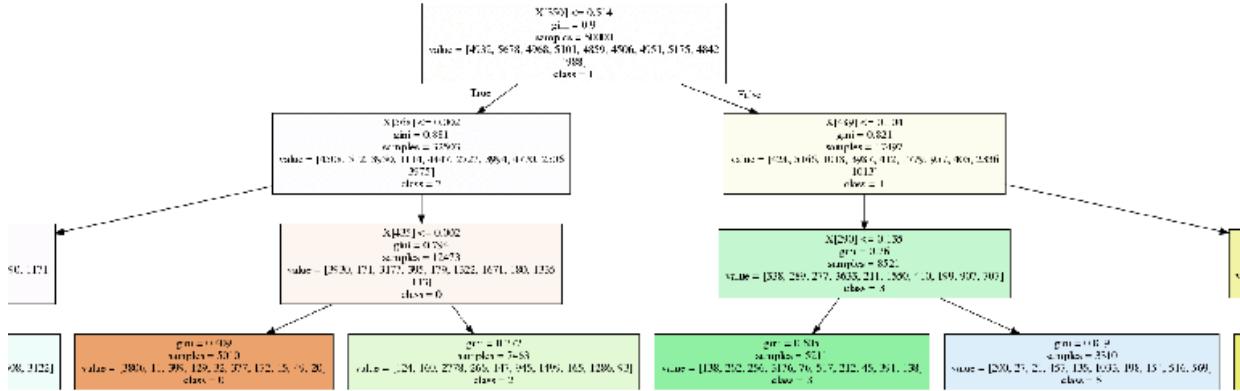
It seems like there would be at least *some* useful information in these features; ones tend to have less area (so avg\_lumin would be lower), eights might have more corners, etc. etc. Yet, the resulting decision tree of depth 3, shown below, however, only achieves 33% training accuracy. Going all the way to depth 10 only bumps it to around 50%.



If we look at the nodes, we can perhaps see what's going on. At the top, we can see that our model will predict a 1 if the width is less than 7.5 pixels, which makes sense as 1 is likely going to be the thinnest digit. Near the bottom, we see that the number of corners is being used to differentiate between 7 and 4. And 4s do have more visual

corners than 7s. But this is very rough, and the overall performance is still not very good.

To compare this with raw features, I also trained a depth 3 decision tree using direct pixel values, i.e. a vector of 784 grayscale values. The resulting model, shown below, gets 50% train and test accuracy.



Here, it's not clear at all why these pixel values were chosen to be the splitting points. And yet the resulting decision tree, for the same number of nodes, does much better. In this simple case, the performance vs interpretability trade-off in representation is quite apparent.

## Does the algorithm itself confer any guarantees? (Algorithmic Transparency)

This asks if our learning algorithm has any properties which make it easy to understand. For example, we might know that the algorithm only outputs sparse models, or perhaps it always converges to a certain type of solution. In these cases, the resulting learned model can be more amenable to analysis. For example, the [Hard Margin SVM](#) is guaranteed to find a unique solution which maximizes the margin. In another vein, the [perceptron](#) is guaranteed to find parameters (not necessarily unique ones, though) that achieve a training loss of 0 if the data are linearly separable.

When it comes to deep learning, I'm less familiar with these kinds of results. My rough understanding is that the equivalence class of models which achieve comparable training error can be quite large, even with regularization, which makes uniqueness results hard to come by.

As I mentioned earlier with KNN, it seems, aside from mechanical transparency, there's another level of understanding regarding "what the algorithm actually does in simple terms". KNN is easy to describe as "it reports the labels of the points closest to the input". The part of this property that's doing the most work here is the way we actually do the describing. Obviously most ML models can be abstracted as "it finds parameters which satisfy certain constraints", but this is very broad. It seems harder to find a description at the same level of granularity for neural nets beyond something like "it learns a high-dimensional manifold that maps onto the input data".

## Post-Hoc Interpretability

These four questions are from Lipton's section on post-hoc interpretability, which focus on things we learn from the model after training has occurred.

## **Can the model give an explanation for its decision, after the fact? (Text Explanation)**

Similar to how humans often give post-hoc justifications for their actions, it could be informative to have models which can also give explanations, perhaps in text. Naive methods of pairing text with decisions, however, are likely going to optimize for something like "how credible the explanation sounds to a human" rather than "how accurate the explanation is at summarizing the internal steps taken".

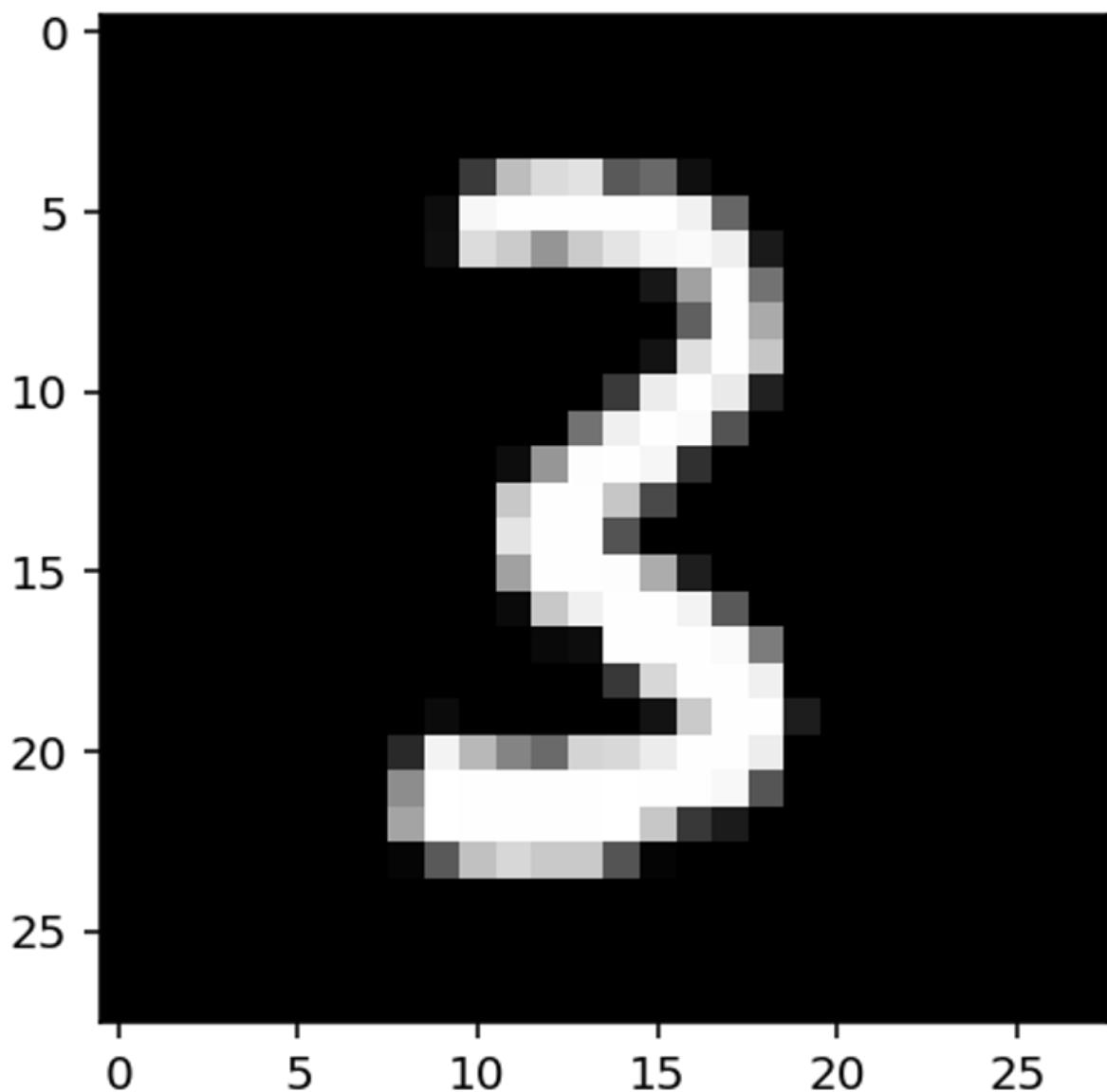
While this seems clearly desirable, I think research in this area is hard to come by, and Lipton only offers one paper that is RL-focused. On ConnectedPapers, I found that said paper is part of a larger related field of [reinforcement learning with human advice](#). This seems to focus on the converse problem—given human explanations, how can models incorporate them into their decision-making? Maybe insights here can eventually be used in the other direction.

## **Can the model identify what is/was important to its decision-making? (Visualization/Local Explanations)**

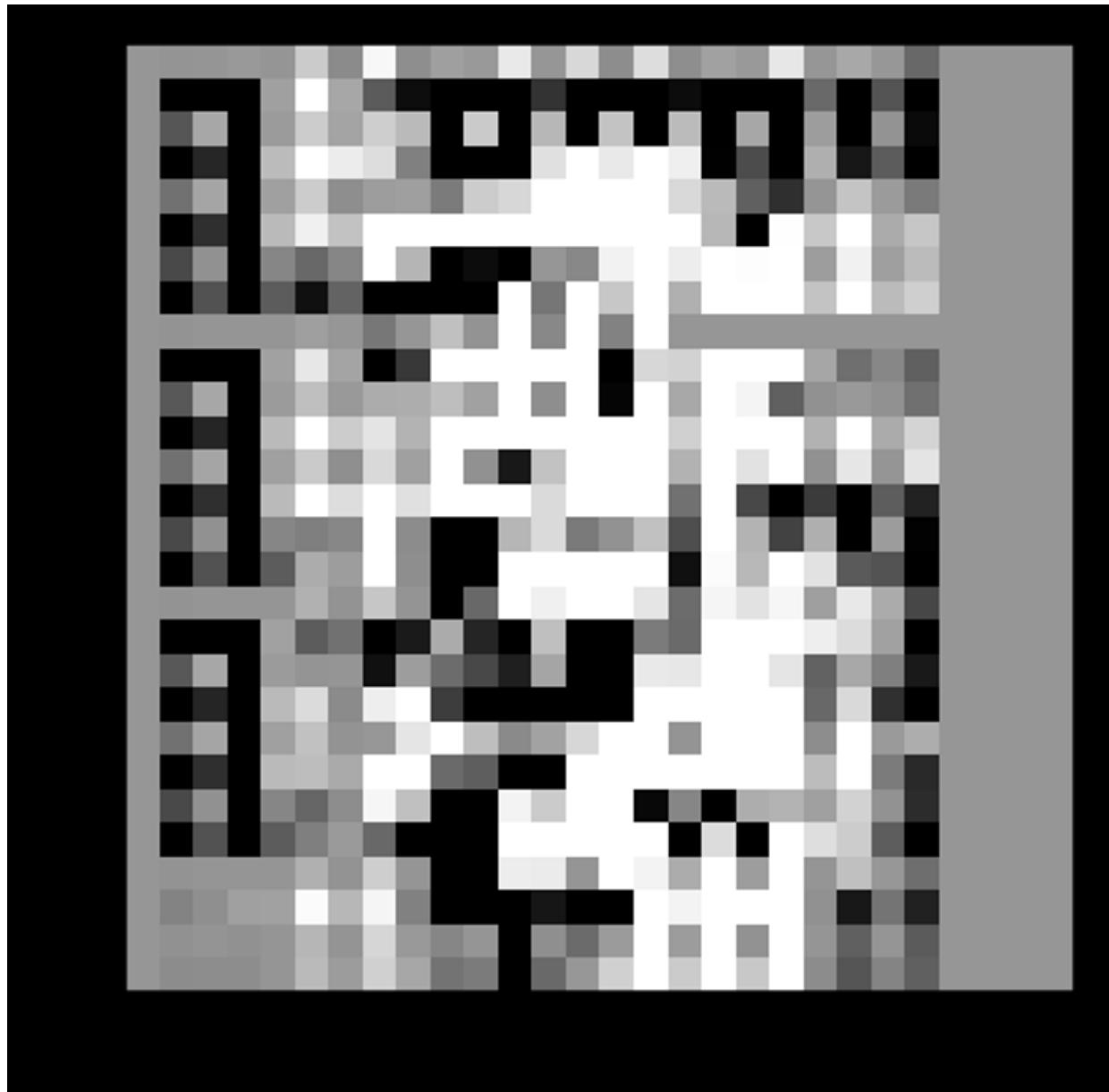
This focuses on how the inputs and outputs change, relative to one another.

Saliency maps are a broad class of approaches that look at where the inputs change in order to change the outputs. A simple way to do this is to take the derivative of the loss function with respect to the input. Past this, there are many modifications which involve averaging the gradient, perturbing the input, or local approximations. [Understanding Deep Networks via Extremal Perturbations and Smooth Masks](#) has a good overview of the work in this area.

For example, I trained a CNN on MNIST and did a simple gradient visualization on an image of this 3:



Using PyTorch, I took the derivative of the logit that corresponds to the class 3 with respect to the input image. This gave me the image below. Here, the white pixels correspond to parts of the image that would increase the logit value for 3, and the black pixels correspond to the reverse. We can see the rough curves of the three come through.

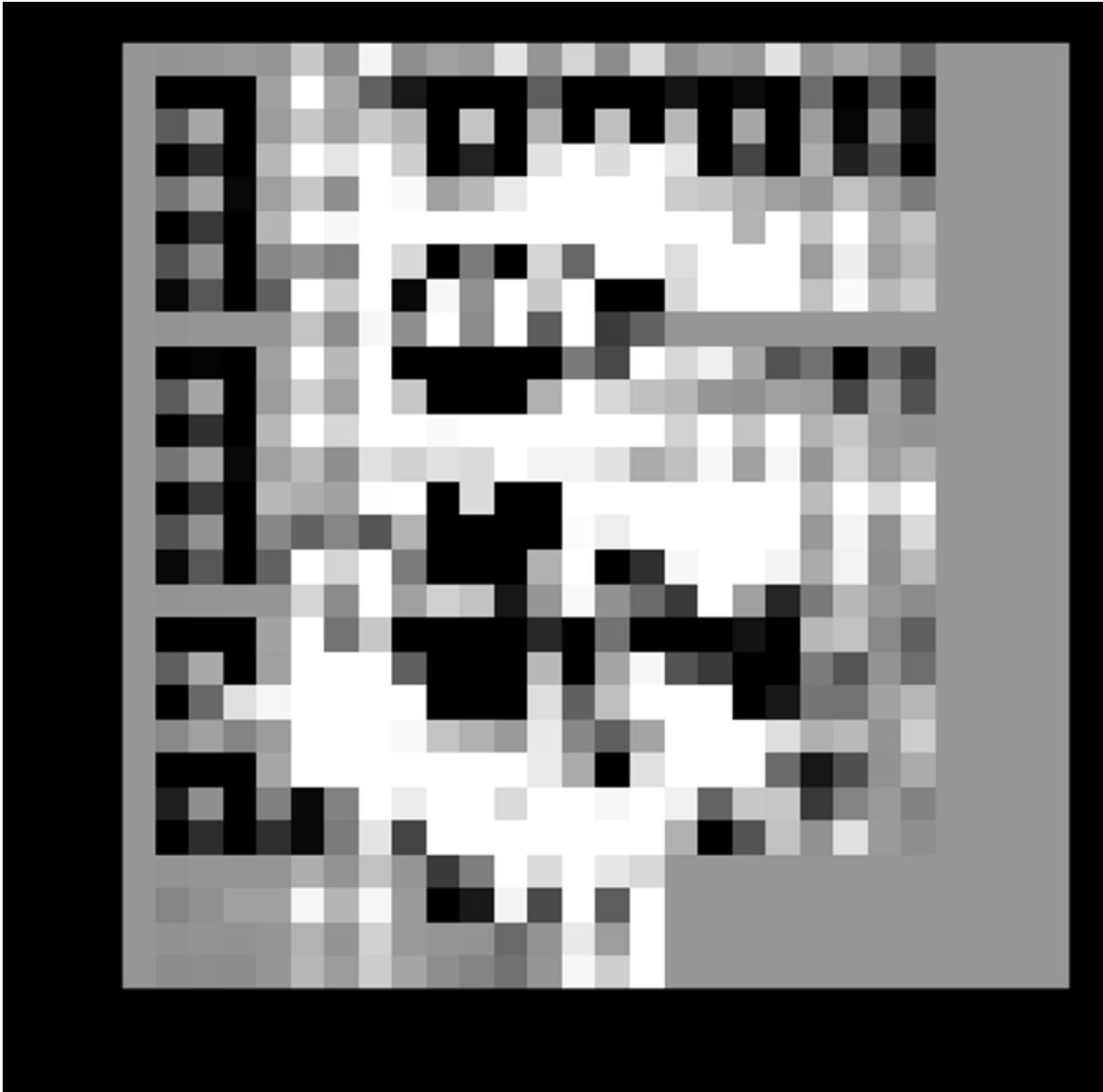


Note how this is different from the visualization we previously had with the linear classifier in red and blue in the first section. Those visuals represented the importance in *aggregate* for the entire input space. The visualization here is only for this specific input. For a different input, e.g. a different 3, the local gradient would look different, as shown below:

This 3:



yields this gradient:



Another group of approaches focus on visualizing with respect to the model parameters themselves, rather than the input. A lot of the work has been done by Chris Olah, Shan Carter, Ludwig Schubert, and others on [distill.pub](#). Their work in this area has gone from [visualizing the activations of specific neurons and layers](#), to entire [maps of activations for many networks](#), to [decomposing models into interpretable building blocks](#). Another great visualization resource for this type of work is the [OpenAI Microscope](#). Progress here has been very exciting, but it remains to be seen if similar approaches can be found for neural nets which focus on tasks other than image recognition.

**Can the model show what else in the training data it thinks are related to this input/output? (Explanation by Example)**

This asks for what other training examples are similar to the current input. When the similarity metric is just distance in the original feature space, this is akin to KNN with  $K = 1$ . More sophisticated methods may look for examples which are similar in whatever representation or latent space the model is using. The human justification for this type of approach is that it is similar to reasoning by analogy, where we present a related scenario to support our actions.

While I think this is useful, it definitely doesn't seem like all we need for understanding, or even most of what we'd need.

## What Else Might Be Important?

These are a mix of other questions I thought of before/after reading the above papers. Some of them are also from Lipton's paper, but from the earlier sections on interpretability desiderata. Because [answering questions is harder than asking them](#), I've also taken the time to give some partial responses to these questions, but these are not well-researched and should be taken as my own thoughts only.

1. What are the relevant features for the model? What is superfluous?
  - We've seen that linear models can easily identify relevant features. Regularization and other approaches to learn sparse models or encodings can also help with this. One interesting direction (that may already be explored) is to evaluate the model on augmented training data that has structured noise or features that correlate with real features and see what happens.
2. How can you describe what the model does in simpler terms?
  - The direct way to approach this question is to focus on approximating the model's performance using fewer parameters. A more interesting approach is to try and summarize what the model does in plain English or some other language. Having a simplified description could help with understanding, at least for our intuition.
3. What can the model tell you to allow you to approximate its performance in another setting or another model?
  - Another way to think about models which are interpretable is that they are doing some sort of modeling of the world. If you asked a person, for example, why they made some decision, they might tell you relevant facts about the world which could help you come to the same decision. Maybe some sort of teacher-learner RL type scenario where we can formalize knowledge transfer? But ultimately it seems important for the insights to be useful for humans; the feedback loop seems too long to make it an objective to optimize for, but maybe there's a clever way to approximate it...There might be a way where we instead train a model to output some representation or distribution that, when added to some other interpretable model (which could be a human's reasoning), leads to improved performance.
4. How informative is this model, relative to another more interpretable model?
  - Currently, deep learning outperforms other more interpretable models on a wide variety of tasks. Past just looking at loss, perhaps there is some way we can formalize how much more information the black box model is using. In the case of learned features versus hand-picked features, it could be useful to understand from an information theory perspective how much more informative the learned features are. Presumably interpretable features would tend to be more correlated with one another.

5. What guarantees does the model have to shifts in distribution?
  - Regularization, data augmentation, and directly training with perturbed examples all help with this issue. But perhaps there are other algorithmic guarantees we could derive for our models.
6. What trips up the model (and also the human)?
  - One interesting sign that our model is reasoning in interpretable ways is to see if examples which trip up humans also trip up the model. There was some work a little while back on adversarial examples which found that certain examples which fooled the network also fooled humans. Lack of divergence on these troubling examples could be a positive sign.
7. What trips up the model (but not the human)?
  - Conversely, we might get better insight into our model by honing in on "easy" examples (for a human) that prove to be difficult for our model. This would likely be indicative of the model using features that we are not, and thus it's learned a different manifold (or whatever) through the input space.
8. What does the model know about the data-generation process?
  - In most cases, this is encoded by our prior, which is then reflected in the class of models we do empirical risk minimization over. Apart from that, it does seem like there are relevant facts about the world which could be helpful to encode. A lot of the symbolic AI approaches to this seem to have failed, and it's unclear to me what a hybrid process would look like. Perhaps some of the human-assisted RL stuff could provide a solution for how to weigh between human advice and learned patterns.
9. Does the model express uncertainty where it should?
  - In cases where the input is something completely nonsensical, it seems perhaps desirable for the model to throw its hands up in the air and say "I don't know", rather than trying its best to give an answer. Humans do this, where we might object to a question on grounds of a type error. For a model, this might require understanding the space of possible inputs.
10. What relationships does the model use?
  - The model could be using direct correlations found in the data. Or it could be modeling some sort of causal graph. Or it could be using latent factors to build an approximate version of what's going on. Understanding what relationships in the data are lending themselves to helping the model and what relationships are stored could be useful.
11. Are the model's results contestable?
  - We touched on this at the very beginning of the post, but there are not many modern approaches which seem to have done this. The most contestable model might look something like an automated theorem prover which uses statements about the world to build an argument. Then we would simply check each line. Past that, one nice-to-have which could facilitate this is to use machine learning systems which build explicit models about the world. In any case, this pushes our models to make their assumptions about the world more explicit.

## What's Next?

Broadly, I think there are two main directions that interpretability research should go, outside of the obvious direction of "find better ways to formalize what we mean by interpretability". These two areas are evaluation and utility.

## Evaluation

The first area is to find better ways of evaluating these numerous interpretability methods. For many of these visualization-based approaches, a default method seems to be sanity-checking with our own eyes, making sure that interpretable features are being highlighted. Indeed, that's what we did for the MNIST examples above. However, [Sanity Checks for Saliency Maps](#), a recent paper, makes a strong case for why this is definitely not enough.

As mentioned earlier, saliency maps represent a broad class of approaches that try to understand what parts of the input are important for the model's output, often through some sort of gradient. The outputs of several of these methods are shown below. Upon visual inspection, they might seem reasonable as they all seem to focus on the relevant parts of the image.

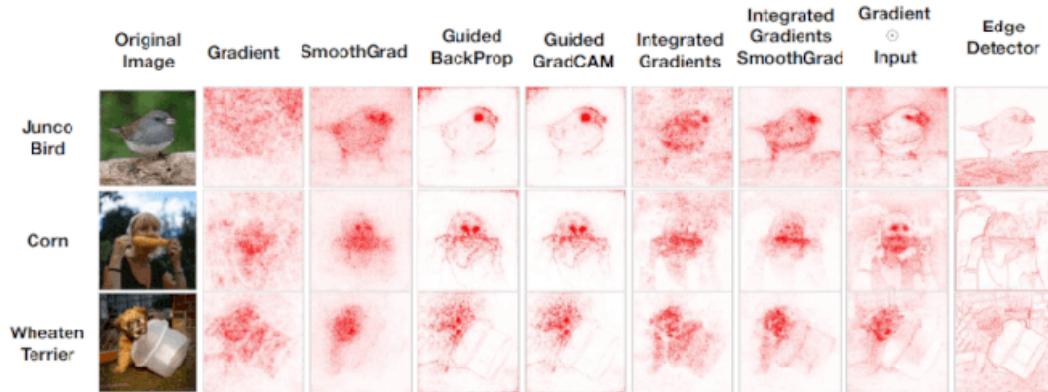


Figure 1: **Saliency maps for some common methods compared to an edge detector.** Saliency masks for 3 inputs for an Inception v3 model trained on ImageNet. We see that an edge detector produces outputs that are strikingly similar to the outputs of some saliency methods. In fact, edge detectors can also produce masks that highlight features which coincide with what appears to be relevant to a model's class prediction. We find that the methods most similar (see Appendix for SSIM metric) to an edge detector, i.e., Guided Backprop and its variants, show minimal sensitivity to our randomization tests.

However, the very last column is the output, not for a saliency map, but for an edge detector applied to the input. This makes it not a function of the model, but merely the input. Yet, it is able to output "saliency maps" which are visually comparable to these other results. This might cause us to wonder if the other approaches are really telling us something about the model. The authors propose several tests to investigate.

The first test compares the saliency map of a trained model with a model that has randomly initialized weights. Here, clearly if the saliency maps look similar, then it really is more dependent on the input and not the model's parameters.

The second test compares the saliency map of a trained model with a trained model that was given randomly permuted labels. Here, once again, if the saliency maps look similar, this is also a sign of input dependence because the same "salient" features have been used to justify two different labels.

Overall, the authors find that the basic gradient map shows desired sensitivity to the above tests, whereas certain other approaches like Guided BackProp do not.

I haven't looked too deep into each one of the saliency map approaches, but I think the evaluation methods here are very reasonable and yet somehow seem to be missed in previous (and later?) papers. For example, the paper on [Grad-CAM](#) goes in-depth over the ways in which their saliency map can help aid in providing explanations or identifying bias for the dataset. But they do not consider the sensitivity of their approach to model parameters.

In the above paper on sanity-checks, they find that Grad-CAM actually is sensitive to changes in the input, which is good, but I definitely would like to see these sanity-checks being applied more frequently. Outside of new approaches, I think additional benchmarks for interpretability that mimic real-world use cases could be of great value to the field.

Another approach is this direction is to back-chain from the explanations that people tend to use in everyday life to derive better benchmarks. [Explanation in Artificial Intelligence: Insights from the Social Sciences](#) provides an overview of where philosophy and social science can meet ML in the middle. Of course, the final arbiter for all this is how well people can actually use and interpret these interpretability results, which brings me to my second point.

## Utility

The second area is to ensure that these interpretability approaches are actually providing value. Even if we find ways of explaining models that are actually sensitive to the learned parameters (and everything else), I think it still remains to be seen if these explanations are actually useful in practice. At least for current techniques, I think the answer is uncertain and possibly even negative.

[Manipulating and Measuring Model Interpretability](#), a large pre-registered from Microsoft Research, found that models which had additional information like model weights were often not useful in helping users decide how to make more accurate judgments on their own or notice when the model was wrong. (Users were given either a black-box model or a more interpretable one.)

They found that:

"[o]n typical examples, we saw no significant difference between a transparent model with few features and a black-box model with many features in terms of how closely participants followed the model's predictions. We also saw that people would have been better off simply following the models rather than adjusting their predictions. Even more surprisingly, we found that transparent models had the unwanted effect of impairing people's ability to correct inaccurate predictions, seemingly due to people being overwhelmed by the additional information that the transparent model presented"

Another paper, [Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for MachineLearning](#), found that even data scientists may not understand what interpretable visualizations tell them, and this can inspire unwarranted confidence in the underlying model, even leading to ad-hoc rationalization of suspicious results.

Lastly, [Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?](#), is a recent study of five interpretability techniques and how they empirically help humans. The authors found very few benefits from any of techniques. Of particular note is that explanations which were rated to be higher quality by participants were not very useful in actually improving human performance.

All of this points to the difficult road ahead for interpretability research. These approaches and visuals are liable to be misused and misinterpreted. Even once we get improved notions of interpretability with intuitive properties, it still remains to be seen if we can use them to achieve the benefits I listed out in the very beginning. While it certainly seems more difficult to formalize interpretability than to use it well, I'm glad that empirical tests are already being done; they can hopefully also guide where the research goes next.

Finally, lurking behind all this is the question of decreased performance and adoption. It's obvious these days that black box models dominate in terms of results for many areas. Any additional work to induce a more interpretable model, or to derive a post-hoc explanation brings an additional cost. At this point in time, all the approaches towards improving interpretability we've seen either increase training / processing time, reduce accuracy, or do some combination of both. For those especially worried about competition, arms races, and multipolar traps, the case to adopt these approaches (past whatever token compliance will satisfy the technology ethics boards of the future) seems weak.

# Highlights from the Blackmail Debate (Robin Hanson vs Zvi Mowshowitz)

At one of our weekly LessWrong events, [we had a lively debate on legalizing blackmail](#) ([video](#), [transcript](#)). Robin Hanson took the pro side, Zvi Mowshowitz took the con, and I moderated. 70 people showed up to watch for ~2 hours.

Here's my overview of their positions.

- Zvi thinks that blackmail would incentivize a whole host of terrible actions, such as trying to trick people into norm violation, and people becoming intensely secretive even around their closest friends and family.
- Robin thinks that blackmail is a weird rule, where you cannot ask for money to keep a secret, but the other person is allowed to offer it (e.g. people can offer you money if you sign an NDAs). This makes no sense and Robin is looking for any clear reason why making one side of this deal should be illegal.

Below are some quotes from their conversation. And of course, there's the [full edited transcript](#) and [video](#) for those who want all the details.

## Highlights

### What's good about blackmail

**Robin Hanson:** I think in the case of say David Letterman, who famously was blackmailed for having affairs, if he could have actually been successfully blackmailed, then people like Letterman would be doing much less of what he was doing. And these weren't just affairs with random people who liked him, these were employees of him and so they are much more morally questionable. And so I think there would just be a lot less sexual harassment if blackmail was legal.

**Robin Hanson:** There are a lot of powerful people who break a lot of rules, actually legal rules in many ways. And then the people around them shut up about it, let them get away with it because they don't feel they actually have a credible threat to report it. And so they don't. And so blackmail would mean a lot more actual reporting or a discouragement of the things powerful people do, that break rules and norms.

### Spreading dirt is often prosocial

**Zvi Mowshowitz:** Essentially your argument is that in today's world, if people online were to find out something about you and decide to cause a lot of trouble in your life based on this thing, that it must've been a bad thing that the public absolutely needed to know.

**Robin Hanson:** On average, letting people know about other people's dirt is a good thing. The incentives to not have dirt, the incentive to expose dirt, are on average a good thing. Yes, they go wrong in many particular ways. But on

average, they're good. That's my fundamental claim about why gossip is generally good, even when people are trying to find dirt in order to make someone look bad, and blackmail is just upping the incentives on that somewhat.

## Banning isn't the only option

**Zvi Mowshowitz:** So what I'm saying, blackmail specifically selects for the scenarios in which there is great harm.

**Robin Hanson:** So does trying to make somebody look bad. That's the thing we're talking about. There are situations now where people are trying to make other people look bad. Either you want to ban those, or you have to accept that on average those are good, even though they contain the problems reporting.

**Zvi Mowshowitz:** No. Okay, so I think there's an important fallacy there, that I can not want to ban something but still think it's in general pretty bad.

## Bad effects of blackmail

**Zvi Mowshowitz:** I think, in a sense, blackmail makes these things much more negative. In particular, the incentive to entrap, the incentive to create negative material, to induce norm violations, is much much stronger under blackmail. And the fear of such things happening, in general, and the cost of navigating blackmail situations, and the fear of having to deal with these things is very bad. I would feel very stressed living under a blackmail legal regime.

**Zvi Mowshowitz:** The idea is that if we had less of them the world would be a better place, but there is no law that I can pass that banning it would not have other effects. But I can try to tax it, right? I can make it more costly. I can make it more inconvenient in ways that discourage the behavior, and maybe that's even better than banning it, because now, when it was really worth doing, it happens anyway. And by tax, I don't mean literally, "You must pay 5%." I want to make this more annoying for you.

## Blackmail leads to doxing?

**Robin Hanson:** Remember one of our other options is just to ban some kinds of gossip. So I said, we do properly ban telling people's passwords or sharing their naked pictures. If you think there's a kind of gossip that's just harmful, you just ban that kind of gossip. We're talking about allowing that gossip, except when one side makes an offer to pay the other, but not vice versa. That's the puzzle we're talking about, not just the generic, that some things you might not want to let people gossip about.

**Zvi Mowshowitz:** One recent example of potential blackmail is what if someone were to blackmail Scott Alexander and threatened to reveal his true name?

**Robin Hanson:** But again, you can have privacy rules about that. If you just want to say, you're not allowed to reveal people's anonymous names, just make that the rule. You're talking about, it would be okay if it was gossip but not if it wasn't,

but that's not true in this case. You think it would be bad even if it was revealed without monetary incentive.

## Blackmail is a weird rule

**Zvi Mowshowitz:** My claim is not that there are no things we want to outright ban you from sharing. I'm saying there'll be a lot of things that we cannot enumerate and ban you from sharing that we nevertheless want to tax the sharing of.

**Robin Hanson:** The key thing isn't whether certain information should be revealed or not, it's whether you should add this extra complexity. So again, we always have the option to ban gossip or to require it, but what we have is this weird rule where you're allowed to do it in trade for other things, but not for money, but you are allowed to do it for money if one side makes the offer, but not the other side makes the offer. This is the thing I find very hard to justify.

**Robin Hanson:** I can understand why something should be private and something should be published, and something should be allowed to be said and some things not. But why this weird combination of not the money unless one side makes the offer but not the other. You haven't addressed this one side versus the other thing at all in your entire conversation here. You haven't addressed the possibility of an NDA, doing all of these examples you don't like.

## Hanson's strongest claim

**Robin Hanson:** My claim was not so much that blackmail is good but that no one had offered concrete, clear, consequentialist arguments for why blackmail should be banned, especially relative to allowing NDAs in terms of who makes the offer. That's my strongest claim. And my claim, especially, is about coming up with explicit reasons and arguments. So it's about the fact that in society, we just have a lot of policies that, if you look at them, you're not sure what their justification is. If you ask people, they give you various justifications that are contradictory. I think it's a sad situation that we don't have clear justification for most of our policies.

## Back-and-Forts

### How much does blackmail punish you?

**Robin Hanson:** Law and blackmail are two different channels. I'm endorsing both channels. We have a formal legal system where we have things formally crimes, and they're formally punished by legislatures deciding the sentence, resources of the police, and fines. But also, we have a system of blackmail, wherein people are paid financially and other ways for their doing things that the audience will disapprove of. Those are two different ways that the larger world disapproves and discourages things.

**Ben Pace:** The level at which society wants to hurt you, or get your private information, is not really often in proportion to how much they think it is just to hurt

you, or I think it is just to hurt you. Whereas a lot of tabloid press that will just hurt you because it gets them attention, and they can join in gossip in a big conversation in a way certainly I don't endorse, and I think most people don't endorse.

**Robin Hanson:** Well, then, why not make that illegal then, Ben?

**Ben Pace:** I think illegal is strong and kind of silencing thing.

**Robin Hanson:** Well, that's true for blackmail as well.

**Zvi Mowshowitz:** No, I don't think it is. Why does blackmail have a chilling effect?

**Robin Hanson:** The mud-raking journalists have a chilling effect. Of course. They all have a chilling effect. The question is whether it's too much or not enough.

**Ben Pace:** But Robin, is your take that the tabloids shouldn't do that to people, and it should be illegal, or is your take that, no, we should have this and encourage it more with more money.

**Robin Hanson:** I'd say that, on average, the tabloids exposing things about people is, on average, a good thing. It goes wrong in many particular ways, but I do not want to ban the tabloids from writing exposes.

**Ben Pace:** I don't want to ban them from writing exposes, but I currently think the situation is kind of like blackmailing in which they will extract way more resources than is proportional, and on that do massive amounts of damage.

**Robin Hanson:** I don't see that.

## What would legalized blackmail actually look like?

**Zvi Mowshowitz:** I don't even think that allowing blackmail would increase the number of such things that were in fact revealed. I think it would decrease it.

**Zvi Mowshowitz:** People would be more secretive, and in fact, sometimes when they were blackmailed, they would in fact pay. Other times they would find credible threats of retaliation and that would prevent the information from coming out. And I think that in particular, right now, when normally determining whether or not to share a gossip, they tend to share net useful gossip more than they tend to share net harmful gossip, and that blackmail reverses this incentive and also causes people to look for net harmful gossip, rather than look for net helpful gossip. Most of the time when people are looking for information, they're looking for information the public needs to know.

**Zvi Mowshowitz:** I believe that most of the time, most people are not, when they seek information, primarily looking to harm someone. They're primarily looking to benefit themselves, or benefit their friends or their allies in some way. Hurting someone else is mostly a side effect.

**Robin Hanson:** That's also true with blackmail. The main effect is the money, not the hurt. Their main motivation is to get the compensation.

**Zvi Mowshowitz:** The main benefit is to gain the power over the person that you extract something of value, whether it's money or something else.

**Robin Hanson:** But that's not the same as hurting.

**Zvi Mowshowitz:** But the way that you do that is you gain the ability to hurt someone.

**Robin Hanson:** I disagree with that whole framing. Verizon, which is my person who supplies my internet and my phone and my TV, they want me to really want their product. So the more that they can make me really desperate for their product, then the more I'm willing to pay for it. Of course they do it, hopefully, by making their product attractive. But that is a way of gaining power over me. In general, all through society, when people are making deals with the other, in anticipation of those deals, they want to be in demand. They want to want the other party to want them. That's basically the same thing. It's all about, before a deal, wanting to have the other party want to make the deal. So that happens in marriages, it happens in jobs, it happens to me with Verizon. Is that harm? Is the effort that Verizon goes through to make sure that I don't want to be without their service, is that harming me because now it makes me more willing to pay for their service?

**Zvi Mowshowitz:** But doesn't Verizon do that by offering a benefit? Verizon creates a service that makes your life better so that you will be willing to buy it. So if Verizon were to, say, cut the wires of their competition so that their competition couldn't come to your house, and then threatened to cut off your service unless you paid them 10 times as much, that seems-

**Robin Hanson:** That's what I said in my initial remarks about the reference point. You have in mind, the reference point is, I say nothing. And so I'm harming you by threatening to say something. But what if the reference was, I was going to say it anyway and you pay me not to say it, well now with respect to that reference point, I'm helping you by letting you pay me not to say it. So it all comes down to what's the reference behavior you thought would have happened instead.

## The cost of having norms

**Ben Pace:** There's a question of scale. I am okay with you telling a bunch of people that I did something bad. I'm not okay if you managed to get it on the front page of the New York Times.

**Robin Hanson:** It depends on who you are. If you're an ordinary person, it won't get on the front page of the Times.

**Ben Pace:** If I'm a rich person who is not very important in a lot of other ways, if I just have a lot of resources to be taken, even though it is not important about how I use those, then I think the blackmail, now, makes it much more likely that that information about me will get to a level of prominence and life-destroying damage that it would not previously, just because I have a lot of resources you can steal.

**Robin Hanson:** Well, why is it bad if New York Times readers find out about it, but not if other people find out about it? Why is that something that makes it bad?

**Ben Pace:** Because I think there's a level of punishment that information, gossiping, should do to you, and in general, people sharing it when it seems useful feels like it

will hit the balance where it will get shared as much as it's useful. But people sharing it for as much resources they can take, I think, will encourage much over-punishment.

**Ben Pace:** Almost no norm violation of mine should be on the front page of New York Times, and if I have enough resources, then it will get there, if you allow blackmail.

**Robin Hanson:** I think you want your public stance to be that you do follow the norms.

**Ben Pace:** No, my public stance is that I do sometimes break norms, and I still should not have my life destroyed by that.

**Zvi Mowshowitz:** Regarding the New York Times, it's interesting that when I worked for a certain corporation which I will not name, we had a principle that we could not put in any written form any statement that we would not want on the front page of the New York Times. And so, the very fact that someone might threaten to cause harm to us, or decide to cause harm to us by sharing this, meant that we had to be much more implicit, keep less records, destroy evidence, be much less rational-

**Robin Hanson:** That's the general cost of norms. I mean, the norm system has cost, okay. It's unique to humans. Other animals didn't have it, and it's part of the power of humanity that we've had and enforced norms, but norm systems definitely have costs. One of them is we sometimes have wrong norms. Sometimes we mis-enforce norms, in that we draw the wrong conclusions about who violated which norms, and we may well punish too much or too little in other situations. But still, on average, norms are good.

## How did the audience's minds change?



Well done to Robin for halving Zvi's support! Better luck next time Zvi.

I myself moved from "blackmail should be illegal" to "I am confused", and would be interested if people could write things to help resolve this debate further.

For the past few months we've had weekly LessWrong events on Sundays, and will continue to do so. We announce them on the frontpage by Thursday each week, check there for announcements of more talks, debates, and double cruxes.

Here is the full [2-hour video \(with Q&A\)](#), and here is [the full edited transcript](#).

# Forecasting AI Progress: A Research Agenda

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

AI forecasting is an important research area but lacking in a general direction. To address this issue we present a research agenda for AI forecasting that has been generated using the Delphi technique to elicit opinion from 15 leading researchers on the topic (the majority of whom are members of this community). The research agenda can be found on arXiv through this link:

**[Forecasting AI Progress: A Research Agenda \(link to arXiv\)](#)**

The agenda was framed so that it can be useful to both members of this community as well as the technological forecasting community more broadly. To these ends we plan to submit the arXiv manuscript to Technological Forecasting and Social Change, however, we will wait for roughly a month to receive comments. Please feel free to give us your thoughts here.