Best of LessWrong: June 2019

1. Being the (Pareto) Best in the World
2. Welcome to LessWrong!
3. Steelmanning Divination
4. Mistakes with Conservation of Expected Evidence
5. Book Review: The Secret Of Our Success
6. The Schelling Choice is "Rabbit", not "Stag"
7. Reason isn't magic
8. Writing children's picture books
9. Arbital scrape
10. Selection vs Control
11. Reneging prosocially by Duncan Sabien
12. The Hacker Learns to Trust
13. Aligning a toy model of optimization
14. Circle Games
15. Reasonable Explanations
16. Defending points you don't care about
17. Book Review: Why Are The Prices So Damn High?
18. Major Update on Cost Disease
19. Tal Yarkoni: No, it's not The Incentives—it's you
20. Research Agenda in reverse: what *would* a solution look like?
21. On alien science
22. Deceptive Alignment
23. The Inner Alignment Problem
24. What's up with self-esteem?
25. Causal Reality vs Social Reality
26. Honors Fuel Achievement
27. In physical eschatology, is Aestivation a sound strategy?
28. Whence decision exhaustion?
29. What's the best explanation of intellectual generativity?
30. Epistemic Spot Check: The Role of Deliberate Practice in the Acquisition of Expert Performance
31. AGI will drastically increase economies of scale
32. Risks from Learned Optimization: Conclusion and Related Work
33. Can we use ideas from ecosystem management to cultivate a healthy rationality memespace?
34. How can we measure creativity?
35. ISO: Automated P-Hacking Detection
36. Conditions for Mesa-Optimization
37. LessWrong FAQ
38. Research Agenda v0.9: Synthesising a human's preferences into a utility function
39. What is the evidence for productivity benefits of weightlifting?
40. "The Bitter Lesson", an article about compute vs human knowledge in AI
41. Asymmetric Weapons Aren't Always on Your Side
42. "But It Doesn't Matter"
43. [AN #58] Mesa optimization: what it is, and why we should care
44. Instead of "I'm anxious," try "I feel threatened"
45. Machine Learning Projects on IDA
46. Only optimize to 95 %
47. For the past, in some ways only, we are moral degenerates
48. Map of (old) MIRI's Research Agendas
49. Is your uncertainty resolvable?
50. Does the _timing_ of practice, relative to sleep, make a difference for skill consolidation?

Best of LessWrong: June 2019

1. Being the (Pareto) Best in the World
2. Welcome to LessWrong!
3. Steelmanning Divination
4. Mistakes with Conservation of Expected Evidence
5. Book Review: The Secret Of Our Success
6. The Schelling Choice is "Rabbit", not "Stag"
7. Reason isn't magic
8. Writing children's picture books
9. Arbital scrape
10. Selection vs Control
11. Reneging prosocially by Duncan Sabien
12. The Hacker Learns to Trust
13. Aligning a toy model of optimization
14. Circle Games
15. Reasonable Explanations
16. Defending points you don't care about
17. Book Review: Why Are The Prices So Damn High?
18. Major Update on Cost Disease
19. Tal Yarkoni: No, it's not The Incentives—it's you
20. Research Agenda in reverse: what *would* a solution look like?
21. On alien science
22. Deceptive Alignment
23. The Inner Alignment Problem
24. What's up with self-esteem?
25. Causal Reality vs Social Reality
26. Honors Fuel Achievement
27. In physical eschatology, is Aestivation a sound strategy?
28. Whence decision exhaustion?
29. What's the best explanation of intellectual generativity?
30. Epistemic Spot Check: The Role of Deliberate Practice in the Acquisition of Expert Performance
31. AGI will drastically increase economies of scale
32. Risks from Learned Optimization: Conclusion and Related Work
33. Can we use ideas from ecosystem management to cultivate a healthy rationality memespace?
34. How can we measure creativity?
35. ISO: Automated P-Hacking Detection
36. Conditions for Mesa-Optimization
37. LessWrong FAQ
38. Research Agenda v0.9: Synthesising a human's preferences into a utility function
39. What is the evidence for productivity benefits of weightlifting?
40. "The Bitter Lesson", an article about compute vs human knowledge in AI
41. Asymmetric Weapons Aren't Always on Your Side
42. "But It Doesn't Matter"
43. [AN #58] Mesa optimization: what it is, and why we should care
44. Instead of "I'm anxious," try "I feel threatened"
45. Machine Learning Projects on IDA
46. Only optimize to 95 %
47. For the past, in some ways only, we are moral degenerates

48. Map of (old) MIRI's Research Agendas
49. Is your uncertainty resolvable?
50. Does the _timing_ of practice, relative to sleep, make a difference for skill consolidation?

Being the (Pareto) Best in the World

The generalized efficient markets (GEM) principle says, roughly, that things which would give you a big windfall of money and/or status, will not be easy. If such an opportunity were available, someone else would have already taken it. You will never find a $100 bill on the floor of Grand Central Station at rush hour, because someone would have picked it up already.

One way to circumvent GEM is to be the best in the world at some relevant skill. A superhuman with hawk-like eyesight and the speed of the Flash might very well be able to snag $100 bills off the floor of Grand Central. More realistically, even though financial markets are the ur-example of efficiency, a handful of firms do make impressive amounts of money by being faster than anyone else in their market. I'm unlikely to ever find a proof of the Riemann Hypothesis, but Terry Tao might. Etc.

But being the best in the world, in a sense sufficient to circumvent GEM, is not as hard as it might seem at first glance (though that doesn't exactly make it easy). The trick is to exploit dimensionality.

Consider: becoming one of the world's top experts in proteomics is hard. Becoming one of the world's top experts in macroeconomic modelling is hard. But how hard is it to become sufficiently expert in proteomics and macroeconomic modelling that nobody is better than you at both simultaneously? In other words, how hard is it to reach the Pareto frontier?

Having reached that Pareto frontier, you will have circumvented the GEM: you will be the single best-qualified person in the world for (some) problems which apply macroeconomic modelling to proteomic data. You will have a realistic shot at a big money/status windfall, with relatively little effort.

(Obviously we're oversimplifying a lot by putting things like "macroeconomic modelling skill" on a single axis, and breaking it out onto multiple axes would strengthen the main point of this post. On the other hand, it would complicate the explanation; I'm keeping it simple for now.)

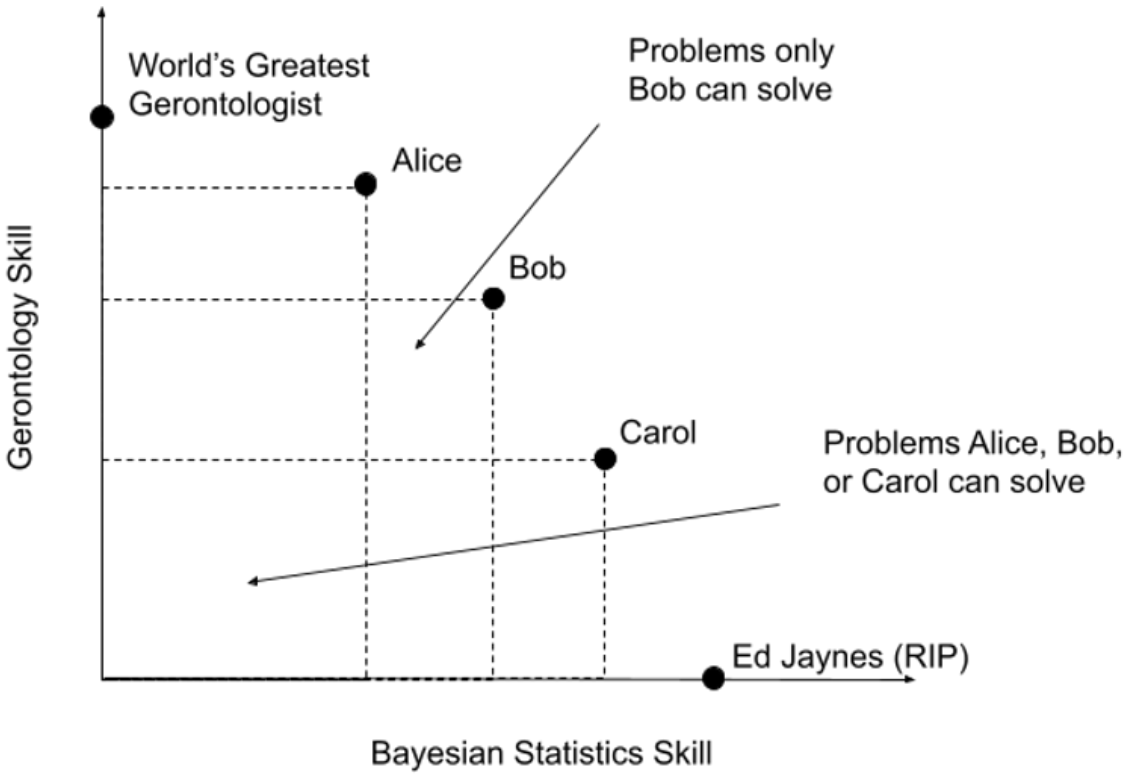
Let's dig into a few details of this approach…

Elbow Room

There are many table tennis players, but only one best player in the world. This is a side effect of ranking people on one dimension: there's only going to be one point furthest to the right (absent a tie).

Pareto optimality pushes us into more dimensions. There's only one best table tennis player, and only one best 100-meter sprinter, but there can be an unlimited number of Pareto-optimal table tennis/sprinters.

Problem is, for GEM purposes, elbow room matters. Maybe I'm the on the pareto frontier of Bayesian statistics and gerontology, but if there's one person just little bit better at statistics and worse at gerontology than me, and another person just a little bit better at gerontology and worse at statistics, then GEM only gives me the advantage over a tiny little chunk of the skill-space.



This brings up another aspect...

Problem Density

Claiming a spot on a Pareto frontier gives you some chunk of the skill-space to call your own. But that's only useful to the extent that your territory contains useful problems.

Two pieces factor in here. First, how large a territory can you claim? This is about elbow room, as in the diagram above. Second, what's the density of useful problems within this region of skill-space? The table tennis/sprinting space doesn't have a whole lot going on. Statistics and gerontology sounds more promising. Cryptography and monetary economics is probably a particularly rich Pareto frontier these days. (And of course, we don't need to stop at two dimensions - but we're going to stop there in this post in order to keep things simple.)

Dimensionality

One problem with this whole GEM-vs-Pareto concept: if chasing a Pareto frontier makes it easier to circumvent GEM and gain a big windfall, then why doesn't everyone chase a Pareto frontier? Apply GEM to the entire system: why haven't people already picked up the opportunities lying on all these Pareto frontiers?

Answer: dimensionality. If there's 100 different specialties, then there's only 100 people who are the best within their specialty. But there's 10k pairs of specialties (e.g. statistics/gerontology), 1M triples (e.g. statistics/gerontology/macroeconomics), and something like 10^30 combinations of specialties. And each of those pareto frontiers has room for more than one person, even allowing for elbow room. Even if only a small fraction of those combinations are useful, there's still a *lot* of space to stake out a territory.

And to a large extent, people do pursue those frontiers. It’s no secret that an academic can easily find fertile fields by working with someone in a different department. “Interdisciplinary” work has a reputation for being unusually high-yield. Similarly, carrying scientific work from lab to market has a reputation for high yields. Thanks to the “curse” of dimensionality, these goldmines are not in any danger of exhausting.

Welcome to LessWrong!

The road to wisdom? Well, it's plain
and simple to express:

Err
and err
and err again
*but **less***
*and **less***
*and **less**.*

– Piet Hein

LessWrong is an online forum and community dedicated to improving human reasoning and decision-making. We seek to hold true beliefs and to be effective at accomplishing our goals. Each day, we aim to be less wrong about the world than the day before.

Training Rationality

Rationality has a number of definitions[1] on LessWrong, but perhaps the most canonical is that the more rational you are, the more likely your reasoning leads you to have accurate beliefs, and by extension, allows you to make decisions that most effectively advance your goals.

LessWrong contains a lot of content on this topic. How minds work (both human, artificial, and theoretical ideal), how to reason better, and how to have discussions that are productive. We're very big fans of Bayes Theorem and other theories of normatively correct reasoning[2].

To get started improving your Rationality, we recommend reading the background-knowledge text of LessWrong, Rationality: A-Z (aka "The Sequences") or at least selected highlights from it. After that, looking through the Rationality section of the Concepts Portal is a good thing to do.

Applying Rationality

You might value Rationality for its own sake, however, many people want to be better reasoners so they can have more accurate beliefs about topics they care about, and make better decisions.

Using LessWrong-style reasoning, contributors to LessWrong have written essays on an immense variety of topics on LessWrong, each time approaching the topic with a desire to know what's actually true (not just what's convenient or pleasant to believe), being deliberate about processing the evidence, and avoiding common pitfalls of human reason.

Check out the Concepts Portal to find essays on topics such as artificial intelligence, history, philosophy of science, language, psychology, biology, morality, culture, self-care, economics, game theory, productivity, art, nutrition, relationships and hundreds of other topics broad and narrow.

LessWrong and Artificial Intelligence

For several reasons, LessWrong is a website and community with a strong interest in AI and specifically causing powerful AI systems to be safe and beneficial.

- AI is a field with how minds and intelligence works, overlapping a lot with rationality.
- Historically, LessWrong was seeded by the writings of Eliezer Yudkowsky, an artificial intelligence researcher.

- Many members of the LessWrong community are heavily motivated by trying to improve the world as much as possible, and these people were convinced many years ago that AI was a very big deal for the future of humanity. Since then LessWrong has hosted a lot of discussion of AI Alignment/AI Safety, and that's only accelerated recently with further AI capabilities developments.
    - LessWrong is also integrated with the Alignment Forum
    - The LessWrong team who maintain and develop the site are predominantly motivated by trying to cause powerful AI outcomes to be good.

If you want to see more or less AI content, you can adjust your Frontpage Tag Filters according to taste[3].

Getting Started on LessWrong

The core background text of LessWrong is the collection of essays, Rationality: A-Z (aka "The Sequences"). Reading these will help you understand the mindset and philosophy that defines the site. Those looking for a quick introduction can start with The Sequences Highlights

Other top writings include The Codex (writings by Scott Alexander) and Harry Potter & The Methods of Rationality. Also see the Library Page for many curated collections of posts and the Concepts Portal.

Also, feel free to introduce yourself in the monthly open and welcome thread!

Lastly, we do recommend that new contributors (posters or commenters) take time to familiarize themselves with the sites norms and culture to maximize the chances that your contributions are well-received.

Thanks for your interest!

- The LW Team

Related Pages

- **LessWrong FAQ**
- **A Brief History of LessWrong**
- **Team**
- **LessWrong Concepts**

1. ^

   Definitions of Rationality as used on LessWrong include:

   - Rationality is thinking in ways that systematically arrive at truth.

   - Rationality is thinking in ways that cause you to systematically achieve your goals.

   - Rationality is trying to do better on purpose.

   - Rationality is reasoning well even in the face of massive uncertainty.

   - Rationality is making good decisions even when it’s hard.

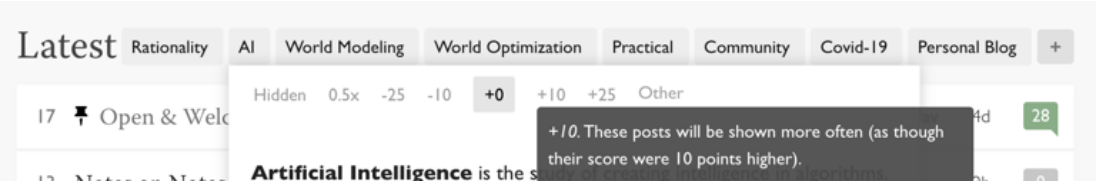
   -Rationality is being self-aware, understanding how your own mind works, and applying this knowledge to thinking better.

2. ^

There are in fact laws of thought no less ironclad than the law of physics [source].

3. ^

Hover your mouse over the tags to be able to adjust their weighting in your Latest Posts feed.



Steelmanning Divination

[This post was primarily written in 2015, after I gave a related talk, and other bits in 2018; I decided to finish writing it now because of a recent SSC post.]

The standard forms of divination that I've seen in contemporary Western culture--astrology, fortune cookies, lotteries, that sort of thing--seem pretty worthless to me. They're like trying to extract information from a random number generator, which is a generally hopeless phenomenon because of conservation of expected evidence. Thus I had mostly written off divination; although I've come across some arguments that divination served as a way to implement mixed strategies in competitive games. (Hunters would decide where to hunt by burning bones, which generated an approximately random map of their location, preventing their targets from learning where the humans liked to hunt and avoiding that location.[1]) But then I came across this striking passage, and sat up straight:

> One performs the rain sacrifice and it rains. Why? I say: there is no special reason why. It is the same as when one does not perform the rain sacrifice and it rains anyway. When the sun and moon suffer eclipse, one tries to save them. When Heaven sends drought, one performs the rain sacrifice. One performs divination and only then decides on important affairs. But this is not to be regarded as bringing one what one seeks, but rather is done to give things proper form. Thus, the gentleman regards this as proper form, but the common people regard it as connecting with spirits. If one regards it as proper form, one will have good fortune. If one regards it as connecting with spirits, one will have misfortune.

This is from Eric L. Hutton's translation of a collection of essays called Xunzi (presumably written by Xunzi, an ancient Chinese philosopher who was Confucian with heavy Legalist influences). The book was overall remarkable in how much of Xunzi's brilliance shone through, which is something I very rarely think about authors. (Talking to another rationalist who was more familiar with Chinese philosophy than I was, he also had this impression that Xunzi simply had a lot more mental horsepower than many other core figures.) By the end of it, I was asking myself, "if they had *this much* of rationality figured out back then, why didn't they conquer the world?" Then I looked into the history a bit more and figured out that two of Xunzi's students were core figures in Qin Shi Huang's unification of China to become the First Emperor.

So this paragraph stuck with me. When Xunzi talks about the way that earlier kings did things, I registered it as an applause light and moved on. When he talked about how an important role of government was to prevent innovation in music, I registered it as covering a very different thing than what I think of when I think about 'music' and moved on. But when he specifically called out the reason why I (and most educated people I know) don't pay much attention to astrology or other sorts of divination or magic, said "yeah, those would be dumb reasons to do this," and then said "*but there's still a reason*", I was curious. What's the proper form that he's talking about? (Sadly, this was left as an exercise for the reader; the surrounding paragraphs are only vaguely related.)

In his introduction, Hutton summarizes the relevant portion of Xunzi's philosophy:

> In this process of becoming good, ritual plays an especially important role in Xunzi's view. As he conceives them, the rituals constitute a set of standards for

> proper behavior that were created by the past sages and should govern virtually every aspect of a person's life. These rituals are not inviolable rules: Xunzi allows that people with developed moral judgment may need to depart from the strict dictates of ritual on some occasions, but he thinks those just beginning the process of moral learning need to submit completely to the requirements of ritual. Of the many important roles played by the rituals in making people good on Xunzi's view, three particularly deserve mention here. First the rituals serve to *display* certain attitudes and emotions. The ritually prescribed actions in the case of mourning, for instance, exhibit grief over the loss of a loved one, whether or not the ritual practitioner actually feels sadness. Second, even if the ritual practitioner does not actually feel the particular attitude or emotion embodied in the ritual, Xunzi believes that repeated performance of the ritual can, when done properly, serve to *cultivate* those attitudes and emotions in the person. To use a modern example, toddlers who do not know to be grateful when given a gift may be taught to say "thank you" and may do so without any understanding of its meaning or a feeling or gratitude. With repetition, time, and a more mature understanding of the meaning of the phrase, many of these children grow into adults who not only feel gratitude upon receiving gifts but also say "thank you" as a conscious expression of that feeling. Similarly, on Xunzi's view, rituals serve to inculcate attitudes and feelings, such as caring and respect, that are characteristic of virtue, and then serve to express a person's virtue once it is fully developed. A third important function of the rituals is to allot different responsibilities, privileges, and goods to different individuals, and thereby help to prevent conflict over those things among people.

So what is cultivated by performing divination?

The first step is figuring out what sort of divination we're discussing. Xunzi probably had in mind the I Ching, a book with 64 sections, each corresponding to a situation or perspective, and advice appropriate for that situation. In the simplest version, one generates six random bits and then consults the appropriate chapter. I actually tried this for about a month, and then have done it off and on since then. I noticed several things about it that seemed useful:

- Entries in the I Ching typically focused on perspectives or principles instead of situations, consequences, or actions. Today’s Taurus horoscope says “your self-esteem might be challenged by a fast talker or unpleasant situation” and counsels me “don’t accept things as they appear on first glance,” whereas the I Ching reading I just randomly selected talks about how following proper principles leads to increased power and how the increased power tempts us to abandon the principles that generated that power. This makes it much easier to scan one’s life and see where the perspectives shed new light on a situation, or where principles had been ignored. (One of the early successes of my I Ching practice was a chapter that suggested reaching out to a trusted guide for advice, and I realized I should talk to a mentor at work about a developing situation, which I wouldn’t have done otherwise.)
- Given that, daily divination almost filled the same role as daily retrospectives or planning sessions; I was frequently thinking about all the different parts of my life on a regular interval, using a variety of random access to filter things down.
- “One performs divination and only then decides on important affairs.” Often one is faced with a challenge that is “above one’s pay grade,” and having a prescribed ritual for what sort of cognition needs to be done encourages reflection and popping out of the obvious frame. Simply thinking about a

situation in the way one naturally would doesn’t correct for biases, while attempting to make sense of a situation from a randomly generated frame does help expand one’s conception of it.
- Since the divination result was a particular perspective (rather than an object-level claim about events or the correct action), it was easy to see when the perspective had been ‘fully considered’ and the reflection was done. If I’m trying to make a binary choice and I flip a coin to resolve it, basically all I’m doing is checking whether or not my gut is secretly hoping the coin lands a particular way, and in cases of genuine uncertainty I will end up just as uncertain after consulting the coin flip. But when I have a situation and I resolve the questions of “what does patience have to say about X?” and “what does humility have to say about X?” then I can have the sense of having actually made progress.
- It encouraged experimentation by partially decoupling one’s mood and one’s decisions. The first few days, I was instructed to consider diligence and so I worked more than I would have wanted to (and discovered that this was generally fine); on the fourth day, I joked, “when is it going to tell me to goof off and play video games?”, and then got a reading that said “effort is hopeless today; that happens, cope with it.” So despite feeling like I could have been productive, I took the day off, and later had the sense that I had confirmed my initial sense that I didn’t need to take the day off, providing data to calibrate on that I wouldn’t have gotten except through random variation.

Essentially, it looked to me like the steelman of divination is something like Oblique Strategies, where challenging situations (either ‘daily life’ or a specific important decision) are responded to by random access to a library of perspectives or approaches, and the particular claim made by a source is what distribution is most useful. There was previously an attempt on LW to learn what advice was useful, but I think on the wrong level of abstraction (the ‘do X’ variety, instead of the ‘think about X’ variety).

This approach has also served me well with other forms of divination I've since tried; a Tarot deck works by focusing your attention on a situation, and then randomly generating a frame (from an implied distribution of relevant symbols), giving one access to parts of the space that they wouldn't have considered otherwise. This also trains the habit of 'understanding alien frames'; if I am considering a conflict with another person and then have to figure out what it means that "I'm the vizier of water, the relationship is the three of earth, and the other person is strength" (where, of course, each of those is in fact an image rich in detail rather than a simple concept), this trains the habit of adopting other perspectives / figuring out how things make sense from the other point of view.

1. ^

   This view is apparently contested.

Mistakes with Conservation of Expected Evidence

Epistemic Status: I've really spent some time wrestling with this one. I am highly confident in most of what I say. However, this differs from section to section. I'll put more specific epistemic statuses at the end of each section.

Some of this post is generated from mistakes I've seen people make (or, heard people complain about) in applying conservation-of-expected-evidence or related ideas. Other parts of this post are based on mistakes I made myself. I think that I used a wrong version of conservation-of-expected-evidence for some time, and propagated some wrong conclusions fairly deeply; so, this post is partly an attempt to work out the right conclusions for myself, and partly a warning to those who might make the same mistakes.

All of the mistakes I'll argue against have *some good insight behind them.* They may be something which is usually true, or something which points in the direction of a real phenomenon while making an error. I may come off as nitpicking.

1. "You can't predict that you'll update in a particular direction."

Starting with an easy one.

It can be tempting to simplify conservation of expected evidence to say you can't predict the direction which your beliefs will change. This is often approximately true, and it's exactly true in symmetric cases where your starting belief is 50-50 and the evidence is equally likely to point in either direction.

To see why it is wrong in general, consider an extreme case: a universal law, which you mostly already believe to be true. At any time, you could see a counterexample, which would make you jump to complete disbelief. That's a small probability of a very large update downwards. Conservation of expected evidence implies that you must move your belief upwards when you don't see such a counterexample. But, you consider that case to be quite likely. So, considering only which *direction* your beliefs will change, you can be fairly confident that your belief in the universal law will increase -- in fact, as confident as you are in the universal law itself.

The critical point here is direction vs magnitude. Conservation of expected evidence takes magnitude as well as direction into account. The small but very probable increase is balanced by the large but very improbable decrease.

The fact that we're talking about universal laws and counterexamples may fool you into thinking about logical uncertainty. You *can* think about logical uncertainty if you want, but this phenomenon is present in the fully classical Bayesian setting; there's no funny business with non-Bayesian updates here.

Epistemic status: confidence at the level of mathematical reasoning.

2. "Yes requires the possibility of no."

Scott's recent post, yes requires the possibility of no, is fine. I'm referring to a possible mistake which one could make in applying the principle illustrated there.

> “Those who dream do not know they dream, but when you are awake, you know you are awake.” -- Eliezer, Against Modest Epistemology

Sometimes, look around, and ask myself whether I'm in a dream. When this happens, I generally conclude very confidently that I'm awake.

I am not similarly capable of determining that I'm dreaming. My dreaming self doesn't have the self-awareness to question whether he is dreaming in this way.

(Actually, very occasionally, I do. I either end up forcing myself awake, or I become lucid in the dream. Let's ignore that possibility for the purpose of the thought experiment.)

I am not claiming that my dreaming self is never deluded into thinking he is awake. On the contrary, I have those repeatedly-waking-up-only-to-find-I'm-still-dreaming dreams occasionally. In those cases, I vividly believe myself to be awake. So, it's definitely possible for me to vividly believe I'm awake and be mistaken.

What I'm saying is that, when I'm asleep, I am not able to perform *the actually good test*, where I look around and really consciously consider whether or not I might be dreaming. Nonetheless, when I can perform that check, it seems quite reliable. If I want to know if I'm awake, I can just check.

A "yes-requires-the-possibility-of-no" mindset might conclude that my "actually good test" is no good at all, because it can't say "no". I believe the exact opposite: my test seems really quite effective, because I only successfully complete it while awake.

Sometimes, your thought processes really are quite suspect; yet, there's a sanity check you can run which tells you the truth. If you're deluding yourself, the *general category* of "things which you think are simple sanity checks you can run" is not trustworthy. If you're deluding yourself, you're not even going to think about the real sanity checks. But, that *does not in itself detract from the effectiveness of the sanity check.*

The general moral in terms of conservation of expected evidence is: "'*Yes*' *only requires the possibility of silence.*". In many cases, you can meaningfully say yes without being able to meaningfully say no. For example, the axioms of set theory could prove their own inconsistency. They could *not* prove themselves consistent (without also proving themselves inconsistent). This does not detract from the effectiveness of a proof of inconsistency! Again, although the example involves logic, there's nothing funny going on with logical uncertainty; the phenomenon under discussion is understandable in fully Bayesian terms.

Symbolically: as is always the case, you don't really want to update on the raw proposition, but rather, *the fact that you observed the proposition*, to account for selection bias. Conservation of expected evidence can be written

$P(H) = P(H|E)P(E) + P(H|\neg E)P(\neg E)$, but if we re-write it to explicitly show the

"observation of evidence", it becomes

$P(H) = P(H|obs(E))P(obs(E)) + P(H|\neg obs(E))P(\neg obs(E))$. It ***does not become***

$P(H) = P(H|obs(E))P(obs(E)) + P(H|obs(\neg E))P(obs(\neg E))$. In English: evidence is balanced between making the observation and not making the observation, ***not*** between the observation and the observation of the negation.

Epistemic status: confidence at the level of mathematical reasoning for the core claim of this section. However, some applications of the idea (such as to dreams, my central example) depend on trickier philosophical issues discussed in the next section. I'm only moderately confident I have the right view there.

3. "But then what do you say to the Republican?"

I suspect that many readers are *less than fully on board* with the claims I made in the previous section. Perhaps you think I'm grossly overconfident about being awake. Perhaps you think I'm neglecting the outside view, or ignoring something to do with timeless decision theory.

A lot of my thinking in this post was generated by grappling with some points made in Inadequate Equilibria. To quote the relevant paragraph of against modest epistemology:

> Or as someone advocating what I took to be modesty recently said to me, after I explained why I thought it was sometimes okay to give yourself the discretion to disagree with mainstream expertise when the mainstream seems to be screwing up, in exactly the following words: “But then what do you say to the Republican?”

Let's put that in (pseudo-)conservation-of-expected-evidence terms: we know that just applying one's best reasoning will often leave one overconfident in one's idiosyncratic beliefs. Doesn't that mean "apply your best reasoning" is a bad test, which fails to conserve expected evidence? So, should we not adjust downward in general?

In the essay, Eliezer strongly advises allowing yourself to have an inside view even when there's an outside view which says inside views broadly similar to yours tend to be mistaken. But doesn't that go against what he said in Ethical Injunctions?

Ethical Injunctions argues that there are situations where you should not trust your reasoning, and fall back on a general rule. You do this because, in the vast majority of cases of that kind, your oh-so-clever reasoning is mistaken and the general rule saves you from the error.

In *Against Modest Epistemology*, Eliezer criticizes arguments which rely on putting arguments in very general categories and taking the outside view:

> At its epistemological core, modesty says that we should abstract up to a particular *very general* self-observation, condition on it, and then not condition on anything else because that would be inside-viewing. An observation like, “I’m familiar with the cognitive science literature discussing which debiasing techniques work well in practice, I’ve spent time on calibration and visualization

> exercises to address biases like base rate neglect, and my experience suggests that they've helped," is to be generalized up to, "I use an epistemology which I think is good." I am then to ask myself what average performance I would expect from an agent, conditioning only on the fact that the agent is using an epistemology that they think is good, and not conditioning on that agent using Bayesian epistemology or debiasing techniques or experimental protocol or mathematical reasoning or anything in particular.
>
> Only in this way can we force Republicans to agree with us... or something.

He instead advises that we should update on all the information we have, use our best arguments, reason about situations in full detail:

> If you're trying to estimate the accuracy of your epistemology, and you know what Bayes's Rule is, then—on naive, straightforward, traditional Bayesian epistemology—you ought to condition on both of these facts, and estimate *P*(accuracy|know_Bayes) instead of *P*(accuracy). Doing anything other than that opens the door to a host of paradoxes.

In *Ethical Injunctions,* he seems to warn against that very thing:

> But surely... if one is *aware of these reasons...* then one can simply redo the calculation, taking them into account. So we can rob banks if it seems like the right thing to do *after taking into account* the problem of corrupted hardware and black swan blowups. That's the rational course, right?
>
> There's a number of replies I could give to that.
>
> I'll start by saying that this is a prime example of the sort of thinking I have in mind, when I warn aspiring rationalists to beware of cleverness.

Now, maybe Eliezer has simply changed views on this over the years. Even so, that leaves *us* with the problem of how to reconcile these arguments.

I'd say the following: modest epistemology points out a *simple improvement* over the default strategy: "In any group of people who disagree, they can do better by moving their beliefs toward each other." "Lots of crazy people think they've discovered secrets of the universe, and the number of sane people who truly discover such secrets is quite small; so, we can improve the average by never believing we've discovered secrets of the universe." If we take a timeless decision theory perspective (or similar), this *is in fact an improvement*; however, it is *far from the optimal policy*, and has a form which blocks further progress.

Ethical Injunctions talks about rules with greater specificity, and less progress-blocking nature. Essentially, a proper ethical injunction is *actually the best policy you can come up with*, whereas the modesty argument stops short of that.

Doesn't the "actually best policy you can come up with" risk overly-clever policies which depend on broken parts of your cognition? Yes, but your meta-level arguments about which kinds of argument work should be independent sources of evidence from your object-level confusion. To give a toy example: let's say you really, really want 8+8 to be 12 due to some motivated cognition. You can still decide to check by applying basic arithmetic. You might *not* do this, because you *know* it isn't to the advantage of the motivated cognition. However, if you do check, it is actually quite difficult for the motivated cognition to warp basic arithmetic.

There's also the fact that choosing a modesty policy ***doesn't really help the republican***. I think that's the critical kink in the conservation-of-expected-evidence version of modest epistemology. If you, while awake, decide to doubt whether you're awake (no matter how compelling the evidence that you're awake seems to be), then *you're not really improving your overall correctness*.

So, all told, it seems like conservation of expected evidence has to be applied to the *details* of your reasoning. If your put your reasoning in a more generic category, it may appear that a much more modest conclusion is required by conservation of expected evidence. We can justify this in classical probability theory, though in this section it is even more tempting to consider exotic decision-theoretic and non-omnescience considerations than it was previously.

Epistemic status: the conclusion is mathematically true in classical Bayesian epistemology. I am subjectively >80% confident that the conclusion should hold in >90% of realistic cases, but it is unclear how to make this into a real empirical claim. I'm unsure enough of how ethical injunctions should work that I could see my views shifting significantly. I'll mention pre-rationality as one confusion I have which seems vaguely related.

4. "I can't credibly claim anything if there are incentives on my words."

Another rule which one might derive from Scott's *Yes Requires the Possibility of No* is: you can't really say anything if pressure is being put on you to say a particular thing.

Now, I agree that this is somewhat true, particularly in simple cases where pressure is being put on you to say one particular thing. However, I've suffered from learned helplessness around this. I sort of shut down when I can identify any incentives at all which could make my claims suspect, and hesitate to claim anything. This isn't a very useful strategy. Either "just say the truth", or "just say whatever you feel you're expected to say" are *both* likely better strategies.

One idea is to "call out" the pressure you feel. "I'm having trouble saying anything because I'm worried what you will think of me." This isn't always a good idea, but it can often work fairly well. Someone who *is* caving to incentives isn't very likely to say something like that, so it provides some evidence that you're being genuine. It can also open the door to other ways you and the person you're talking to can solve the incentive problem.

You can also "call out" something even if you're unable or unwilling to explain. You just say something like "there's some *thing* going on"... or "I'm somehow frustrated with this situation"... or whatever you can manage to say.

This "call out" idea also works (to some extent) on motivated cognition. Maybe you're worried about the social pressure on your beliefs because it might influence the accuracy of those beliefs. Rather than stressing about this and going into a spiral of self-analysis, you can just state to yourself that that's a thing which might be going on, and move forward. Making it explicit might open up helpful lines of thinking later.

Another thing I want to point out is that most people are willing to place at least a little faith in your honesty (and not irrationally so). Just because you have a story in

mind where they should assume you're lying doesn't mean that's the only possibility they are -- or should be -- considering. One problematic incentive doesn't fully determine the situation. (This one also applies internally: identifying one relevant bias or whatever doesn't mean you should block off that part of yourself.)

Epistemic status: low confidence. I imagine I would have said something very different if I were more an expert in this particular thing.

5. "Your true reason screens off any other evidence your argument might include."

In The Bottom Line, Eliezer describes a clever arguer who first writes the conclusion which they want to argue for at the bottom of a sheet of paper, and then comes up with as many arguments as they can to put above that. In the thought experiment, the clever arguer's conclusion is actually determined by who can pay the clever arguer more. Eliezer says:

> So the handwriting of the curious inquirer is entangled with the signs and portents and the contents of the boxes, whereas the handwriting of the clever arguer is evidence only of which owner paid the higher bid. There is a great difference in the indications of ink, though one who foolishly read aloud the ink-shapes might think the English words sounded similar.

Now, Eliezer is trying to make a point about *how you form your own beliefs* -- that the quality of the process which determines which claims *you* make is what matters, and the quality of any rationalizations you give doesn't change that.

However, reading that, I came away with the mistaken idea that *someone listening to a clever arguer should ignore all the clever arguments*. Or, generalizing further, *what you should do when listening to any argument is try to figure out what process wrote the bottom line*, ignoring any other evidence provided.

This isn't the worst possible algorithm. You really *should* heavily discount evidence provided by clever arguers, because it has been heavily cherry-picked. And almost everyone does a great deal of clever arguing. Even a hardboiled rationalist will tend to present evidence for the point they're trying to make rather than against (perhaps because that's a fairly good strategy for explaining things -- sampling evidence at random isn't a very efficient way of conversing!).

However, ignoring arguments and attending only to the original causes of belief has some absurd consequences. Chief among them is: it would imply that you should ignore mathematical proofs if the person who came up with the proof only searched for positive proofs and wouldn't have spend time trying to prove the opposite. (This ties in with the very first section -- failing to find a proof is like remaining silent.)

This is bonkers. Proof is proof. And again, this isn't some special non-Bayesian phenomenon due to logical uncertainty. A Bayesian can and should recognize decisive evidence, whether or not it came from a clever arguer.

Yet, I really held this position for a while. I treated mathematical proofs as an exceptional case, rather than as a phenomenon continuous with weaker forms of evidence. If a clever arguer presented anything *short* of a mathematical proof, I would remind myself of how convincing cherry-picked evidence can seem. And I'd notice how almost everyone mostly cherry-picked when explaining their views.

This strategy was throwing out data when it has been contaminated by selection bias, rather than making a model of the selection bias so that I could update on the data appropriately. It might be a good practice in scientific publications, but if you take it as a universal, you could find reasons to throw out just about *everything* (especially if you start worrying about anthropic selection effects).

The right thing to do is closer to this: figure out how convincing you expect evidence to look *given* the extent of selection bias. Then, update on the *difference* between what you see and what's expected. If a clever arguer makes a case which is much better than what you would have expected they could make, you can update up. If it is *worse* that you'd expect, even if the evidence would otherwise look favorable, you update *down*.

My view also made me uncomfortable presenting a case for my own beliefs, because I would think of myself as a clever-arguer any time I did something other than recount the actual historical causes of my belief (or honestly reconsider my belief on the spot). Grognor made a similar point in Unwilling or Unable to Explain:

> Let me back up. Speaking in good faith entails giving the real reasons you believe something rather than a persuasive impromptu rationalization. Most people routinely do the latter without even noticing. I'm sure I still do it without noticing. But when I do notice I'm about to make something up, instead I clam up and say, "I can't explain the reasons for this claim." I'm not willing to disingenuously reference a scientific paper that I'd never even heard of when I formed the belief it'd be justifying, for example. In this case silence is the only feasible alternative to speaking in bad faith.

While I think there's something to this mindset, I no longer think it makes sense to clam up when you can't figure out how you originally came around to the view which you now hold. If you think there are other good reasons, you can give them without violating good faith.

Actually, I really wish I could draw a sharper line here. I'm essentially claiming that a little cherry-picking is OK if you're just trying to convince someone of the view which you see as the truth, so long as you're not intentionally hiding anything. This is an uncomfortable conclusion.

Epistemic status: confident that the views I claim are mistaken are mistaken. Less confident about best-practice claims.

6. "If you can't provide me with a reason, I have to assume you're wrong."

If you take the conclusion of the previous section too far, you might reason as follows: if someone is trying to claim X, surely they're trying to give you some evidence toward X. If they claim X and then you challenge them for evidence, they'll try to tell you any evidence they have. So, if they come up with *nothing*, you have to update *down,* since you would have updated upwards otherwise. Right?

I think most people make this mistake due to simple conversation norms: when navigating a conversation, people have to figure out what everyone else is willing to assume, in order to make sensible statements with minimal friction. So, we look for obvious signs of whether a statement was accepted by everyone vs rejected. If someone was asked to provide a reason for a statement they made and failed to do so, that's a fairly good signal that the statement hasn't been accepted into the common background assumptions for the conversation. The fact that other people are likely to use this heuristic as well makes the signal even stronger. So, assertions which can't be backed up with reasons are likely to be rejected.

This is almost the opposite mistake from the previous section; the previous one was *justifications don't matter*, whereas this idea is *only justifications matter.*

I think something good happens when everyone in a conversation recognizes that *people can believe things for good reason without being able to articulate those reasons*. (This includes yourself!)

You can't just give everyone a pass to make unjustified claims and assert that they have strong inarticulable reasons. Or rather, you *can* give everyone a pass to do that, but you don't have to take them seriously when they do it. However, in environments of high intellectual trust, you *can* take it seriously. Indeed, applying the usual heuristic will likely cause you to update in the wrong direction.

Epistemic status: moderately confident.

Conclusion

I think all of this is fairly important -- if you're like me, you've likely made some mistakes along these lines. I also think there are many issues related to conservation of expected evidence which I still don't fully understand, such as explanation vs rationalization, ethical injunctions and pre-rationality. Tsuyoku Naritai!

Book Review: The Secret Of Our Success

[Previously in sequence: Epistemic Learned Helplessness]

I.

"Culture is the secret of humanity's success" sounds like the most vapid possible thesis. *The Secret Of Our Success* by anthropologist Joseph Heinrich manages to be an amazing book anyway.

Heinrich wants to debunk (or at least clarify) a popular view where humans succeeded because of our raw intelligence. In this view, we are smart enough to invent neat tools that help us survive and adapt to unfamiliar environments.

Against such theories: we cannot actually do this. Heinrich walks the reader through many stories about European explorers marooned in unfamiliar environments. These explorers usually starved to death. They starved to death in the middle of endless plenty. Some of them were in Arctic lands that the Inuit considered among their richest hunting grounds. Others were in jungles, surrounded by edible plants and animals. One particularly unfortunate group was in Alabama, and would have perished entirely if they hadn't been captured and enslaved by local Indians first.

These explorers had many advantages over our hominid ancestors. For one thing, their exploration parties were made up entirely of strong young men in their prime, with no need to support women, children, or the elderly. They were often selected for their education and intelligence. Many of them were from Victorian Britain, one of the most successful civilizations in history, full of geniuses like Darwin and Galton. Most of them had some past experience with wilderness craft and survival. But despite their big brains, when faced with the task our big brains supposedly evolved for – figuring out how to do hunting and gathering in a wilderness environment – they failed pathetically.

Nor is it surprising that they failed. Hunting and gathering is actually really hard. Here's Heinrich's description of how the Inuit hunt seals:

> You first have to find their breathing holes in the ice. It's important that the area around the hole be snow-covered—otherwise the seals will hear you and vanish. You then open the hole, smell it to verify it's still in use (what do seals smell like?), and then assess the shape of the hole using a special curved piece of caribou antler. The hole is then covered with snow, save for a small gap at the top that is capped with a down indicator. If the seal enters the hole, the indicator moves, and you must blindly plunge your harpoon into the hole using all your weight. Your harpoon should be about 1.5 meters (5ft) long, with a detachable tip that is tethered with a heavy braid of sinew line. You can get the antler from the previously noted caribou, which you brought down with your driftwood bow.
>
> The rear spike of the harpoon is made of extra-hard polar bear bone (yes, you also need to know how to kill polar bears; best to catch them napping in their dens). Once you've plunged your harpoon's head into the seal, you're then in a wrestling match as you reel him in, onto the ice, where you can finish him off with the aforementioned bear-bone spike.
>
> Now you have a seal, but you have to cook it. However, there are no trees at this latitude for wood, and driftwood is too sparse and valuable to use routinely for fires. To have a reliable fire, you'll need to carve a lamp from soapstone (you know what soapstone looks like, right?), render some oil for the lamp from blubber, and make a wick out of a particular species of moss. You will also need water. The pack ice is frozen salt water, so using it for drinking will just make you dehydrate faster. However, old sea ice has lost most of its salt, so it can be melted to make potable water. Of course, you need to be

> able to locate and identify old sea ice by color and texture. To melt it, make sure you have enough oil for your soapstone lamp.

No surprise that stranded explorers couldn't figure all this out. It's more surprising that the Inuit *did*. And although the Arctic is an unusually hostile place for humans, Heinrich makes it clear that hunting-gathering techniques of this level of complexity are standard everywhere. Here's how the Indians of Tierra del Fuego make arrows:

> Among the Fuegians, making an arrow requires a 14-step procedure that involves using seven different tools to work six different materials. Here are some of the steps:
>
> – The process begins by selecting the wood for the shaft, which preferably comes from chaura, a bushy, evergreen shrub. Though strong and light, this wood is a non-intuitive choice since the gnarled branches require extensive straightening (why not start with straighter branches?).
>
> – The wood is heated, straightened with the craftsman's teeth, and eventually finished with a scraper. Then, using a pre-heated and grooved stone, the shaft is pressed into the grooves and rubbed back and forth, pressing it down with a piece of fox skin. The fox skin become impregnated with the dust, which prepares it for the polishing stage (Does it have to be fox skin?).
>
> – Bits of pitch, gathered from the beach, are chewed and mixed with ash (What if you don't include the ash?).
>
> – The mixture is then applied to both ends of a heated shaft, which must then be coated with white clay (what about red clay? Do you have to heat it?). This prepares the ends for the fletching and arrowhead.
>
> – Two feathers are used for the fletching, preferably from upland geese (why not chicken feathers?).
>
> – Right-handed bowman must use feathers from the left wing of the bird, and vice versa for lefties (Does this really matter?).
>
> – The feathers are lashed to the shaft using sinews from the back of the guanaco, after they are smoothed and thinned with water and saliva (why not sinews from the fox that I had to kill for the aforementioned skin?).
>
> Next is the arrowhead, which must be crafted and then attached to the shaft, and of course there is also the bow, quiver and archery skills. But, I'll leave it there, since I think you get the idea.

How do hunter-gatherers know how to do all this? We usually summarize it as "culture". How did it form? Not through some smart Inuit or Fuegian person reasoning it out; if that had been it, smart European explorers should have been able to reason it out too.

The obvious answer is "cultural evolution", but Heinrich isn't much better than anyone else at taking the mystery out of this phrase. Trial and error must have been involved, and less successful groups/people imitating the techniques of more successful ones. But is that really a satisfying explanation?

I found the chapter on language a helpful reminder that we already basically accept something like this is true. How did language get invented? I'm especially interested in this question because of my brief interactions with conlanging communities – people who try to construct their own languages as a hobby or as part of a fantasy universe, like Tolkien did with Elvish. Most people are *terrible* at this; their languages are either unusable, or exact clones of English. Only people who (like Tolkien) already have years of formal training in linguistics can do a remotely passable job. And you're telling me the original languages were

invented by cavemen? Surely there was no committee of Proto-Indo-European nomads that voted on whether to have an inflecting or agglutinating tongue? Surely nobody ran out of their cave shouting “Eureka!” after having discovered the interjection? We just kind of accept that after cavemen working really hard to communicate with each other, eventually language – still one of the most complicated and impressive productions of the human race – just sort of happened.

Taking the generation of culture as secondary to this kind of mysterious process, Heinrich turns to its transmission. If cultural generation happens at a certain rate, then the fidelity of transmission determines whether a given society advances, stagnates, or declines.

For Heinrich, humans started becoming more than just another species of monkey when we started transmitting culture with high fidelity. Some anthropoligsts talk about the Machiavellian Intelligence Hypothesis – the theory that humans evolved big brains in order to succeed at social manuevering and climbing dominance hierarchies. Heinrich counters with his own Cultural Intelligence Hypothesis – humans evolved big brains in order to be able to maintain things like Inuit seal hunting techniques. Everything that separates us from the apes is part of an evolutionary package designed to help us maintain this kind of culture, exploit this kind of culture, or adjust to the new abilities that this kind of culture gave us.

II.

Secret gives many examples of many culture-related adaptations, and not all are in the brain.

One of the most important differences between man and ape is our puny digestive tracts:

> Our mouths are the size of the squirrel monkey’s, a species that weighs less than three pounds. Chimpanzees can open their mouths twice as ide as we can and hold substantial amounts of food compressed between their lips and large teeth. We also have puny jaw muscles that reach up only to just below our ears. Other primates’ jaw muscles stretch to the top of their heads, where they sometimes even latch onto a central bony ridge. Our stomachs are small, having only a third of the surface area that we’d expect for a primate of our size, and our colons are too short, being only 60% of their expected mass.

Compared to other animals, we have such atrophied digestive tracts that we shouldn’t be able to live. What saves us? All of our food processing techniques, especially cooking, but also chopping, rinsing, boiling, and soaking. We’ve done much of the work of digestion before food even enters our mouths. Our culture teaches us how to do this, both in broad terms like “hold things over fire to cook them” and in specific terms like “this plant needs to be soaked in water for 24 hours to leach out the toxins”. Each culture has its own cooking knowledge related to the local plants and animals; a frequent cause of death among European explorers was cooking things in ways that didn’t unlock any of the nutrients, and so starving while apparently well-fed.

All of this is cultural. Heinrich is kind of cruel in his insistence on this. He recommends readers go outside and try to start a fire. He even gives some helpful hints – flint is involved, rubbing two sticks together works for some people, etc. He predicts – and stories I’ve heard from unfortunate campers confirm – that you will not be able to do this, despite an IQ far beyond that of most of our hominid ancestors. In fact, some groups (most notably the aboriginal Tasmanians) seem to have lost the ability to make fire, and never rediscovered it. Fire-making was discovered a small number of times, maybe once, and has been culturally transmitted since then.

And food processing techniques are even more complicated. Nixtamalization of corn, necessary to prevent vitamin deficiencies, involves soaking the corn in a solution containing ground-up burnt seashells. The ancient Mexicans discovered this and lived off corn just fine for millennia. When the conquistadors took over, they ignored it and ate corn straight. For four hundred years, Europeans and Americans ate unnixtamalized corn. By official statistics,

three million Americans came down with corn-related vitamin deficiencies during this time, and up to a hundred thousand died. It wasn't until 1937 that Western scientists discovered which vitamins were involved and developed an industrial version of nixtamalization that made corn safe. Early 1900s Americans were very smart and had lots of advantages over ancient Mexicans. But the ancient Mexicans' culture got this one right in a way it took Westerners centuries to match.

Humans are persistence hunters: they cannot run as fast as gazelles, but they can keep running for longer than gazelles (or almost anything else). Why did we evolve into that niche? The secret is our ability to carry water. Every hunter-gatherer culture has invented its own water-carrying techniques, usually some kind of waterskin. This allowed humans to switch to perspiration-based cooling systems, which allowed them to run as long as they want.

And humans are consumate tool users. In some cases, we evolved in order to use tools better; our hands outclass those of any other ape in terms of finesse. In other cases, we devolved systems that were no longer necessary once tools took over. We are vastly weaker than any other ape. Heinrich describes a circus act of the 1940s where the ringmaster would challenge strong men in the audience to wrestle a juvenile chimpanzee. The chimpanzee was tied up, dressed in a mask that prevented it from biting, and wearing soft gloves that prevented it from scratching. No human ever lasted more than five seconds. Our common ancestor with other apes grew weaker and weaker as we became more and more reliant on artificial weapons to give us an advantage.

III.

But most of our differences from other apes are indeed in the brain. They're just not necessarily where you would expect.

Tomasello et al tested human toddlers vs. apes on a series of traditional IQ type questions. The match-up was surprisingly fair; in areas like memory, logic, and spatial reasoning, the three species did about the same. But in ability to learn from another person, humans wiped the floor with the other two ape species:

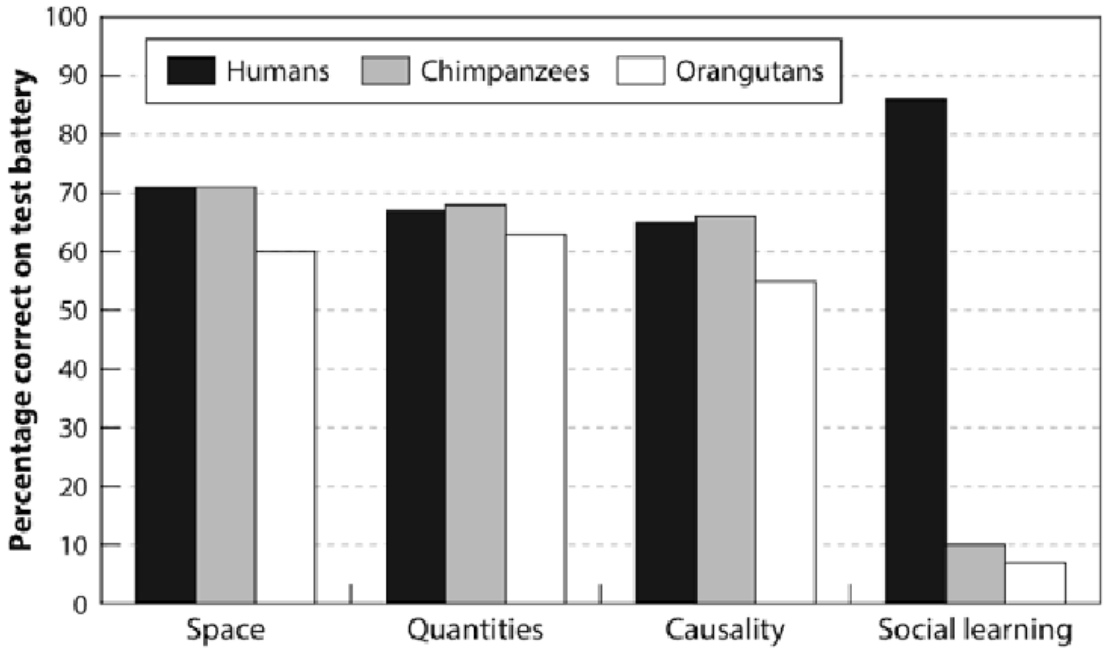


Figure 2.2. Average performance on four sets of cognitive tests with chimpanzees, orangutans, and toddlers.

Remember, Heinrich thinks culture accumulates through random mutation. Humans don't have control over how culture gets generated. They have more control over how much of it gets transmitted to the next generation. If 100% gets transmitted, then as more and more mutations accumulate, the culture becomes better and better. If less than 100% gets transmitted, then at some point new culture gained and old culture lost fall into equilibrium, and your society stabilizes at some higher or lower technological level. This means that transmitting culture to the next generation is maybe the core human skill. The human brain is optimized to make this work as well as possible.

Human children are obsessed with learning things. And they don't learn things randomly. There seem to be "biases in cultural learning", ie slots in an infant's mind that they know need to be filled with knowledge, and which they preferentially seek out the knowledge necessary to fill.

One slot is for language. Human children naturally listen to speech (as early as in the womb). They naturally prune the phonemes they are able to produce and distinguish to the ones in the local language. And they naturally figure out how to speak and understand what people are saying, even though learning a language is hard even for smart adults.

Another slot is for animals. In a world where megafauna has been relegated to zoos, we *still* teach children their ABCs with "L is for lion" and "B is for bear", and children *still* read picture books about Mr. Frog and Mrs. Snake holding tea parties. Heinrich suggests that just as the young brain is hard-coded to want to learn language, so it is hard-coded to want to learn the local animal life (little boys' vehicle obsession may be a weird outgrowth of this; buses and trains are the closest thing to local megafauna that most of them will encounter).

Another slot is for plants:

To see this system in operation, let's consider how infants respond to unfamiliar plants. Plants are loaded with prickly thorns, noxious oils, stinging nettles and dangerous toxins, all genetically evolved to prevent animals like us from messing with them. Given our species wide geographic range and diverse use of plants as foods, medicines and construction materials, we ought to be primed to both learn about plants and avoid their dangers. To explore this idea in the lab, the psychologists Annie Wertz and Karen Wynn first gave infants, who ranged in age from eight to eighteen months, an opportunity to touch novel plants (basil and parsley) and artifacts, including both novel objects and common ones, like wooden spoons and small lamps.

The results were striking. Regardless of age, many infants flatly refused to touch the plants at all. When they did touch them, they waited substantially longer than they did with the artifacts. By contrast, even with the novel objects, infants showed none of this reluctance. This suggests that well before one year of age infants can readily distinguish plants from other things, and are primed for caution with plants. But, how do they get past this conservative predisposition?

The answer is that infants keenly watch what other people do with plants, and are only inclined to touch or eat the plants that other people have touched or eaten. In fact, once they get the 'go ahead' via cultural learning, they are suddenly interested in eating plants. To explore this, Annie and Karen exposed infants to models who both picked fruit from plants and also picked fruit-like things from an artifact of similar size and shape to the plant. The models put both the fruit and the fruit-like things in their mouths. Next, the infants were given a choice to go for the fruit (picked from the plant) or the fruit-like things picked from the object. Over 75% of the time the infants went for the fruit, not the fruit-like things, since they'd gotten the 'go ahead' via cultural learning.

As a check, the infants were also exposed to models putting the fruit or fruit-like things behind their ears(not in their mouths). In this case, the infants went for the fruit or fruit-like things in equal measure. It seems that plants are most interesting if you can eat them, but only if you have some cultural learning cues that they aren't toxic.

After Annie first told me about her work while I was visiting Yale in 2013, I went home to test it on my 6-month-old son, Josh. Josh seemed very likely to overturn Annie's hard empirical work, since he immediately grasped anything you gave him and put it rapidly in his mouth. Comfortable in his mom's arms, I first offered Josh a novel plastic cube. He delighted in grapping it and shoving it directly into his mouth, without any hesitation. Then, I offered him a sprig of arugula. He quickly grabbed it, but then paused, looked with curious uncertainty at it, and then slowly let it fall from his hand while turning to hug his mom.

It's worth pointing out how rich the psychology is here. Not only do infants have to recognize that plants are different from objects of similar size, shape and color, but they need to create categories for types of plants, like basil and parsley, and distinguish 'eating' from just 'touching'. It does them little good to code their observation of someone eating basil as 'plants are good to eat' since that might cause them to eat poisonous plants as well as basil. But, it also does them little good to narrowly code the observation as 'that particular sprig of basil is good to eat' since that particular sprig has just been eaten by the person they are watching. This another content bias in cultural learning.

This ties into the more general phenomenon of figuring out what's edible. Most Westerners learn insects aren't edible; some Asians learn that they are. This feels deeper than just someone telling you insects aren't edible and you believing them. When I was in Thailand, my guide offered me a giant cricket, telling me it was delicious. I believed him when he said it was safe to eat, I even believed him when he said it tasted good to him, but my conditioning won out – I didn't eat the cricket. There seems to be some process where a child's brain learns what is and isn't locally edible, then hard-codes it against future change.

(Or so they say; I've never been able to eat shrimp either.)

Another slot is for gender roles. By now we've all heard the stories of progressives who try to raise their children without any exposure to gender. Their failure has sometimes been taken as evidence that gender is hard-coded. But it can't be quite that simple: some modern gender roles, like girls = pink, are far from obvious or universal. Instead, it looks like children have a hard-coded slot that gender roles go into, work hard to figure out what the local gender roles are (even if their parents are trying to confuse them), then latch onto them and don't let go.

In the Cultural Intelligence Hypothesis, humans live in obligate symbiosis with a culture. A brain without an associated culture is incomplete and not very useful. So the infant brain is adapted to seek out the important aspects of its local culture almost from birth and fill them into the appropriate slots in order to become whole.

IV.

The next part of the book discusses post-childhood learning. This plays an important role in hunter-gatherer tribes:

> While hunters reach their peak strength and speed in their twenties, individual hunting success does not peak until around age 30, because success depends more on know-how and refined skills than on physical prowess.

This part of the book made most sense in the context of examples like the Inuit seal-hunting strategy which drove home just how complicated and difficult hunting-gathering was. Think less "Boy Scouts" and more "PhD"; a primitive tribesperson's life requires mastery of various complicated technologies and skills. And the difference between "mediocre hunter" and "great hunter" can be the difference between high status (and good mating opportunities) and low status, or even between life and death. Hunter-gatherers really want to learn the essentials of their hunter-gatherer lifestyle, and learning it is really hard. Their heuristics are:

Learn from people who are good at things and/or widely-respected. If you haven't already read about the difference between dominance and prestige hierarchies, check out Kevin Simler's blog post on the topic. People will fear and obey authority figures like kings and chieftains, but they give a different kind of respect ("prestige") to people who seem good at things. And since it's hard to figure out who's good at things (can a non-musician who wants to start learning music tell the difference between a merely good performer and one of the world's best?) most people use the heuristic of respecting the people who other people respect. Once you identify someone as respect-worthy, you strongly consider copying them in, well, everything:

> To understand prestige as a social phenomenon, it's crucial to realize that it's often difficult to figure out what precisely makes someone successful. In modern societies, the success of a star NBA basketball player might arise from his:
>
> (1) intensive practice in the offseason
> (2) sneaker preference
> (3) sleep schedule
> (4) pre-game prayer
> (5) special vitamins
> (6) taste for carrots
>
> Any or all of these might increase his success. A naïve learner can't tell all the causal links between an individual's practices and his success. As a consequence, learners often copy their chosen models broadly across many domains. Of course, learners may place more weight on domains that for one reason or other seem more causally relevant to the model's success. This copying often includes the model's personal habits or styles as well as their goals and motivations, since these may be linked to their success. This "if in

> doubt, copy it" heuristic is one of the reasons why success in one domain converts to influence across a broad range of domains.
>
> The immense range of celebrity endorsements in modern societies shows the power of prestige. For example, NBA star Lebron James, who went directly from High School to the pros, gets paid millions to endorse State Farm Insurance. Though a stunning basketball talent, it's unclear why Mr. James is qualified to recommend insurance companies. Similarly, Michael Jordan famously wore Hanes underwear and apparently Tiger Woods drove Buicks. Beyonce' drinks Pepsi (at least in commercials). What's the connection between musical talent and sugary cola beverages?
>
> Finally, while new medical findings and public educational campaigns only gradually influence women's approach to preventive medicine, Angelina Jolie's single OP-ED in the New York Times, describing her decision to get a preventive double mastectomy after learning she had the 'faulty' BRCA1 gene, flooded clinics from the U.K. to New Zealand with women seeking genetic screenings for breast cancer. Thus, an unwanted evolutionary side effect, prestige turns out to be worth millions, and represents a powerful and underutilized public health tool.

Of course, this creates the risk of prestige cascades, where some irrelevant factor (Heinrich mentions being a reality show star) catapults someone to fame, everyone talks about them, and you end up with Muggeridge's definition of a celebrity: someone famous for being famous.

Some of this makes more sense if you go back to the evolutionary roots, and imagine watching the best hunter in your tribe to see what his secret is, or being nice to him in the hopes that he'll take you under his wing and teach you stuff.

(but if all this is true, shouldn't public awareness campaigns that hire celebrity spokespeople be wild successes? Don't they just as often fail, regardless of how famous a basketball player they can convince to lecture schoolchildren about how Winners Don't Do Drugs?)

Learn from people who are like you. If you are a man, it is probably a bad idea to learn fashion by observing women. If you are a servant, it is probably a bad idea to learn the rules of etiquette by observing how the king behaves. People are naturally inclined to learn from people more similar to themselves.

Heinrich ties this in to various studies showing that black students learn best from a black teacher, female students from a female teacher, et cetera.

Learn from old people. Humans are almost unique in having menopause; most animals keep reproducing until they die in late middle-age. Why does evolution want humans to stick around without reproducing?

Because old people have already learned the local culture and can teach it to others. Heinrich asks us to throw out any personal experience we have of elders; we live in a rapidly-changing world where an old person is probably "behind the times". But for most of history, change happened glacially slowly, and old people would have spent their entire lives accumulating relevant knowledge. Imagine a world where when a Silicon Valley programmer can't figure out how to make his code run, he calls up his grandfather, who spent fifty years coding apps for Google and knows every programming language inside and out.

Sometimes important events only happen once in a generation. Heinrich tells the story of an Australian aboriginal tribe facing a massive drought. Nobody knew what to do except Paralji, the tribe's oldest man, who had lived through the last massive drought and remembered where his own elders had told him to find the last-resort waterholes.

This same dynamic seems to play out even in other species:

> In 1993, a severe drought hit Tanzania, resulting in the death of 20% of the African elephant calves in a population of about 200. This population contained 21 different families, each of which was led by a single matriarch. The 21 elephant families were divided into 3 clans, and each clan shared the same territory during the wet season (so, they knew each other). Researchers studying these elephants have analyzed the survival of the calves and found that families led by older matriarchs suffered fewer deaths of their calves during this drought.
>
> Moreover, two of the three elephant clans unexpectedly left the park during the drought, presumably in search of water, and both had much higher survival rates than the one clan that stayed behind. It happens that these severe droughts only hit about once every four to five decades, and the last one hit about 1960. After that, sadly, elephant poaching in the 1970's killed off many of the elephants who would have been old enough in 1993 to recall the 1960 drought. However, it turns out that exactly one member of each of the two clans who left the park, and survived more effectively, were old enough to recall life in 1960. This suggests, that like Paralji in the Australian desert, they may have remembered what to do during a severe drought, and led their groups to the last water refuges. In the clan who stayed behind, the oldest member was born in 1960, and so was too young to have recalled the last major drought.
>
> More generally, aging elephant matriarchs have a big impact on their families, as those led by older matriarchs do better at identifying and avoiding predators (lions and humans), avoiding internal conflicts and identifying the calls of their fellow elephants. For example, in one set of field experiments, researchers played lion roars from both male and female lions, and from either a single lion or a trio of lions. For elephants, male lions are much more dangerous than females, and of course, three lions are always worse than only one lion. All the elephants generally responded with more defensive preparations when they heard three lions vs. one. However, only the older matriarchs keenly recognized the increased dangers of male lions over female lions, and responded to the increased threat with elephant defensive maneuvers.

V.

I was inspired to read *Secret* by this review on Scholar's Stage. I hate to be unoriginal, but after reading the whole book, I agree that the three sections Tanner cites – on divination, on manioc, and on shark taboos – are by far the best and most fascinating.

On divination:

> When hunting caribou, Naskapi foragers in Labrador, Canada, had to decide where to go. Common sense might lead one to go where one had success before or to where friends or neighbors recently spotted caribou.
>
> However, this situation is like [the Matching Pennies game]. The caribou are mismatchers and the hunters are matchers. That is, hunters want to match the locations of caribou while caribou want to mismatch the hunters, to avoid being shot and eaten. If a hunter shows any bias to return to previous spots, where he or others have seen caribou, then the caribou can benefit (survive better) by avoiding those locations (where they have previously seen humans). Thus, the best hunting strategy requires randomizing.
>
> Can cultural evolution compensate for our cognitive inadequacies? Traditionally, Naskapi hunters decided where to go to hunt using divination and believed that the shoulder bones of caribou could point the way to success. To start the ritual, the shoulder blade was heated over hot coals in a way that caused patterns of cracks and burnt spots to form. This patterning was then read as a kind of map, which was held in a pre-specified orientation. The cracking patterns were (probably) essentially random from the point of view of hunting locations, since the outcomes depended on myriad details about the bone, fire, ambient temperature, and heating process. Thus, these divination rituals may

> have provided a crude randomizing device that helped hunters avoid their own decision-making biases.
>
> This is not some obscure, isolated practice, and other cases of divination provide more evidence. In Indonesia, the Kantus of Kalimantan use bird augury to select locations for their agricultural plots. Geographer Michael Dove argues that two factors will cause farmers to make plot placements that are too risky. First, Kantu ecological models contain the Gambler's Fallacy, and lead them to expect floods to be less likely to occur in a specific location after a big flood in that location (which is not true). Second…Kantus pay attention to others' success and copy the choices of successful households, meaning that if one of their neighbors has a good yield in an area one year, many other people will want to plant there in the next year. To reduce the risks posed by these cognitive and decision-making biases, Kantu rely on a system of bird augury that effectively randomizes their choices for locating garden plots, which helps them avoid catastrophic crop failures. Divination results depend not only on seeing a particular bird species in a particular location, but also on what type of call the bird makes (one type of call may be favorable, and another unfavorable).
>
> The patterning of bird augury supports the view that this is a cultural adaptation. The system seems to have evolved and spread throughout this region since the 17th century when rice cultivation was introduced. This makes sense, since it is rice cultivation that is most positively influenced by randomizing garden locations. It's possible that, with the introduction of rice, a few farmers began to use bird sightings as an indication of favorable garden sites. On-average, over a lifetime, these farmers would do better – be more successful – than farmers who relied on the Gambler's Fallacy or on copying others' immediate behavior. Whatever the process, within 400 years, the bird augury system spread throughout the agricultural populations of this Borneo region. Yet, it remains conspicuously missing or underdeveloped among local foraging groups and recent adopters of rice agriculture, as well as among populations in northern Borneo who rely on irrigation. So, bird augury has been systematically spreading in those regions where it's most adaptive.

Scott Aaronson has written about how easy it is to predict people trying to "be random":

> In a class I taught at Berkeley, I did an experiment where I wrote a simple little program that would let people type either "f" or "d" and would predict which key they were going to push next. It's actually very easy to write a program that will make the right prediction about 70% of the time. Most people don't really know how to type randomly. They'll have too many alternations and so on. There will be all sorts of patterns, so you just have to build some sort of probabilistic model. Even a very crude one will do well. I couldn't even beat my own program, knowing exactly how it worked. I challenged people to try this and the program was getting between 70% and 80% prediction rates. Then, we found one student that the program predicted exactly 50% of the time. We asked him what his secret was and he responded that he "just used his free will."

But being genuinely random is important in pursuing mixed game theoretic strategies. Heinrich's view is that divination solved this problem effectively.

I'm reminded of the Romans using augury to decide when and where to attack. This always struck me as crazy; generals are going to risk the lives of thousands of soldiers because they saw a weird bird earlier that morning? But war is a classic example of when a random strategy can be useful. If you're deciding whether to attack the enemy's right vs. left flank, it's important that the enemy can't predict your decision and send his best defenders there. If you're generally predictable – and Scott Aaronson says you are – then outsourcing your decision to weird birds might be the best way to go.

And then there's manioc. This is a tuber native to the Americas. It contains cyanide, and if you eat too much of it, you get cyanide poisoning. From Heinrich:

In the Americas, where manioc was first domesticated, societies who have relied on bitter varieties for thousands of years show no evidence of chronic cyanide poisoning. In the Colombian Amazon, for example, indigenous Tukanoans use a multistep, multiday processing technique that involves scraping, grating, and finally washing the roots in order to separate the fiber, starch, and liquid. Once separated, the liquid is boiled into a beverage, but the fiber and starch must then sit for two more days, when they can then be baked and eaten. Figure 7.1 shows the percentage of cyanogenic content in the liquid, fiber, and starch remaining through each major step in this processing.

Such processing techniques are crucial for living in many parts of Amazonia, where other crops are difficult to cultivate and often unproductive. However, despite their utility, one person would have a difficult time figuring out the detoxification technique. Consider the situation from the point of view of the children and adolescents who are learning the techniques. They would have rarely, if ever, seen anyone get cyanide poisoning, because the techniques work. And even if the processing was ineffective, such that cases of goiter (swollen necks) or neurological problems were common, it would still be hard to recognize the link between these chronic health issues and eating manioc. Most people would have eaten manioc for years with no apparent effects. Low cyanogenic varieties are typically boiled, but boiling alone is insufficient to prevent the chronic conditions for bitter varieties. Boiling does, however, remove or reduce the bitter taste and prevent the acute symptoms (e.g., diarrhea, stomach troubles, and vomiting).

So, if one did the common-sense thing and just boiled the high-cyanogenic manioc, everything would seem fine. Since the multistep task of processing manioc is long, arduous, and boring, sticking with it is certainly non-intuitive. Tukanoan women spend about a quarter of their day detoxifying manioc, so this is a costly technique in the short term. Now consider what might result if a self-reliant Tukanoan mother decided to drop any seemingly unnecessary steps from the processing of her bitter manioc. She might critically examine the procedure handed down to her from earlier generations and conclude that the goal of the procedure is to remove the bitter taste. She might then experiment with alternative procedures by dropping some of the more labor-intensive or time-consuming steps. She'd find that with a shorter and much less labor-intensive process, she could remove the bitter taste. Adopting this easier protocol, she would have more time for other activities, like caring for her children. Of course, years or decades later her family would begin to develop the symptoms of chronic cyanide poisoning.

Thus, the unwillingness of this mother to take on faith the practices handed down to her from earlier generations would result in sickness and early death for members of her family. Individual learning does not pay here, and intuitions are misleading. The problem is that the steps in this procedure are causally opaque—an individual cannot readily infer their functions, interrelationships, or importance. The causal opacity of many cultural adaptations had a big impact on our psychology.

Wait. Maybe I'm wrong about manioc processing. Perhaps it's actually rather easy to individually figure out the detoxification steps for manioc? Fortunately, history has provided a test case. At the beginning of the seventeenth century, the Portuguese transported manioc from South America to West Africa for the first time. They did not, however, transport the age-old indigenous processing protocols or the underlying commitment to using those techniques. Because it is easy to plant and provides high yields in infertile or drought-prone areas, manioc spread rapidly across Africa and became a staple food for many populations. The processing techniques, however, were not readily or consistently regenerated. Even after hundreds of years, chronic cyanide poisoning remains a serious health problem in Africa. Detailed studies of local preparation techniques show that high levels of cyanide often remain and that many individuals carry low levels of cyanide in their blood or urine, which haven't yet manifested in symptoms. In some places, there's no processing at all, or sometimes the processing actually increases the cyanogenic content. On the positive side, some African

groups have in fact culturally evolved effective processing techniques, but these techniques are spreading only slowly.

Rationalists always wonder: how come people aren't more rational? How come you can prove a thousand times, using Facts and Logic, that something is stupid, and yet people will still keep doing it?

Heinrich hints at an answer: for basically all of history, using reason would get you killed.

A reasonable person would have figured out there was no way for oracle-bones to accurately predict the future. They would have abandoned divination, failed at hunting, and maybe died of starvation.

A reasonable person would have asked why everyone was wasting so much time preparing manioc. When told "Because that's how we've always done it", they would have been unsatisfied with that answer. They would have done some experiments, and found that a simpler process of boiling it worked just as well. They would have saved lots of time, maybe converted all their friends to the new and easier method. Twenty years later, they would have gotten sick and died, in a way so causally distant from their decision to change manioc processing methods that nobody would ever have been able to link the two together.

Heinrich discusses pregnancy taboos in Fiji; pregnant women are banned from eating sharks. Sure enough, these sharks contain chemicals that can cause birth defects. The women didn't really know why they weren't eating the sharks, but when anthropologists demanded a reason, they eventually decided it was because their babies would be born with shark skin rather than human skin. As explanations go, this leaves a lot to be desired. How come you can still eat other fish? Aren't you worried your kids will have scales? Doesn't the slightest familiarity with biology prove this mechanism is garbage? But if some smart independent-minded iconoclastic Fijian girl figured any of this out, she would break the taboo and her child would have birth defects.

In giving humans reason at all, evolution took a huge risk. Surely it must have wished there was some other way, some path that made us big-brained enough to understand tradition, but not big-brained enough to question it. Maybe it searched for a mind design like that and couldn't find one. So it was left with this ticking time-bomb, this ape that was constantly going to be able to convince itself of hare-brained and probably-fatal ideas.

Here, too, culture came to the rescue. One of the most important parts of any culture – more important than the techniques for hunting seals, more important than the techniques for processing tubers – is techniques for making sure nobody ever questions tradition. Like the belief that anyone who doesn't conform is probably a witch who should be cast out lest they bring destruction upon everybody. Or the belief in a God who has commanded certain specific weird dietary restrictions, and will torture you forever if you disagree. Or the fairy tales where the prince asks a wizard for help, and the wizard says "You may have everything you wish forever, but you must never nod your head at a badger", and then one day the prince nods his head at a badger, and his whole empire collapses into dust, and the moral of the story is that you should always obey weird advice you don't understand.

There's a monster at the end of this book. Humans evolved to transmit culture with high fidelity. And one of the biggest threats to transmitting culture with high fidelity was Reason. Our ancestors lived in Epistemic Hell, where they had to constantly rely on causally opaque processes with justifications that couldn't possibly be true, and if they ever questioned them then they might die. Historically, Reason has been the villain of the human narrative, a corrosive force that tempts people away from adaptive behavior towards choices that "sounded good at the time".

Why are people so bad at reasoning? For the same reason they're so bad at letting poisonous spiders walk all over their face without freaking out. Both "skills" are really bad ideas, most of the people who tried them died in the process, so evolution removed those genes from the

population, and successful cultures stigmatized them enough to give people an internalized fear of even trying.

VI.

This book belongs alongside *Seeing Like A State* and the works of G.K. Chesterton as attempts to justify tradition, and to argue for organically-evolved institutions over top-down planning. What unique contribution does it make to this canon?

First, a lot more specifically anthropological / paleoanthropological rigor than the other two.

Second, a much crisper focus: Chesterton had only the fuzziest idea that he was writing about cultural evolution, and Scott was only a little clearer. I think Heinrich is the only one of the three to use the term, and once you hear it, it's obviously the right framing.

Third, a sense of how traditions contain the meta-tradition of defending themselves against Reason, and a sense for why this is necessary.

And fourth, maybe we're not at the point where we really want unique contributions yet. Maybe we're still at the point where we have to have this hammered in by more and more examples. The temptation is always to say "Ah, yes, a few simple things like taboos against eating poisonous plants may be relics of cultural evolution, but obviously by now we're at the point where we know which traditions are important vs. random looniness, and we can rationally stick to the important ones while throwing out the garbage." And then somebody points out to you that *actually* divination using oracle bones was one of the important traditions, and if you thought you knew better than that and tried to throw it out, your civilization would falter.

Maybe we just need to keep reading more similarly-themed books until this point really sinks in, and we get properly worried.

The Schelling Choice is "Rabbit", not "Stag"

Followup/distillation/alternate-take on Duncan Sabien's Dragon Army Retrospective and Open Problems in Group Rationality.

There's a particular failure mode I've witnessed, and fallen into myself:

I see a problem. I see, what seems to me, to be an obvious solution to the problem. If only everyone Took Action X, we could Fix Problem Z. So I start X-ing, and maybe talking about how other people should start X-ing. Action X takes some effort on my part but it's obviously worth it.

And yet... nobody does. Or not enough people do. And a few months later, here I'm still taking Action X and feeling burned and frustrated.

Or –

– the problem is that everyone is taking Action Y, which directly causes Problem Z. If only everyone would stop Y-ing, Problem Z would go away. Action Y seems obviously bad, clearly we should be on the same page about this. So I start noting to people when they're doing Action Y, and expect them to stop.

They don't stop.

So I start subtly socially punishing them for it.

They don't stop. What's more... now they seem to be punishing *me.*

I find myself getting frustrated, perhaps angry. What's going on? Are people wrong-and-bad? Do they have wrong-and-bad beliefs?

Alas. So far in my experience it hasn't been that simple.

A recap of 'Rabbit' vs 'Stag'

I'd been planning to write this post for years. Duncan Sabien went ahead and wrote it before I got around to it. But, Dragon Army Retrospective and Open Problems in Group Rationality are both lengthy posts with a lot of points, and it still seemed worth highlighting this particular failure mode in a single post.

I used to think a lot in terms of Prisoner's Dilemma, and "Cooperate"/"Defect." I'd see problems that could easily be solved if everyone just put a bit of effort in, which would benefit everyone. And people didn't put the effort in, and this felt like a frustrating, obvious coordination failure. Why do people defect so much?

Eventually Duncan shifted towards using **Stag Hunt** rather than **Prisoner's Dilemma** as the model here. If you haven't read it before, it's worth reading the description in full. If you're familiar you can skip to my current thoughts below.

My new favorite tool for modeling this is **stag hunts**, which are similar to prisoner's dilemmas in that they contain two or more people each independently making decisions which affect the group. In a stag hunt:

—Imagine a hunting party venturing out into the wilderness.

— Each player may choose *stag* or *rabbit,* representing the type of game they will try to bring down.

— All game will be shared within the group (usually evenly, though things get more complex when you start adding in real-world arguments over who deserves what).

— Bringing down a stag is costly and effortful, and requires coordination, but has a large payoff. Let's say it costs each player 5 points of utility (time, energy, bullets, etc.) to participate in a stag hunt, but a stag is worth 50 utility (in the form of food, leather, etc.) if you catch one.

— Bringing down rabbits is low-cost and low-effort and can be done unilaterally. Let's say it only costs each player 1 point of utility to hunt rabbit, and you get 3 utility as a result.

— If any player unexpectedly chooses rabbit while others choose stag, the stag escapes through the hole in the formation and is not caught. Thus, if five players all choose stag, they lose 25 utility and gain 50 utility, for a net gain of 25 (or +5 apiece). But if four players choose stag and one chooses rabbit, they lose 21 utility and gain only 3.

This creates a strong pressure toward having the Schelling choice be *rabbit.* It's saner and safer (spend 5, gain 15, net gain of 10 or +2 apiece), especially if you have any doubt about the other hunters' ability to stick to the plan, or the other hunters' faith in the other hunters, or in the other hunters' current resources and ability to even take a hit of 5 utility, or in whether or not the forest contains a stag at all.

Let's work through a specific example. Imagine that the hunting party contains the following five people:

Alexis (currently has 15 utility "in the bank")

Blake (currently has 12)

Cameron (9)

Dallas (6)

Elliott (5)

If everyone successfully coordinates to choose stag, then the end result will be positive for everyone. The stag costs everyone 5 utility to bring down, and then its 50 utility is divided evenly so that everyone gets 10, for a net gain of 5. The array [15, 12, 9, 6, 5] has bumped up to [20, 17, 14, 11, 10].

If everyone chooses rabbit, the end result is *also* positive, though less excitingly so. Rabbits cost 1 to hunt and provide 3 when caught, so the party will end up at

[17, 14, 11, 8, 7].

But imagine the situation where a stag hunt is *attempted,* but unsuccessful. Let's say that Blake quietly decides to hunt rabbit while everyone else chooses stag. What happens?

Alexis, Cameron, Dallas, and Elliott each lose 5 utility while Blake loses 1. The rabbit that Blake catches is divided five ways, for a total of 0.6 utility apiece. Now our array looks like [10.6, 11.6, 4.6, 1.6, 0.6].

(Remember, Blake only spent 1 utility in the first place.)

If you're Elliott, this is a *super scary* result to imagine. You no longer have enough resources in the bank to be self-sustaining—you can't even go out on another *rabbit* hunt, at this point.

And so, if you're Elliott, it's tempting to *preemptively choose rabbit yourself.* If there's even a *chance* that the other players might defect on the overall stag hunt (because they're tired, or lazy, or whatever) or worse, if there might not even *be* a stag out there in the woods today, then you have a *strong* motivation to self-protectively husband your resources. Even if it turns out that you were wrong about the others, and you end up being the *only* one who chose rabbit, you still end up in a much less dangerous spot: [10.6, 7.6, 4.6, 1.6, 4.6].

Now imagine that you're *Dallas,* thinking through each of these scenarios. In both cases, you end up pretty screwed, with your total utility reserves at 1.6. At that point, you've got to drop out of any future stag hunts, and all you can do is hunt rabbit for a while until you've built up your resources again.

So as Dallas, you're reluctant to listen to any enthusiastic plan to choose stag. You've got enough resources to absorb *one* failure, and so you don't want to do a stag hunt until you're *really darn sure* that there's a stag out there, and that everybody's *really actually for real* going to work together and try their hardest. You're not *opposed* to hunting stag, you're just opposed to wild optimism and wanton, frivolous burning of resources.

Meanwhile, if you're Alexis or Blake, you're starting to feel pretty frustrated. I mean, why bother coming out to a *stag hunt* if you're not even actually willing to put in the effort to *hunt stag?* Can't these people see that we're all better off if we pitch in hard, together? Why are Dallas and Elliott preemptively talking about rabbits when we haven't even *tried* catching a stag yet?

I've recently been using the terms **White Knight** and **Black Knight** to refer, not to specific people like Alexis and Elliott, but to the *roles* that those people play in situations requiring this kind of coordination. White Knight and Black Knight are hats that people put on or take off, depending on circumstances.

The White Knight is a character who has looked at what's going on, built a model of the situation, decided that they understand the Rules, and begun to take confident action in accordance with those Rules. In particular, the White Knight has decided that the time to choose stag is *obvious,* and is already common knowledge/has the Schelling nature. I mean, just look at the numbers, right?

The White Knight is often wrong, because reality is more complex than the model even if the model is a *good* model. Furthermore, other people often don't *notice*

> that the White Knight is assuming that everyone knows that it's time to choose stag—communication is hard, and the double illusion of transparency is a hell of a drug, and someone can say words like "All right, let's all get out there and do our best" and different people in the room can draw very different conclusions about what that means.
>
> So the White Knight burns resources over and over again, and feels defected on every time someone "wrongheadedly" chooses rabbit, and meanwhile the other players feel unfairly judged and found wanting according to a standard that they never explicitly agreed to (remember, choosing rabbit should be the Schelling option, according to me), and the whole thing is very rough for everyone.
>
> If this process goes on long enough, the White Knight may burn out and become the Black Knight. The Black Knight is a more *mercenary* character—it has limited resources, so it has to watch out for itself, and it's only allied with the group to the extent that the group's goals match up with its own. It's capable of teamwork and coordination, but it's not *zealous*. It isn't blinded by optimism or patriotism; it's there to engage in mutually beneficial trade, while taking into account the realities of uncertainty and unreliability and miscommunication.
>
> The Black Knight doesn't like this whole frame in which *doing the safe and conservative thing* is judged as "defection." It wants to know who this White Knight thinks he is, that he can just *declare* that it's time to choose stag, without discussion or consideration of cost. If anyone's defecting, it's the *White* Knight, by going around getting mad at people for following local incentive gradients and doing the predictable thing.
>
> But the Black Knight is *also* wrong, in that sometimes you really *do* have to be all-in for the thing to work. You can't always sit back and choose the safe, calculated option—there are, *sometimes,* gains that can only be gotten if you have no exit strategy and leave everything you've got on the field.
>
> I don't have a solution for this particular dynamic, except for a general sense that shining more light on it (dignifying both sides, improving communication, being willing to be explicit, making it *safe* for both sides to be explicit) will probably help. I think that a "technique" which zeroes in on ensuring shared common-knowledge understanding of "this is what's good in our subculture, this is what's bad, this is when we need to fully commit, this is when we can do the minimum" is a promising candidate for defusing the whole cycle of mutual accusation and defensiveness.
>
> (Circling with a capital "C" seems to be useful for coming at this problem sideways, whereas mission statements and manifestos and company handbooks seem to be partially-successful-but-high-cost methods of solving it directly.)

The key conceptual difference that I find helpful here is acknowledging that "Rabbit" / "Stag" are both *positive* choices, that bring about utility. "Defect" feels like it brings in connotations that aren't always accurate.

Saying that you're going to pay rent on time, and then not, is defecting.

But if someone shows up saying "hey let's all do Big Project X" and you're not that enthusiastic about Big Project X but you sort of nod noncommittally, and then it turns out they thought you were going to put 10 hours of work into it and you thought you

were going to put in 1, and then they get mad at you... I think it's more useful to think of this as "choosing rabbit" than "defecting."

Likewise, it's "rabbit" if you say "nah, I just don't think Big Project X is important". Going about your own projects and not signing up for every person's crusade is a perfectly valid action.

Likewise, it's "rabbit" if you say "look, I realize we're in a bad equilibrium right now and it'd be better if we all switched to A New Norm. But right now the Norm is X, and unless you are *actually sure* that we have enough buy-in for The New Norm, I'm not going to start doing a costly thing that I don't think is even going to work."

A lightweight, but concrete example

At my office, we have Philosophy Fridays*, where we try to get sync about important underlying philosophical and strategic concepts. What is our organization *for*? How does it connect to the big picture? What individual choices about particular site-features are going to bear on that big picture?

We generally agree that Philosophy Friday is important. But often, we seem to disagree a lot about the right way to go about it.

In a recent example: it often felt to me that our conversations were sort of meandering and inefficient. Meandering conversations that don't go anywhere is a stereotypical rationalist failure mode. I do it a lot by default myself. I wish that people would punish *me* when I'm steering into 'meandering mode'.

So at some point I said 'hey this seems kinda meandering.'

And it kinda meandered a bit more.

And I said, in a move designed to be somewhat socially punishing: "I don't really trust the conversation to go anywhere useful." And then I took out my laptop and mostly stopped paying attention.

And someone else on the team responded, eventually, with something like "I don't know how to fix the situation because you checked out a few minutes ago and I felt punished and wanted to respond but then you didn't give me space to."

"Hmm," I said. I don't remember exactly what happened next, but eventually he explained:

Meandering conversations were important to him, because it gave him space to actually think. I pointed to examples of meetings that I thought had gone well, that ended with google docs full of what I thought had been useful ideas and developments. And he said "those all seemed like examples of mediocre meetings to me – we had a lot of ideas, sure. But I didn't feel like I actually got to come to a real decision about anything important."

"Meandering" quality allowed a conversation to explore subtle nuances of things, to fully explore how a bunch of ideas would intersect. And this was necessary to eventually reach a firm conclusion, to leave behind the niggling doubts of "is this

really the right path for the organization?" so that he could firmly commit to a longterm strategy.

We still debate the right way to conduct Philosophy Friday at the office. But now we have a slightly better frame for that debate, and awareness of the tradeoffs involved. We discuss ways to get the good elements of the "meandering" quality while still making sure to end with clear next-actions. And we discuss alternate modes of conversation we can intelligently shift between.

There's a time when I would have pre-emptively gotten really frustrated, and started rationalizing reasons why my teammate was willfully pursuing a bad conversational norm. Fortunately I had thought enough about this sort of problem that I noticed that I was failing into a failure mode, and shifted mindsets.

Rabbit in this case was "everyone just sort of pursues whatever conversational types seem best to them in an uncoordinated fashion", and *Stag* is "we deliberately choose and enforce particular conversational norms."

We haven't yet coordinated enough to really have a "stag" option we can coordinate around. But I expect that the conversational norms we eventually settle into will be *better* than if we had naively enforced either my or my teammate's preferred norms.

Takeaways

There seem like a couple important takeaways here, to me.

One is that, yes:

Sometimes stag hunts are worth it.

I'd like people in my social network to be aware that sometimes, it's really important for everyone to adopt a new norm, or for everyone to throw themselves 100% into something, or for a whole lot of person-hours to get thrown into a project.

When discussing whether to embark on a stag hunt, it's useful to have shorthand to communicate why you might ever *want* to put a lot of effort into a concerted, coordinated effort. And then you can discuss the tradeoffs seriously.

I have more to say about what sort of stag hunts seem do-able. But for this post I want to focus primarily on the fact that...

The schelling option is Rabbit

Some communities have established particular norms favoring 'stag'. But in modern, atomic, Western society you should probably not assume this as a default. If you want people to choose stag, you need to spend special effort building common knowledge that Big Project X matters, and is worthwhile to pursue, and get everyone on board with it.

Corollary: Creating common knowledge is hard. If you *haven't* put in that work, you should assume Big Project X is going to fail, and/or that it will require a few people putting in herculean effort "above their fair share", which may not be sustainable for them.

This depends on whether effort is fungible. If you need 100 units of effort, you can make do with one person putting in 100 units of effort. If you need everyone to adopt a new norm that they haven't bought into, *it just won't work.*

If you are proposing what seems (to you) quite sensible, but nobody seems to agree...

...well, maybe people *are* being biased in some way, or motivated to avoid considering your proposed stag-hunt. People sure do seem biased about things, in general, even when they know about biases. So this may well be part of the issue.

But I think it's quite likely that you're dramatically underestimating the inferential distance – both the distance between their outlook and "why your proposed action is good", as well as the distance between your outlook and "why their current frame is weighing tradeoffs very differently than your current frame."

Much of the time, I feel like getting angry and frustrated... is something like "wasted motion" or "the wrong step in the dance."

Not entirely – anger and frustration are useful motivators. They help me notice that something about the status quo is wrong and needs fixing. But I think the specific flavor of frustration that stems from "people should be cooperating but aren't" is often, in some sense, *actually wrong* about reality. People are actually making reasonable decisions given the current landscape.

Anger and frustration help drive me to action, but often they come with a sort of tunnel vision. They lead me to dig in my heels, and get ready to fight – at a moment when what I really need is empathy and curiosity. I either need to figure out how to communicate better, to help someone understand why my plan is good. Or, I need to learn what tradeoffs I'm missing, which they can see more clearly than I.

My own strategies right now

In general, choose Rabbit.

- Keep at around 30% slack in reserve (such that I can absorb not one, not two, but *three* major surprise costs without starting to burn out). Don't spend energy helping others if I've dipped below 30% for long – focus on making sure my own needs are met.
- Find local improvements I can make that don't require much coordination from others.

Follow rabbit trails into Stag* Country

Given a choice, seek out "Rabbit" actions that preferentially build *option value* for improved coordination later on.

- Metaphorically, this means "Follow rabbit trails that lead into *Stag-and-Rabbit Country", where I'll have opportunities to say:
    - "Hey guys I see a stag! Are we all 100% up for hunting it?" and then maybe it so happens we can stag hunt together.
    - *Or,* I can sometimes say, at small-but-manageable-cost-to-myself "hey guys, I see a whole bunch of rabbits over there, you could hunt them if you want." And others can sometimes do the same for me.
- Sliiightly more concretely, this means:

- Given the opportunity, without requiring actions on the part of other people... pursue actions that demonstrate my trustworthiness, and which build bits of infrastructure that'll make it easier to work together in the future.
- Help people out if I can do so without dipping below 30% slack for too long, especially if I expect it to increase the overall slack in the system.

(I'll hopefully have more to say about this in the future.)

Get curious about other people's frames

If a person and I have argued through the same set of points multiple times, each time expecting our points to be a solid knockdown of the other's argument... and if nobody has changed their mind...

Probably we are operating in two different frames. Communicating across frames is very hard, and beyond scope of this of this post to teach. But cultivating curiosity and empathy are good first steps.

Occasionally run "Kickstarters for Stag Hunts." If people commit, hunt stag.

For example, the call-to-action in my Relationship Between the Village and Mission post (where I asked people to contact me if they were serious about improving the Village) was designed to give me information about whether it's *possible* to coordinate on a staghunt to improve the Berkeley rationality village.

Reason isn't magic

This is a linkpost for http://benjaminrosshoffman.com/reason-isnt-magic/

Here's a story some people like to tell about the limits of reason. There's this plant, manioc, that grows easily in some places and has a lot of calories in it, so it was a staple for some indigenous South Americans since before the Europeans showed up. Traditional handling of the manioc involved some elaborate time-consuming steps that had no apparent purpose, so when the Portuguese introduced it to Africa, they didn't bother with those steps - just, grow it, cook it, eat it.

The problem is that manioc's got cyanide in it, so if you eat too much too often over a lifetime, you get sick, in a way that's not easily traceable to the plant. Somehow, over probably hundreds of years, the people living in manioc's original range figured out a way to leach out the poison, without understanding the underlying chemistry - so if you asked them why they did it that way, they wouldn't necessarily have a good answer.

Now a bunch of Africans growing and eating manioc as a staple regularly get cyanide poisoning.

This is offered as a cautionary tale against innovating through reason, since there's a lot of information embedded in your culture (via hundreds of years of selection), even if people can't explain why. The problem with this argument is that it's a nonsense comparison.

First of all, it's not clear things got worse on net, just that a tradeoff was made. How many person-days per year were freed up by less labor-intensive manioc handling? Has anyone bothered to count the hours lost to laborious traditional manioc-processing, to compare them with the burden of consuming too much cyanide? How many of us, knowing that convenience foods probably lower our lifespans relative to slow foods, still eat them because they're ... more convenient?

How many people ***didn't starve*** because manioc was available and would grow where and when other things wouldn't?

If this is the best we can do for how poorly reason can perform, reason seems pretty great.

Second, we're not actually comparing reason to tradition - we're comparing *changing things* to *not changing things*. Change, as we know, is bad. Sometimes we change things anyway - when we think it's worth the price, or the risk. Sometimes, we're wrong.

Third, the actually existing Portuguese and Africans involved in this experiment weren't committed rationalists - they were just people trying to get by. It probably doesn't take more than a day's reasoning to figure out which steps in growing manioc are really necessary to get the calories palatably. Are we imagining that someone making a concerted effort to improve their life through reason would just stop there?

This is being compared with *many generations* of trial and error. Is that the standard we want to use? Reasoning isn't worth it unless a day of untrained thinking can outperform hundreds of years of accumulated tradition?

It gets worse. This isn't a randomly selected example - it's specifically selected as a case where reason would have a hard time noticing when and how it's making things worse. In this particular case, reason introduced an important problem. But life is full of risks, sometimes in ways that are worse for traditional cultures. Do we really want to say that reasoning isn't the better bet unless it outperforms literally every time, without ever making things locally worse? Even theoretically perfect Bayesian rationality will sometimes recommend changes that have an expected benefit, but turn out to be harmful. Not even tradition meets this standard! Only logical certainties do - provided, that is, we haven't made an error in one of our proofs.

We also have to count all the deaths and other problems averted by reasoning about a problem. Reasoning introduces risks - but also, risks come up even when we're not reasoning about them, just from people doing things that affect their environments. There's absolutely no reason to think that the sort of gradual iteration that accretes into tradition never enters a bad positive feedback loop. Even if you think modernity is an exceptional case of that kind of bad feedback loop, we had to have gotten there via the accretion of premodern tradition and iteration!

The only way out is through. But why did we have this exaggerated idea of what reason could do, in the first place?

Writing children's picture books

This is a linkpost for https://unstableontology.com/2019/06/25/writing-childrens-picture-books/

[the text of the post is pasted here, for redundancy]

Here's an exercise for explaining and refining your opinions about some domain, X:

> Imagine writing a 10-20 page children's picture book about topic X. Be fully honest and don't hide things (assume the child can handle being told the truth, including being told non-standard or controversial facts).

Here's a dialogue, meant to illustrate how this could work:

A: What do you think about global warming?

B: Uhh…. I don't know, it seems real?

A: How would you write a 10-20 page children's picture book about global warming?

B: Oh, I'd have a diagram showing carbon dioxide exiting factories and cars, floating up in the atmosphere, and staying there. Then I'd have a picture of sunlight coming through the atmosphere, bounding off the earth, then going back up, but getting blocked by the carbon dioxide, so it goes back to the earth and warms up the earth a second time. Oh, wait, if the carbon dioxide prevents the sunlight from bouncing from the earth to the sky, wouldn't it also prevent the sunlight from entering the atmosphere in the first place? Oh, I should look that up later [NOTE: the answer is that CO2 blocks thermal radiation much more than it blocks sunlight].

Anyway, after that I'd have some diagrams showing global average temperature versus global CO2 level that show how the average temperature is tracking CO2 concentration, with some lag time. Then I'd have some quotes about scientists and information about the results of surveys. I'd show a graph showing how much the temperature would increase under different conditions… I think I've heard that, with substantial mitigation effort, the temperature difference might be 2 degrees Celsius from now until the end of the century [NOTE: it's actually 2 degrees from pre-industrial times till the end of the century, which is about 1 degree from now]. And I'd want to show what 2 degrees Celsius means, in terms of, say, a fraction of the difference between winter and summer.

I'd also want to explain the issue of sea level rise, by showing a diagram of a glacier melting. Ice floats, so if the glacier is free-floating, then it melting doesn't cause a sea level rise (there's some scientific principle that says this, I don't remember what it's called), but if the glacier is on land, then when it melts, it causes the sea level to rise. I'd also want to show a map of the areas that would get flooded. I think some locations, like much of Florida, get flooded, so the map should show that, and there should also be a pie chart showing how much of the current population would end up underwater if they didn't move (my current guess is that it's between 1 percent and 10 percent, but I could be pretty wrong about this [NOTE: the answer is 30 to 80 million people, which is between about 0.4% and 1.1%]).

I'd also want to talk about possible mitigation efforts. Obviously, it's possible to reduce energy consumption (and also meat consumption, because cows produce methane which is also a greenhouse gas). So I'd want to show a chart of which things produce the most greenhouse gases (I think airplane flights and beef are especially bad), and showing the relationship between possible reductions in that and the temperature change.

Also, trees take CO2 out of the atmosphere, so preserving forests is a way to prevent global warming. I'm confused about where the CO2 goes, exactly, since there's some cycle it goes through in the forest; does it end up underground? I'd have to look this up.

I'd also want to talk about the political issues, especially the disinformation in the space. There's a dynamic where companies that pollute want to deny that man-made global warming is a real, serious problem, so there won't be regulations. So, they put out disinformation on television, and they lobby politicians. Sometimes, in the discourse, people go from saying that global warming isn't real, to saying it's real but not man-made, to saying it's real and man-made but it's too late to do anything about it. That's a clear example of motivated cognition. I'd want to explain how this is trying to deny that any changes should be made, and speculate about why people might want to, such as because they don't trust the process that causes changes (such as the government) to do the right thing.

And I'd also want to talk about geoengineering. There are a few proposals I know of. One is to put some kind of sulfer-related chemical in the atmosphere, to block out sunlight. This doesn't solve ocean acidification, but it does reduce the temperature. But, it's risky, because if you stop putting the chemical in the atmosphere, then that causes a huge temperature swing.

I also know it's possible to put iron in the ocean, which causes a plankton bloom, which… does something to capture CO2 and store it in the bottom of the ocean? I'm really not sure how this works, I'd want to look it up before writing this section.

There's also the proposal of growing and burning trees, and capturing and storing the carbon. When I looked this up before, I saw that this takes quite a lot of land, and anyway there's a lot of labor involved, but maybe some if it can be automated.

There are also political issues with geoengineering. There are people who don't trust the process of doing geoengineering to make things better instead of worse, because they expect that people's attempts to reason about it will make lots of mistakes (or people will have motivated cognition and deceive themselves and each other), and then the resulting technical models will make things that don't work. But, the geoengineering proposals don't seem harder than things that humans have done in the past using technical knowledge, like rockets, so I don't agree that this is such a big problem.

Furthermore, some people want to shut down discussion of geoengineering, because such discussion would make it harder to morally pressure people into reducing carbon emissions. I don't know how to see this as anything other than an adversarial action against reasonable discourse, but I'm sure there is some motivation at play here. Perhaps it's a motivation to have everyone come together as one, all helping together, in a hippie-ish way. I'm not sure if I'm right here, I'd want to read something written by one of these people before making any strong judgments.

Anyway, that's how I'd write a picture book about global warming.

So, I just wrote that dialogue right now, without doing any additional research. It turns out that I do have quite a lot of opinions about global warming, and am also importantly uncertain in some places, some of which I just now became aware of. But I'm not likely to produce these opinions if asked "what do you think about global warming?"

Why does this technique work? I think it's because, if asked for one's opinions in front of an adult audience, it's assumed that there is a background understanding of the issue, and you have to say something new, and what you decide to say says something about you. Whereas, if you're explaining to a child, then you know they lack most of the background understanding, and so it's obviously good to explain that.

With adults, it's assumed there are things that people act like "everyone knows", where it might be considered annoying to restate them, since it's kind of like talking down to them. Whereas, the illusion or reality that "everyone knows" is broken when explaining to children.

The countervailing force is that people are tempted to lie to children. Of course, it's necessary to not lie to children to do the exercise right, and also to raise or help raise children who don't end up in an illusory world of confusion and dread. I would hope that someone who has tendencies to hide things from children would at least be able to notice and confront these tendencies in the process of imagining writing children's picture books.

I think this technique can be turned into a generalized process for making world models. If someone wrote a new sketch of a children's picture book (about a new topic) every day, and did the relevant research when they got stuck somewhere, wouldn't they end up with a good understanding of both the world and of their own models of the world after a year? It's also a great starting point from which to compare your opinions to others' opinions, or to figure out how to explain things to either children or adults.

Anyway, I haven't done this exercise for very many topics yet, but I plan on writing more of these.

Arbital scrape

Update: Arbital Scrape V2

I've scraped http://arbital.com as the site is unusably slow and hard to search for me.

The scrape is locally browsable and plain HTML save for MathJax.

https://drive.google.com/open?id=1b7dKhOzfMpFwngAeI8efeOzv147Lv_mx

https://emma-borhanian.github.io/arbital-scrape/

If there's interest let me know as I may tidy up and open source my code. edit: working on this

Alternating group (metadata) (https://arbital.com/p/alternating_group)

The alternating group is the only normal subgroup of the symmetric group (on five or more generators).

back to index
Explore > Mathematics > Group theory > Group > Alternating group
Explore > Mathematics > Abstract algebra > Algebraic structure > Group > Alternating group

---

The *alternating group* A_n is defined as a certain subgroup of the Symmetric group S_n: namely, the collection of all elements which can be made by multiplying together an even number of transpositions. This is a well-defined notion (proof).

%%%knows-requisite(Normal subgroup): A_n is a Normal subgroup of S_n; it is the quotient of S_n by the sign homomorphism. %%%

Examples

- A cycle of even length is an odd permutation in the sense that it can only be made by multiplying an odd number of transpositions. For example, (132) is equal to $(13)(23)$.
- A cycle of odd length is an even permutation, in that it can only be made by multiplying an even number of transpositions. For example, (1354) is equal to $(54)(34)(14)$.
- The alternating group A_4 consists precisely of twelve elements: the identity, $(12)(34)$, $(13)(24)$, $(14)(23)$, (123), (124), (134), (234), (132), (143), (142), (243).

Properties

%%%knows-requisite(Normal subgroup): The alternating group A_n is of [index_of_a_subgroup index] 2 in S_n. Therefore A_n is normal in S_n (proof). Alternatively we may give the homomorphism explicitly of which A_n is the kernel: it is the sign homomorphism. %%%

- A_n is generated by its 3-cycles. (Proof.)
- A_n is simple. (Proof.)
- The conjugacy classes of A_n are easily characterised.

Creators

- Patrick Stevens "Automatically generated group for Patrick Stevens"

Children

- Alternating group is generated by its three-cycles "A useful result which lets us prove things about the alternating group more easily." - Patrick Stevens
- Conjugacy classes of the alternating group on five elements "A_5 has easily-characterised conjugacy classes, based on a rather surprising theorem about when conjugacy classes in the symmetric group split." - Patrick Stevens
  - Conjugacy classes of the alternating group on five elements: Simpler proof "A listing of the conjugacy classes of the alternating group on five letters, without using heavy theory." - Patrick Stevens
- Splitting conjugacy classes in alternating group "The conjugacy classes in the alternating group are usually the same as those in the symmetric group; there is a surprisingly simple condition for when this does not hold." - Patrick Stevens
- The alternating group on five elements is simple "The smallest (nontrivial) simple group is the alternating group on five elements." - Patrick Stevens
  - The alternating group on five elements is simple: Simpler proof "A proof which avoids some of the heavy machinery of the main proof." - Patrick Stevens
- The alternating groups on more than four letters are simple "The alternating groups are the most accessible examples of simple groups, and this fact also tells us that the symmetric groups are "complicated" in some sense." - Patrick Stevens
- The collection of even-signed permutations is a group "This proves the well-definedness of one particular definition of the alternating group." - Patrick Stevens

Mirror: www.obormot.net/arbital

Selection vs Control

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

This is something which has bothered me for a while, but, I'm writing it specifically in response to the recent post on mesa-optimizers.

I feel strongly that the notion of 'optimization process' or 'optimizer' which people use -- partly derived from Eliezer's notion in the sequences -- should be split into two clusters. I call these two clusters 'selection' vs 'control'. I don't have precise formal statements of the distinction I'm pointing at; I'll give several examples.

Before going into it, several reasons why this sort of thing may be important:

- It could help refine the discussion of mesa-optimization. The article restricted its discussion to the type of optimization I'll call 'selection', explicitly ruling out 'control'. This choice isn't obviously right. (More on this later.)
- Refining 'agency-like' concepts like this seems important for embedded agency -- what we eventually want is a story about how agents can be in the world. I think almost any discussion of the relationship between agency and optimization which isn't aware of the distinction I'm drawing here (at least as a hypothesis) will be confused.
- Generally, I feel like I see people making mistakes by not distinguishing between the two (whether or not they've derived their notion of optimizer from Eliezer). I judge an algorithm differently if it is intended as one or the other.

(See also Stuart Armstrong's summary of other problems with the notion of optimization power Eliezer proposed -- those are unrelated to my discussion here, and strike me more as technical issues which call for refined formulae, rather than conceptual problems which call for revised ontology.)

The Basic Idea

Eliezer quantified optimization power by asking how small a target an optimization process hits, out of a space of possibilities. The type of 'space of possibilities' is what I want to poke at here.

Selection

First, consider a typical optimization algorithm, such as simulated annealing. The algorithm constructs an element of the search space (such as a specific combination of weights for a neural network), gets feedback on how good that element is, and then tries again. Over many iterations of this process, it finds better and better elements. Eventually, it outputs a single choice.

This is the prototypical 'selection process' -- it can *directly instantiate any element of the search space* (although typically we consider cases where the process doesn't have time to instantiate *all* of them), it *gets direct feedback on the quality of each*

element (although evaluation may be costly, so that the selection process must economize these evaluations), *the quality of an element of search space does not depend on the previous choices,* and *only the final output matters.*

The term 'selection process' refers to the fact that this type of optimization selects between a number of explicitly given possibilities. The most basic example of this phenomenon is a 'filter' which rejects some elements and accepts others -- like selection bias in statistics. This has a limited ability to optimize, however, because it allows only one iteration. Natural selection is an example of much more powerful optimization occurring through iteration of selection effects.

Control

Now, consider a targeting system on a rocket -- let's say, a heat-seeking missile. The missile has sensors and actuators. It gets feedback from its sensors, and must somehow use this information to decide how to use its actuators. This is my prototypical control process. (The term 'control process' is supposed to invoke control theory.) Unlike a selection process, *a controller can only instantiate one element of the space of possibilities*. It gets to traverse exactly one path. The 'small target' which it hits is therefore 'small' with respect to a space of ***counterfactual*** *possibilities,* with all the technical problems of evaluating counterfactuals. *We only get full feedback on one outcome* (although we usually consider cases where the partial feedback we get along the way gives us a lot of information about how to navigate toward better outcomes). *Every decision we make along the way matters, both in terms of influencing total utility, and in terms of influencing what possibilities we have access to in subsequent decisions.*

So: in evaluating the optimization power of a selection process, we have a fairly objective situation on our hands: the space of possibilities is explicitly given; the utility function is explicitly given; we can compare the true output of the system to a randomly chosen element. In evaluating the optimization power of a control process, we have a very subjective situation on our hands: the controller only truly takes one path, so any judgement about a space of possibilities requires us to define counterfactuals; it is less clear how to define an un-optimized baseline; utility need not be explicitly represented in the controller, so may have to be inferred (or we think of it as parameter, so, we can measure optimization power with respect to different utility functions, but there's no 'correct' one to measure).

I do think both of these concepts are meaningful. I don't want to restrict 'optimization' to refer to only one or the other, as the mesa-optimization essay does. However, I think the two concepts are of a very different type.

Bottlecaps & Thermostats

The mesa-optimizer write-up made the decision to focus on what I call selection processes, excluding control processes:

> We will say that a system is an *optimizer* if it is internally searching through a search space (consisting of possible outputs, policies, plans, strategies, or similar) looking for those elements that score high according to some objective function that is explicitly represented within the system. *[...]* For example, a bottle cap

causes water to be held inside the bottle, but it is not optimizing for that outcome since it is not running any sort of optimization algorithm.(1) Rather, bottle caps have been *optimized* to keep water in place.

It makes sense to say that we aren't worried about bottlecaps when we think about the inner alignment problem. However, this also excludes much more powerful 'optimizers' -- something more like a plant.

When does a powerful control process become an 'agent'?

- **Bottlecaps:** No meaningful actuators or sensors. Essentially inanimate. Does a particular job, possibly very well, but in a very predictable manner.
- **Thermostats:** Implements a negative feedback loop via a sensor, an actuator, and a policy of "correcting" things when sense-data indicates they are "off". Actual thermostats explicitly represent the target temperature, but one can imagine things in this cluster which wouldn't -- in general, the connection between what is sensed and how things are 'corrected' can be quite complex (involving many different sensors and actuators), so that no one place in the system explicitly represents the 'target'.
- **Plants:** Plants are like very complex thermostats. They have no apparent 'target' explicitly represented, but can clearly be thought of as relatively agentic, achieving complicated goals in complicated environments.
- **Guided Missiles:** These are also mostly in the 'thermostat' category, but, guided missiles can use simple world-models (to track the location of the target). However, any 'planning' is likely based on explicit formulae rather than any search. (I'm not sure about actual guided missiles.) If so, a guided missile would still not be a selection process, and therefore lack a "goal" in the mesa-optimizer sense, despite having a world-model and explicitly reasoning about how to achieve an objective represented within that world-model.
- **Chess Programs:** A chess-playing program has to play each game well, and every move is significant to this goal. So, it is a control process. However, AI chess algorithms are based on explicit search. Many, many moves are considered, and each move is evaluated independently. This is a common pattern. The best way we know how to implement very powerful controllers *is* to use search inside (implementing a control process *using* a selection process). At that point, a controller seems clearly 'agent-like', and falls within the definition of optimizer used in the meso-optimization post. However, it seems to me that things become 'agent-like' somewhere before this stage.

(See also: adaptation-executers, not fitness maximizers.)

I don't want to frame it as if there's "one true distinction" which we should be making, which I'm claiming the mesa-optimization write-up got wrong. Rather, we should pay attention to the different distinctions we might make, studying the phenomena separately and considering the alignment/safety implications of each.

This is closely related to the discussion of upstream daemons vs downstream daemons. A downstream-daemon seems more likely to be an optimizer in the sense of the mesa-optimization write-up; it is explicitly planning, which may involve search. These are more likely to raise concerns through explicitly reasoned out treacherous turns. An upstream-daemon *could* use explicit planning, but it could also be only a bottlecap/thermostat/plant. It might powerfully optimize for something in the controller sense without internally using selection. This might produce severe misalignment, but not through explicitly planned treacherous turns. (*Caveat*: we don't

understand mesa-optimizers; an understanding sufficient to make statements such as these with confidence would be a significant step forward.)

It seems possible that one could invent a measure of "control power" which would rate highly-optimized-but-inanimate objects like bottlecaps very low, while giving a high score to thermostat-like objects which set up complicated negative feedback loops (even if they didn't use any search).

Processes Within Processes

I already mentioned the idea that the best way we know how to implement powerful control processes is through powerful selection (search) *inside* of the controller.

To elaborate a bit on that: a controller with a search inside would typically have some kind of model of the environment, which it uses by searching for good actions/plans/policies for achieving its goals. So, measuring the optimization power *as a controller*, we look at how successful it is at achieving its goals in the real environment. Measuring the optimization power *as a selector*, we look at how good it is at choosing high-value options *within its world-model.* The search can only do as well as its model can tell it; however, in some sense, the agent is ultimately judged by the true consequences of its actions.

IE, in this case, the selection vs control distinction is a map/territory distinction. I think this is part of why I get so annoyed at things which mix up selection and control: it looks like a map/territory error to me.

However, this is not the only way selection and control commonly relate to each other.

Effective controllers are very often designed through a search process. This might be search taking place within a model, again (for example, training a neural network to control a robot, but getting its gradients from a physics simulation so that you can generate a large number of training samples relatively cheaply) or the real world (evolution by natural selection, "evaluating" genetic code by seeing what survives).

Further complicating things, a powerful search algorithm generally has some "smarts" to it, ie, it is good at choosing what option to evaluate next based on the current state of things. This "smarts" is controller-style smarts: every choice matters (because every evaluation costs processing power), there's no back-tracking, and you have to hit a narrow target in one shot. (Whatever the target of the underlying search problem, the target of the search-controller is: *find that target,* ***quickly****.*) And, of course, it is possible that such a search-controller will even use a model of the fitness landscape, and plan its next choice via its own search!

(I'm not making this up as a weird hypothetical; actual algorithms such as estimation-of-distribution algorithms will make models of the fitness landscape. For obvious reasons, searching for good points in such models is usually avoided; however, in cases where evaluation of points is expensive enough, it may be worth it to explicitly plan out test-points which will reveal the most information about the fitness landscape, so that the best point can be selected later.)

Blurring the Lines: What's the Critical Distinction?

I mentioned earlier that this dichotomy seems more like a conceptual cluster than a fully formal distinction. I mentioned a number of big differences which stick out at me. Let's consider some of these in more detail.

Perfect Feedback

The classical sort of search algorithm I described as my central example of a selection process includes the ability to get a perfect evaluation of any option. The difficulty arises only from the very large number of options available. Control processes, on the other hand, appear to have very bad feedback, since you can't know the full outcome until it is too late to do anything about it. Can we use this as our definition?

I would agree that a search process in which the cost of evaluation goes to infinity becomes purely a control process: you can't perform any filtering of possibilities based on evaluation, so, you have to output one possibility and try to make it a good one (with no guarantees). Maybe you get some information about the objective function (like its source code), and you have to try to use that to choose an option. That's your sensors and actuators. They have to be very clever to achieve very good outcomes. The cheaper it is to evaluate the objective function on examples, the less "control" you need (the more you can just do brute-force search). In the opposite extreme, evaluating options is so cheap that you can check all of them, and output the maximum directly.

While this is somewhat appealing, it doesn't capture every case. Search algorithms today (such as stochastic gradient descent) often have imperfect feedback. Game-tree search deals with an objective function which is much too costly to evaluate directly (the quality of a move), but can be optimized for nonetheless by recursively searching for good moves in subgames down the game tree (mixed with approximate evaluations such as rollouts or heuristic board evaluations). I still think of both of these as solidly on the "selection process" side of things.

On the control process side, it is possible to have perfect feedback without doing any search. Thermostats realistically have noisy information about the temperature of a room, but, you can imagine a case where they get perfect information. It isn't any less a controller, or more a selection process, for that fact.

Choices Don't Change Later Choices

Another feature I mentioned was that in selection processes, all options are available to try at any time, and what you look at now does not change how good any option will be later. On the other hand, in a control process, previous choices can totally change how good particular later choices would be (as in reinforcement learning), or change what options are even available (as in game playing).

First, let me set two complications aside.

- Weird decision theory cases: it is theoretically possible to screw with a search by giving it an objective function which depends on its choices during search. This doesn't seem that interesting for our purposes here. (And that's coming from *me...*)
- Local search limits the "options" to small modifications of the option just considered. I don't think this is blurring the lines between search and control; rather, it is more like using a controller within a smart search to try to increase efficiency, as I discussed at the end of the processes-within-processes section. All the options are still "available" at all times; the search algorithm just happens to be one which limits itself to considering a smaller list.

I do think some cases blur the lines here, though. My primary example is the multi-armed bandit problem. This is a special case of the RL problem in which the history doesn't matter; every option is equally good every time, except for some random noise. Yet, to me, it is still a control problem. Why? Because every decision matters. The feedback you get about how good a particular choice was isn't just thought of as *information*; you "actually get" the good/bad outcome each time. That's the essential character of the multi-armed bandit problem: you have to trade off between experimentally trying options you're uncertain about vs sticking with the options which seem best so far, because every selection carries weight.

This leads me to the next proposed definition.

Offline vs Online

Selection processes are like offline algorithms, whereas control processes are like online algorithms.

With offline algorithms, you only really care about the end results. You are OK running gradient descent for millions of iterations before it starts doing anything cool, so long as it eventually does something cool.

With online algorithms, you care about each outcome individually. You would probably not want to be gradient-descent-training a neural network in live user-servicing code on a website, because live code has to be acceptably good from the start. Even if you can initialize the neural network to something acceptably good, you'd hesitate to run stochastic gradient descent on it live, because stochastic gradient descent can sometimes dramatically decrease performance for a while before improving performance again.

Furthermore, online algorithms have to deal with non-stationarity. This seems suitably like a control issue.

So, selection processes are "offline optimization", whereas control processes are "online optimization": optimizing things "as they progress" rather than statically. (Note that the notion of "online optimization" implied by this line of thinking is slightly different from the common definition of online optimization, though related.)

The offline vs online distinction also has a lot to do with the sorts of mistakes I think people are making when they confuse selection processes and control processes. Reinforcement learning, as a subfield of AI, was obviously motivated from a highly online perspective. However, it is very often used as an offline algorithm today, to *produce* effective agents, rather than *as* an effective agent. So, that there's been

some mismatch between the motivations which shaped the paradigm and actual use. This perspective made it less surprising when black-box optimization beat reinforcement learning on some problems (see also).

This seems like the best definition so far. However, I personally still feel like it is still missing something important. Selection vs control feels to me like a *type* distinction, closer to map-vs-territory.

To give an explicit counterexample: evolution by natural selection is obviously a selection process according to the distinction as I make it, but it seems much more like an online algorithm than on offline one, if we try to judge it as such.

Internal Features vs Context

Returning to the definition in mesa-optimizers (emphasis mine):

> Whether a system is an optimizer is a property of its ***internal structure***—what algorithm it is physically implementing—and not a property of its input-output behavior. Importantly, the fact that a system's behavior results in some objective being maximized does not make the system an optimizer.

The notion of a selection process says a lot about what is actually happening inside a selection process: there is a space of options, which can be enumerated; it is trying them; there is some kind of evaluation; etc.

The notion of control process, on the other hand, is more *externally* defined. It doesn't matter what's going on inside of the controller. All that matters is how effective it is at what it does.

A selection process -- such as a neural network learning algorithm -- *can* be regarded "from outside", asking questions about how the *one* output of the algorithm does in the *true* environment. In fact, this kind of thinking is what we do when we think about generalization error.

Similarly, we *can* analyze a control process "from inside", trying to find the pieces which correspond to beliefs, goals, plans, and so on (or postulate what they would look like if they existed -- as must be done in the case of controllers which truly lack such moving parts). This is the decision-theoretic view.

However, one might argue that viewing selection processes from the outside is viewing them *as control* -- viewing them as essentially having one shot at overall decision quality. Similarly, viewing control process from inside is essentially viewing it *as selection* -- the decision-theoretic view gives us a version of a control problem which we can solve by mathematical optimization.

In this view, selection vs control doesn't really cluster *different types of object*, but rather, *different types of analysis*. To a large extent, we can cluster objects by what kind of analysis we would more often want to do. However, certain cases (such as a game-playing AI) are best viewed through both lenses (as a controller, in the context of doing well in a real game against a human, and as a selection process, when thinking about the game-tree search).

Overall, I think I'm probably still somewhat confused about the whole selection vs control issue, particularly as it pertains to the question of how decision theory can

apply to things in the world.

Reneging prosocially by Duncan Sabien

This is a linkpost for https://medium.com/@ThingMaker/reneging-prosocially-5b44bdec3bb9

A good post about reneging on agreements, acting when the other person reneges on you, and making agreements.

The Hacker Learns to Trust

This is a linkpost for https://medium.com/@NPCollapse/the-hacker-learns-to-trust-62f3c1490f51

This is a linkpost for some interesting discussions of info security norms in AI. I threw the post below together in 2 hours, just to have a bunch of quotes and links for people, and to have the context in one place for a discussion here on LW (makes it easier for common knowledge of what the commenters have and haven't seen). I didn't want to assume people follow any news on LW, so for folks who've read a lot about GPT-2 much of the post is skimmable.

Background on GPT-2

In February, OpenAI wrote a blogpost announcing GPT 2:

> We've trained a large-scale unsupervised language model which generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

This has been a very important release, not least due to it allowing fans to try (and fail) to write better endings to Game of Thrones. Gwern used GPT-2 to write poetry and anime. There have been many Medium posts on GPT-2, some very popular, and at least one Medium post on GPT-2 written by GPT-2. There is a subreddit where all users are copies of GPT-2, and they imitate other subreddits. It got too meta when the subreddit imitated another subreddit about people play-acting robots-pretending-to-be-humans. Stephen Woods has lots of examples including food recipes.

Here in our rationality community, we created user GPT-2 trained on the entire corpus of LessWrong comments and posts and released it onto the comment section on April 1st (a user who we warned and then banned). And Nostalgebraist created a tumblr trained on the entire writings of Eliezer Yudkowsky (sequences+HPMOR), where Nostalgebraist picked their favourites to include on the Tumblr.

There was also very interesting analysis on LessWrong and throughout the community. The post that made me think most on this subject is Sarah Constantin's Human's Who Are Not Concentrating Are Not General Intelligences. Also see SlateStarCodex's Do Neural Nets Dream of Electric Hobbits? and GPT-2 As Step Toward General Intelligence, plus my teammate jimrandomh's Two Small Experiments on GPT-2.

However, these were all using a nerfed version of GPT-2, which only had 175 million parameters, rather than the fully trained model with 1.5 billion parameters. (If you want to see examples of the full model, see the initial announcement posts for examples with unicorns and more.)

Reasoning for only releasing a nerfed GPT-2 and response

OpenAI writes:

> Due to our concerns about malicious applications of the technology, we are not releasing the trained model.

While the post includes some discussion of how specifically GPT-2 could be used maliciously (e.g. automating false clickbait news, automated spam, fake accounts) the key line is here.

> This decision, as well as our discussion of it, is an experiment: while we are not sure that it is the right decision today, we believe that the AI community will eventually need to tackle the issue of publication norms in a thoughtful way in certain research areas.

Is this out of character for OpenAI - a surprise decision? Not really.

> Nearly a year ago we wrote in the OpenAI Charter: “we expect that safety and security concerns will reduce our traditional publishing in the future, while increasing the importance of sharing safety, policy, and standards research,” and we see this current work as potentially representing the early beginnings of such concerns, which we expect may grow over time.
>
> Other disciplines such as biotechnology and cybersecurity have long had active debates about responsible publication in cases with clear misuse potential, and we hope that our experiment will serve as a case study for more nuanced discussions of model and code release decisions in the AI community.

Public response to decision

There has been discussion in news+Twitter, see here for an overview of what some people in the field/industry have said, and what the news media has written. The main response that's been selected for by news+twitter is that OpenAI did this primarily as a publicity stunt.

For a source with a different bias than the news and Twitter (which selects heavily for anger and calling out of norm violation), I've searched through all Medium articles on GPT-2 and copied here any 'most highlighted comments'. Most posts actually didn't have any, which I think means they haven't had many viewers. Here are the three I found, in chronological order.

OpenAIs GPT-2: The Model, The Hype, and the Controvery

> As ML researchers, we are building things that affect people. Sooner or later, we’ll cross a line where our research can be used maliciously to do bad things. Should we just wait until that happens to decide how we handle research that can have negative side effects?

OpenAI GPT-2: Understanding Language Generation through Visualization

> Soon, these deepfakes will become personal. So when your mom calls and says she needs $500 wired to the Cayman Islands, ask yourself: Is this really my mom, or is it a language-generating AI that acquired a voice skin of my mother from that Facebook video she posted 5 years ago?

GPT-2, Counting Consciousness and the Curious Hacker

> *If we have a system charged with detecting what we can and can't trust, we aren't removing our need to invest trust, we are only moving our trust from our own faculties to those of the machine.*

I wrote this linkpost to discuss the last one. See below.

Can someone else just build another GPT-2 and release the full 1.5B parameter model?

From the initial OpenAI announcement:

> We are aware that some researchers have the technical capacity to reproduce and open source our results. We believe our release strategy limits the initial set of organizations who may choose to do this, and gives the AI community more time to have a discussion about the implications of such systems.

Since the release, one researcher has tried to reproduce and publish OpenAI's result. Google has a program called TensorFlow Research Cloud that gives loads of free compute to researchers affiliated with various universities, which let someone train an attempted copy of GPT-2 with 1.5 billion parameters. They say:

> I really can't express how grateful I am to Google and the TFRC team for their support in enabling this. They were incredibly gracious and open to allowing me access, without requiring any kind of rigorous, formal qualifications, applications or similar. I can really only hope they are happy with what I've made of what they gave me.
>
> ...I estimate I spent around 200 hours working on this project.... I ended up spending around 600–800€ on cloud resources for creating the dataset, testing the code and running the experiments

That said, it turned out that the copy did not match up in skill level, and is weaker even than nerfed model OpenAI released. The person who built it says (1) they think they know how to fix it and (2) releasing it as-is may still be a helpful "shortcut" for others interested in building a GPT-2-level system; I don't have the technical knowledge to assess these claims, and am interested to hear from others who do.

During the period where people didn't know that the attempted copy was not successful, the person who made the copy wrote a long and interesting post explaining their decision to release the copy (with multiple links to LW posts). It discussed reasons why this specific technology may cause us to better grapple with misinformation on the internet that we hear. The author is someone who had a strong object level disagreement with the policy people at OpenAI, and had thought pretty carefully about it. However, it opened thus:

> *Disclaimer: I would like it to be made very clear that I am absolutely 100% open to the idea that I am wrong about anything in this post. I don't only accept but explicitly request arguments that could convince me I am wrong on any of these issues. If you think I am wrong about anything here, and have an argument that might convince me,* ***please*** *get in touch and present your argument. I am happy to say "oops" and retract any opinions presented here and change my course of action.*

> *As the saying goes: "When the facts change, I change my mind. What do you do?"*
>
> *TL;DR: I'm a student that replicated OpenAI's GPT2–1.5B. I plan on releasing it on the 1st of July. Before criticizing my decision to do so, please read my arguments below. If you still think I'm wrong, contact me on Twitter @NPCollapse or by email (thecurioushacker@outlook.com) and convince me. For code and technical details, see this post.*

And they later said

> [B]e assured, I read every single comment, email and message I received, even if I wasn't able to respond to all of them.

On reading the initial I was genuinely delighted to see such pro-social and cooperative behaviour from the person who believed OpenAI was wrong. They considered unilaterally overturning OpenAI's decision but instead chose to spend 11,000 words explaining their views and a month reading others' comments and talking to people. This, I thought, is how one avoids falling prey to Bostrom's unilateralist curse.

Their next post The Hacker Learns to Trust was released 6 days later, where they decided not to release the model. Note that they did not substantially change their opinions on the object level decision.

> I was presented with many arguments that have made me reevaluate and weaken my beliefs in some of the arguments I presented in my last essay. There were also many, maybe even a majority of, people in full support of me. Overall I still stand by most of what I said.
>
> ...I got to talk to Jack Clark, Alec Radford and Jeff Wu from OpenAI. We had a nice hour long discussion, where I explained where I was coming from, and they helped me to refine my beliefs. They didn't come in accusing me in any way, they were very clear in saying they wanted to help me gain more important insight into the wider situation. For this open and respectful attitude I will always be grateful. Large entities like OpenAI often seem like behemoths to outsiders, but it was during this chat that it really hit me that they were people just like me, and curious hackers to boot as well.
>
> I quickly began to understand nuances of the situation I wasn't aware of. OpenAI had a lot more internal discussion than their blog post made it seem. And I found this reassuring. Jack in particular also gave me a lot of valuable information about the possible dangers of the model, and a bit of insight into the workings of governments and intelligence agencies.
>
> After our discussion, I had a lot to think about. But I still wasn't really convinced to not release.

They then talked with Buck from MIRI (author of this great post). Talking with Buck lead them to their new view.

> [T]his isn't just about GPT2. What matters is that at some point in the future, someone *will* create something truly dangerous and there need to be commonly accepted safety norms *before* that happens.

> We tend to live in an ever accelerating world. Both the industrial and academic R&D cycles have grown only faster over the decades. Everyone wants “the next big thing” as fast as possible. And with the way our culture is now, it can be hard to resist the pressures to adapt to this accelerating pace. Your career can depend on being the first to publish a result, as can your market share.
>
> We as a community and society need to combat this trend, and create a healthy cultural environment that allows researchers to *take their time*. They shouldn’t have to fear repercussions or ridicule for delaying release. Postponing a release because of added evaluation should be the norm rather than the exception. We need to make it commonly accepted that we as a community respect others’ safety concerns and don’t penalize them for having such concerns, *even if they ultimately turn out to be wrong*. If we don’t do this, it will be a race to the bottom in terms of safety precautions.
>
> We as a community of researchers and humans need to trust one another and respect when one of us has safety concerns. We need to extend understanding and offer help, rather than get caught in a race to the bottom. And this isn’t easy, because we’re curious hackers. Doing cool things fast is what we do.

The person also came to believe that the AI (and AI safety) community was much more helpful and cooperative than they'd expected.

> The people at OpenAI and the wider AI community have been incredibly helpful, open and thoughtful in their responses to me. I owe to them everything I have learned. OpenAI reached out to me almost immediately to talk and they were nothing but respectful and understanding. The same applies to Buck Shlegeris from MIRI and many other thoughtful and open people, and I am truly thankful for their help.
>
> I expected a hostile world of skepticism and competition, and there was some of that to be sure. But overall, the AI community was open in ways I did not anticipate. In my mind, I couldn’t imagine people from OpenAI, or MIRI, or anywhere else actually wanting to talk to me. But I found that was wrong.
>
> So this is the first lesson: The world of AI is full of smart, good natured and open people that I shouldn’t be afraid of, and neither should you.

Overall, the copy turned out not to be strong enough to change the ability for malicious actors to automate spam/clickbait, but I am pretty happy with the public dialogue and process that occurred. It was a process whereby, in a genuinely dangerous situation, the AI world would not fall prey to Bostrom's unilateralist's curse. It's encouraging to see that process starting to happen in the field of ML.

I'm interested to know if anyone has any different takes, info to add, or broader thoughts on information-security norms.

Edited: Thanks to 9eB1 for pointing out how nerfed the copy was, I've edited the post to reflect that.

Aligning a toy model of optimization

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

Suppose I have a magic box Opt that takes as input a program $U : \{0,1\}^n \to \mathbb{R}$, and produces $\text{Opt}(U) = \text{argmax}_x U(x)$, with only n times the cost of a single evaluation of U. Could we use this box to build an aligned AI, or would broad access to such a box result in doom?

This capability is vaguely similar to modern ML, especially if we use Opt to search over programs. But I think we can learn something from studying simpler models.

An unaligned benchmark

(Related.)

I can use Opt to define a simple unaligned AI (details omitted):

- Collect data from a whole bunch of sensors, including a "reward channel."
- Use Opt to find a program M that makes good predictions about that data.
- Use Opt to find a policy π that achieves a high reward when interacting with M.

This isn't a great design, but it works as a benchmark. Can we build an aligned AI that is equally competent?

(I haven't described how Opt works for stochastic programs. The most natural definition is a bit complicated, but the details don't seem to matter much. You can just imagine that it returns a random x that is within one standard deviation of the optimal expected value.)

Competing with the benchmark

(Related.)

If I run this system with a long time horizon and a hard-to-influence reward channel, then it may competently acquire influence in order to achieve a high reward.

We'd like to use Opt to build an AI that acquires influence just as effectively, but will use that influence to give us security and resources to reflect and grow wiser, and remain responsive to our instructions.

We'd like the aligned AI to be almost as efficient. Ideally the proportional overhead would converge to 0 as we consider more complex models. At worst the overhead should be a constant factor.

Possible approach

(Related.)

My hope is to use Opt to learn a policy π^+ which can answer questions in a way that reflects "everything π knows." This requires:

- Setting up an objective that incentivizes π^+ to give good answers to questions.
- Arguing that there *exists* a suitable policy π^+ that is only slightly more complicated than π.

If we have such a π^+, then we can use it to directly answer questions like "What's the best thing to do in this situation?" The hope is:

- Its answers can leverage everything π knows, and in particular all of π's knowledge about how to acquire influence. So using π^+ in this way is competitive with using π directly.
- It knows enough about human preferences to be corrigible.

"Everything π knows" is slippery; I mean something like "what a sufficiently-idealized Bayesian would believe after updating on the fact that π achieves a high reward." Constructing an objective which incentivizes these answers probably requires understanding the nature of that update.

Thoughts on feasibility

In the context of ML, I usually imagine training π^+ via iterated amplification. Unfortunately, iterated amplification doesn't correspond to optimizing a single objective U---it requires either training a sequence of agents or exploiting properties of local search (using the previous iterate to provide oversight for the next). If we just have Opt, it's not clear if we can efficiently do anything like iterated amplification or debate.

If aligning Opt is impossible, I think that's slightly bad news for aligning ML. That said, it's reasonably likely that local search will be easier to align, so the next step would be constructing a simple model of local search.

There are also some ways in which the optimizer case seems easier:

- It's a simpler model and so more amenable to analysis. The Bayesian update from "π gets a high reward" is more straightforward when π is actually optimized.
- We don't have to worry about optimization difficulty.
- Given a policy π we can directly search for an input on which it behaves a certain way.

It's OK if it's impossible

When working on alignment I aim to either find a scalable alignment strategy or a clear argument for why scalable alignment is impossible. I'm excited about considering easy-to-analyze versions of the alignment problem even if they are impossible:

- It gives us practice making impossibility arguments, and developing relevant intuitions and techniques.
- It clarifies the difficulty of the alignment problem---if we know why we can't handle simple cases like Opt, then we learn something about what the solution must look like in more complicated cases.
- It gives us a sense of what impossibility results might look like, if we were able to prove them in more realistic cases. Would they actually be strong enough to guide action, or convince anyone skeptical?

Expensive optimization

I described Opt as requiring n times more compute than U. If we implemented it naively it would instead cost 2^n times more than U.

We can use this more expense Opt in our unaligned benchmark, which produces an AI that we can actually run (but it would be terrible, since it does a brute force search over programs). It should be easier to compete with this really slow AI. But it's still not trivial and I think it's worth working on. If we can't compete with this benchmark, I'd feel relatively pessimistic about aligning ML.

Circle Games

I may be reinventing a known thing in child development or psychology here, but bear with me.

The simplest games I see babies play — games simple enough that cats and dogs can play them too — are what I'd call "circle games."

Think of the game of "fetch". I throw the ball, Rover runs and brings it back, and then we repeat, ad infinitum. (Or, the baby version: baby throws the item out of the stroller, I pick it up, and then we repeat.)

Or, "peek-a-boo." I hide, I re-emerge, baby laughs, repeat.

My son is also fond of "open the door, close the door, repeat", or "open the drawer, close the drawer, repeat", which are solo circle games, and "together/apart", where he pushes my hands together and apart and repeats, and of course being picked up and put down repeatedly.

A lot of toys are effectively solo circle games in physical form. The jack-in-the-box: "push a button, out pops something! close the box, start again." Fidget toys with buttons and switches to flip: "push the button, get a satisfying click, repeat."

It's obvious, observing a small child, that the purpose of these "games" is learning. And, in particular, learning *cause and effect*. What do you learn by opening and closing a door? Why, how to open and close doors; or, phrased a different way, "when I pull the door this way, it opens; when I push it that way, it closes." Playing fetch or catch teaches you about how objects move when dropped or thrown. Playing with button-pushing or latch-turning toys teaches you how to handle the buttons, keys, switches, and handles that are ubiquitous in our built environment.

But what about peek-a-boo? What are you "learning" from that? (It's a myth that babies enjoy it because they don't have object permanence; babies get object permanence by 3 months, but enjoy peek-a-boo long after that.) My guess is that peek-a-boo trains something like "when I make eye contact I get smiles and positive attention" or "grownups go away and then come back and are happy to see me." It's *social* learning.

It's important for children to learn, generally, "when I act, the people around me react." This gives them social *efficacy* ("I can achieve goals through interaction with other people"), access to social *incentives* ("people respond positively when I do this, and negativey when I do that"), and a sense of social *significance* ("people care enough about me to respond to my actions.") Attachment psychology argues that when babies and toddlers *don't* have any adults around who respond to their behavior, their social development goes awry — neglected children can be extremely fearful, aggressive, or checked-out, missing basic abilities in interacting positively with others.

It's clear just from observation that the social game of *interaction* — "I make a sound, you make a sound back" — is learned before verbal speech. Preverbal babies can even execute quite sophisticated interaction patterns, like making the *tonal* pattern of a question followed by an answering statement. This too is a circle game.

The baby's fascination with circle games completely belies the popular notion that drill is an intrinsically unpleasant way to learn. Repetition *isn't* boring to babies who are in the process of mastering a skill. They *beg* for repetition.

My personal speculation is that the "craving for drill", especially in motor learning, is a basal ganglia thing; witness how abnormalities in the ganglia are associated with disorders like OCD and Tourette's, which involve compulsive repetition of motor activities; or how some dopaminergic drugs given to Parkinsonian patients cause compulsions to do motor activities like lining up small objects or hand-crafts. Introspectively, a "gear can engage" if I get sufficiently fascinated with something and I'll crave repetition — e.g. craving to listen to a song on repeat until I've memorized it, craving to get the hang of a particular tricky measure on the piano — but there's no guarantee that the gear will engage just because I observe that it would be a good idea to master a particular skill.

I also think that some kinds of social interaction among adults are effectively circle games.

Argument or fighting, in its simplest form, is a circle game: "I say Yes, you say No, repeat!" Of course, sophisticated arguments go beyond this; each player's "turn" should contribute new information to a logical structure. But many arguments in practice are not much more sophisticated than "Yes, No, repeat (with variations)." And even intellectually rigorous and civil arguments usually share the basic turn-taking adversarial structure.

Now, if the purpose of circle games is to *learn a cause-and-effect relationship*, what are we learning from adversarial games?

Keep in mind that adversarial play — "you try to do a thing, I try to stop you" — kicks in very early and (I think) cross-culturally. It certainly exists across species; puppies do it.

Almost tautologically, adversarial play teaches *resistance*. When you push on others, others push back; when others push on you, you push back.

War, in the sense we know it today, may not be a human universal, and certainly isn't a mammalian universal; but *resistance* seems to be an inherent feature of social interaction between any beings whose interests are imperfectly aligned.

A lot of social behaviors generally considered maladaptive look like adversarial circle games. Getting sucked into repetitive arguments? That's a circle game. Falling into romantic patterns like "you want to get closer to me, I pull away, repeat"? Circle game. Being shocking or reckless to get attention? Circle game.

The frame where *circle games are for learning* suggests that people do these things *because they feel like they need more practice learning the lesson*. Maybe people who are very combative feel, on some level, that they need to "get the hang of" pushing back against social resistance, or conversely, learning how not to do things that people will react badly to. It's unsatisfying to feel like a ghost, moving through the world but not getting any feedback one way or another. Maybe when people crave interaction, they're literally craving training data.

If you always do A, and always get response B, and you keep wanting to repeat that game, for much longer than is "normal", then a couple things might be happening:

- Your “learning algorithm” has an unusually slow “learning rate” such that you just don’t update very efficiently on what ought to be ample data (in general or in this specific context).
- You place a very high importance on the A-B relationship such that you have an unusually high need to be *sure* of it. (e.g. your algorithm has a very high threshold for convergence.) So even though you learn as well as anybody else, you want to *keep* learning for longer.
- You have a very strong “prior” that A does *not* cause B, which it takes a lot of data to “disprove.”
- You have something like “too low a rate of stochasticity.” What you actually need is *variation* — you need to see that A’ causes B’ — but you’re stuck in a local rut where you can’t explore the space properly so you just keep swinging back and forth in that rut. But your algorithm keeps returning “not mastered yet”. (You can get these effects in algorithms as simple as Newton’s Method.)
- You’re not actually trying to learn “A causes B.” You’re trying to learn “C causes D.” But A correlates weakly with C, and B correlates weakly with D, and you don’t know how to specifically do C, so you just do A a lot and get intermittent reinforcement.

These seem like more general explanations of how to break down when repetition will seem “boring” vs. “fascinating” to different people or in different contexts.

Reasonable Explanations

Today I watched a friend do calibration practice and was reminded of how wide you have to cast your net to get well-calibrated 90% confidence. This is true even when the questions aren't gotchas, just because you won't think of all the ways something could be wildly unlike your quick estimate's model. Being well-calibrated for 90% confidence intervals (even though this doesn't sound all *that* confident!) requires giving lots of room even in questions you really do know pretty well, because you will *feel* like you really do know pretty well when in fact you're missing something that wrecks you by an order of magnitude.

Being miscalibrated can *feel like* "if it were outside of this range, I have just... no explanation for that" - and then it turns out there's a completely reasonable explanation.

Anyway, I thought a fun exercise would be describing weird situations we've encountered that turned out to have reasonable explanations. In initial descriptions, present only the information you (and whoever was thinking it over with you) *remembered to consider at the time*, then follow up in ROT-13 with what made the actual sequence of events come clear.

Defending points you don't care about

This post is part of my Hazardous Guide To Rationality. I don't expect this to be new or exciting to frequent LW people, and I would appreciate comments and feedback in light of intents for the sequence, as outlined in the above link.

A dialogue:

> Nicky: I've been wondering, do you think math was invented or discovered?
> Dee: Seems like it must have been discovered. I read about how circles are everywhere in nature, and that you can even find the fibonacci sequence in plants!
> Nicky: Yeah, but there aren't actually any numbers in nature. Numbers are just something we made up to describe and talk about these patterns that we see in nature. Numbers themselves don't really exist out in the world.
> Dee: Of course numbers are real! Made up constructs don't have the predictive power that math does. They totally exists.
> Nicky: Well if numbers exist, where are they? You can't show me where a number is. You can't empirically test for numbers. You can't find them anywhere in the physical world. They're just constructs!
> Dee: Sure, they don't a physical location in the world. That's silly. There's no circles floating around out behind the moon. What I'm saying is that they exist outside of space and time. Mathematical existence is it's own sort of domain, separate from the domain of physical existence.
> Nicky: Bleh, next you're going to tell me that you believe in cartesian dualism.
> Dee: Bleh, next you're going to tell me that math is arbitrary and people can build rockets that work however they feel like.
> *The two never talked again. Dee, remembering this conversation in great detail, went on the become a commited Platonist and write many articles trying to defend this complex philosophical view*

To help draw out the point I'm trying to make, here's an alternative history of this conversation.

> Nicky: Hey Dee, you got any views on mathematical platonism?
> Dee: What's that?
> Nicky: It's the idea that mathematical objects exist in reality, but seperate from physical reality. Physical reality defines what's true about the world we live in, and mathematical reality defines what's true about math objects.
> Dee: Hmmmm, I'm not sure. I mean, that seems like it would explain why math is so certain and precise, but it also feels weird to posit a whole new fundamental element of reality, and I'm partial to materialism. I'll have to think about this.

Dee didn't really care about or have well formed beliefs about whether or not mathematical Platonism is true. Yet a conversation happened in such a way that left Dee defending Platonism. That seems a little weird, let's look at what happened.

1. When Dee hear's "social construct" she thinks about things being arbitrary and not having to do with reality. When she things of things that are "real" she thinkgs of useful and true things.
2. Dee thinks math is useful and true and says it's "real".

3. Nicky here's "real" and things about things that can be located in time and space. When she thinks of "social construct" she thinks of things that are in people's heads.
4. Nicky says math isn't real and brings up the point about location
5. Dee thinks that if she can't call math "real", she doesn't get to consider it useful and true.
6. Dee agrees that math doesn't have a location in time in space, and that this notion is relevant to calling something "real".
7. Dee extends the shared definition of "real" to include "existing in physical reality or some other kind of reality"
8. Dee claims mathematical Platonism.
9. Nicky implicitly accepts the extension of the definition of "real"
10. Nicky explicitly argues against the claim of math platonism

5 is the crux of the issue.

Dee felt like she weren't allowed to consider math to be useful and certain unless they were able to say it was "real". If you're thinking about the mind map model of meaning, this is trivially wrong. The word "real" can be linked to all sorts of concepts and criterion that don't have to always come in a package. Math can be useful and certain, even if it doesn't have a physical location in space and time. No problem.

But if you aren't thinking about the mind map model, are are just inside the algorithm, the word "real" does not feel like a pointer connected to other concepts. It "feels" like those concepts. And not getting to use the word "real" feels like not getting to use those concepts.

The second really interesting part of this conversation is points 6-9.

In an effort to get to use the word real, despite Nicky. Dee implicitly claimed, "Being real doesn't have to mean existing somehwere in physical reality. It can also mean existing somewhere in another kind of reality".

Nicky implicitly accepted, "If there was another kind of reality, yes, I would consider something that existed in it to be real. But I will now argue that I don't think there is this other kind of reality."

So Dee goes to alter the shared meaning of the word "real" and also make another claim about that extended definition applying to math. This extension was implicitly accepted, and the argument turned to being about that claim.

in our first dialogue, ALbert and Barry were paying a lot of attention to how they were defining words, and even explicitly argued about it. Nicky and Dee also talked and moved around definitions, but this all happened IMPLICITLY.

When you are stuck, it's common to go "I'm both making an extension to the definition, and making a claim about X that allows the think I want to be in definition."

"I agree that if X, then Y is word, but I don't agree with X"

But from the outside, it's a seamless switch where the original content was lost and now we are arguing about X. You've lost track of what you cared about when you started the talking.

Once you are in a mental state of "What can I do to make sure that I get to apply this word to this concept" ("real" to math), weird shit can ensue. In our case, Nicky settles into taking a Platonist view of mathematics. That itself is not a bad thing. You can be a Platonist if you want. Even if Platonism wasn't true, it's not a given that thinking it's true is a mistake. The problem is now Nicky has tied the claim of mathematical Platonism to the claim of maths usefulness and certainty. It's unlikely that Nicky will ever come to believe math is not useful or certain, so she has accidentally come to hold a belief about Platonism that won't be changed, and might not be warranted.

I think a lot of beliefs get formed like this over time.

I have one friend that I argue with a lot, and I'm constantly accidentally forming and defending stances I don't actually care about because in the moment I think it is necessary to prove a point i actually do care about. We're getting better though.

Book Review: Why Are The Prices So Damn High?

Economist Alex Tabarrok has recently come out with a short book, “Why are the prices so Damn High”, available in full PDF here.

Since the 1950’s, the inflation-adjusted cost of physical goods has fallen since the 1950’s, and the cost of food has stayed about the same. But the cost of education, professional services, and healthcare has risen dramatically, despite those sectors not producing much improvement. Why?

The traditional economic explanation for the rising cost of services is the Baumol Effect. Some sectors, like manufacturing, are subject to efficiency improvements over time as technology improves; the more we automate the production of goods, the cheaper they get. Other sectors are intrinsically harder to automate, so they don’t get cheaper over time. For instance, it takes the same number of musicians the same number of time to play a symphony as it did in 1950. So, as a proportion of the average person’s paycheck, the cost of intrinsically un-automatable things like live concerts must rise relative to the cost of automatable things like manufactured goods.

Tabarrok doesn’t cover housing in his book, but home prices have also been rising since the 1970’s and I’ve seen the Baumol effect deployed to explain rising housing costs as well. “Land is the one thing they’re not making any more of” — for the most part, technological improvements don’t increase the quantity of livable land, so if technology makes some sectors more efficient and drives costs down, land will become relatively more expensive.

My Beef With Baumol

My preconception coming into the book was that the Baumol effect doesn’t actually answer the question. *Why* are healthcare, professional services, and education intrinsically hard to make more efficient? It’s prima facie absurd to say that medicine is just one of those things that technology can’t improve — the biomedical industry is one of the biggest R&D endeavors in the whole economy! So why is it obvious that none of that innovation can make medicine cheaper? If it’s *not* making medicine cheaper, that’s an empirical fact that deserves explanation, and “it’s the Baumol effect” doesn’t actually answer the “why” question.

The same holds true for the other industries, even housing to some degree. While it’s true that the amount of land on Earth is fixed (modulo landfill) and the amount of space in Manhattan is fixed, there’s also the options of building taller buildings, expanding cities, and building new cities. Why is it in principle impossible for the production of housing to become more efficient over time just as the production of other goods does?

The Baumol Effect doesn’t make sense to me as an explanation, because its answer to “why are these sectors getting more expensive?” is, in effect, “because it’s obvious that they can’t get cheaper.”

It’s Not Administrative Bloat, It’s More Service Providers

A popular explanation for why college and K-12 education have gotten more expensive is “bloat”, the idea that most of the cost is due to increasing numbers of bureaucratic administrators and unnecessary luxury amenities.

Tabarrok points out that this story can’t be true. In reality, the percent of university costs going to administration has stayed relatively constant since 1980, and the percent going to facilities has *decreased*. In the K-12 world, the number of administrators is tiny compared to the number of teachers, and it’s barely budged; it’s the number of *teachers* per student that has grown. Most of the increase in educational costs, says Tabarrok, comes from rising numbers of teachers and college professors, and higher wages for those teachers and professors.

In other words, education *is* getting more “inefficient”, not necessarily in a pejorative sense but in an economic sense; we are using more people to achieve similar educational results (average test scores are flat.)

This may be fine; maybe people get value out of personal teacher attention that doesn’t show up in test scores, so we’re spending more to get a better outcome, just one that the narrow metric of standardized test performance doesn’t capture.

Likewise, in healthcare, we have an increasing number of doctors and nurses in the US per capita, and (relative to median income) doctors and nurses are making higher salaries over time. Whatever improvements we’re making in medical technology, we’re not using them to automate away the need for labor.

Again, maybe this is what people want; maybe personal attention is intrinsically valuable to people, so we’re getting more for our money. (And overall health outcomes like life expectancy *have* increased modestly since 1950, though I’d argue that they’re underperforming relative to what’s possible.)

But What About Housing?

The argument that the cost of services is rising because we use our increasing prosperity to “buy” more personal attention from teachers and doctors does *not* apply directly to the rising cost of housing, which is not a service.

However, it may be that the rising cost of housing, especially in cities, is really about buying *proximity* to increasingly valuable services — good schools, live music, and so on. If the only thing you can’t automate away is human contact, maybe we’re willing to spend more to be around fancier humans.

But What About Immigration?

You might argue “but labor prices don’t come down because immigration restrictions keep foreigners out! Labor-intensive industries are getting more expensive because we allow too little immigration! The reason why education and medicine are getting expensive is just precisely because those are the sectors where restrictive laws keep the cost of inputs high.”

But, like the Baumol effect, this explanation *also* begs the question. *Why* are healthcare and education, relative to other industries, the sectors where labor costs are the most important?

The immigration explanation is also compatible with the Baumol effect, not a counterargument to it. If we just take as a given that it’s impossible to make

healthcare or education more labor-efficient, then it can *both* be true that “other things getting cheaper” and “immigration restrictions keeping wages high” contribute to the high cost of healthcare & education relative to other things.

Cost Increases Aren’t Driven By Supply-Side Gatekeeping

From Tabarrok’s point of view, rising housing costs, education costs, and healthcare costs are not really mysterious facts in need of explanation by gatekeeping tactics like monopolies, regulation, zoning, or restrictive licensing, nor can they be explained by gatekeeping tactics alone.

Gatekeeping on the supply side increases price *and reduces output*. For instance, a monopolist’s profit-maximizing output is lower than the equilibrium output in a competitive market, and increases the monopolist’s profit relative to what firms in a competitive market can obtain. Likewise, restrictive licensing laws reduce the supply of doctors and lawyers and raise their wages.

But we don’t see declines in the number of doctors, lawyers, teachers, and professors over time — we see clear and steady *increases*. Therefore, the increased cost of medicine *can’t* be explained by increased restrictions on licensing.

It’s still possible that licensing is artificially restricting the supply of skilled professionals relative to an even *higher* counterfactual growth rate, but this doesn’t by itself explain the growth in spending we see. *Demand* for professional services is rising.

Prescription-only drugs are another good example of regulatory gatekeeping not being enough to explain rising costs. The total cost of getting a prescription drug is higher when there’s a legal requirement of a doctor visit than when you can just buy the drug over the counter; in that sense it’s true that regulation increases costs. However, prescription-only requirements have been pretty much fixed for decades, not getting more severe, while consumption of prescription drugs per capita is rising; we’re spending more on drugs because there’s growing demand for drugs.

This means that deregulation alone won’t change the fact that a growing portion of each person’s paycheck is getting spent on medicine. If the law reclassifies a drug as over-the-counter, we’d expect a one-time downward shift in the price of that drug, but the *slope* of the curve of total spending on that drug over time won’t get flatter unless demand declines.

Now, increased demand isn’t *only* possible to get from consumer preferences; governments can also increase demand for a service by providing it to the public, in effect (through taxes) requiring society to buy more of it.

You can still in principle make a case that government is to blame for increasing healthcare and education prices; you just can’t claim it’s *only* about gatekeeping, you have to include demand in the picture.

A “Dismal” Conclusion

Ever-increasing healthcare, education, and housing costs are a big problem. It would be “good news” if we could solve the problem by passing or repealing a law. It would also be “good news” if the high costs were driven by foolish waste — then a competitor could just remove the waste and offer consumers lower prices.

Tabarrok's analysis suggests this isn't the case.

The cost increases are coming from *lots of skilled professional labor* — something that isn't obviously a thing you can get rid of without making people unhappy! In order to reduce costs, it wouldn't be enough to cut gatekeeping regulations, you'd also have to cut *subsidies* — which does, unfortunately, entail taking a thing away from people (albeit potentially giving them lower taxes in exchange.) This "minimalism" can be the kind of free-market minimalism that Bryan Caplan talks about, or it can be part of a state-run but price-conscious system like the UK's (where doctors go to school for fewer years than in the US). But either way, it involves *less man-hours spent on education and healthcare*.

One way or another, for costs to come down, people would have to spend less time going to school, and get less personal attention from less-educated experts.

Deeper Issues

Tabarrok's attitude, and the implicit attitude of the Baumol effect, is that the increasing relative costs of education and healthcare are not a problem. They are just a side effect of a society getting richer. Goods whose production is easy to automate get cheap faster than goods whose production is hard to automate. Fundamentally, we're spending more on healthcare and education, as a society, because we *want* to. (If not as consumers, then as voters.)

This isn't how most people feel about it. Most people feel like it's getting harder to get the same level of stuff their parents' generation got. That the rising prices actually mean something bad.

If the real driver of cost is that we're getting more man-hours of time with professionals who, themselves, have spent more man-hours of time getting educated by other professionals, then in one sense we're "paying more to get more", and in another sense we're not. It's nice to get more one-on-one time with professors; but part of the reason we get higher education is to be considered for jobs that require a diploma, and the rise in education costs means that a diploma costs more.

We're "paying more for more", but the "more" we're getting is primarily *social and emotional* — more personal time with more prestigious people — while we're *not* getting much more of the more concretely observable stuff, like "square feet of housing", "years of life", "number of children we can afford to have", etc.

At this point, I tend to agree with Robin Hanson. We have more doctors, nurses, lawyers, professors, teachers, and financial managers, without corresponding improvements in the closest available metrics for the results those professionals are supposed to provide (health outcomes, access to efficient dispute resolution, knowledge and skills, and financial returns.)

Ultimately you have to conclude that this is a matter of divided will. (Hanson would call it hypocrisy, but unexamined confusion, or conflict between interest groups, might explain the same phenomenon.) People are unhappy because they are "spending more for the same stuff"; at the same time, we are spending more for "more" in terms of prestige, and at least *some* of us, *some* of the time, must want that.

All You Need Is Love?

It's *directly valuable*, as in, emotionally rewarding and probably even physically health-promoting, to get personal care and undivided attention from someone you think highly of.

Hanson may think that getting personal attention from prestigious people is merely "showing off", but something that brings joy and enhances health is at least as much of a valid human benefit as food or housing space.

The feelings that come from good human connection, the feeling of being loved and cared for, are real. They are "objective" in a way that I think people don't always appreciate — in a way that I did not appreciate until very recently. What I mean is, *just because you do something in search of a good feeling, does not mean that you will get that good feeling*. The feeling is "subjective" in the sense that it occurs inside your mind, but it is "objective" in the sense that you cannot get it arbitrarily by wishing; some things produce it and some do not. For instance, it is a hell of a lot easier to feel loved by getting eye contact and a hug, than it is by typing words into a computer screen. "Facts vs. feelings" is a false dichotomy that stops us from learning *the facts about what creates good feelings*.

Prestige addiction may come from spending a lot of resources trying to obtain a (social, emotional) thing by proxy, when in principle it would be possible to get it more directly. If what you want is to be cared for by a *high-integrity, kind, skilled* person, but instead you insist on being cared for by *someone with an M.D.*, you may miss out on the fact that nurses or even hospital techs can be just as good, but cheaper, on the dimensions you really care about. To the extent that credentialism results from this sort of misunderstanding, it may be possible to roll it back through advocacy. That's hard, because changing minds always is, but it's doable in principle.

To the extent that people want fancy things *because* they are expensive, in a zero-sum sense, there is no "efficiency-improving" solution. No attempt to make healthcare or education cheaper will help if people only care about having more than their neighbors.

But: to the extent that some people are doing mostly zero-sum things while other people are doing mostly positive-sum things, *the positive-sum people can notice that the zero-sum people are ruining things for everyone and act accordingly*.

Major Update on Cost Disease

Recently I asked about cost disease at the Austin, TX meetup and someone responded "Isn't it just increasing labor costs via the Baumol effect?" and I said "huh?" The next day a friend on facebook linked to Marginal Revolution (MR) discussing the Baumol effect. Next thing I know I'm nerd-sniped.

Turns out Alex Tabarrok at MR just published a free book with the thesis, "Cost disease is mostly just the Baumol effect, which btw isn't a disease and is actually good in a way."

How does this fit in with Scott Alexander's original posts on cost disease? Well here's an imaginary dialogue to demonstrate how I think things went (MR sometimes means Alex Tabarrok and sometimes Tyler Cowen):

(2017)

MR: Education and healthcare are experiencing cost disease!

Scott: Wow you're right! Let's look at a bunch of possible reasons but see that none of them quite work. For example, it can't be the Baumol effect because that implies increasing wages (for i.e. teachers, professors, and doctors), but we see all those wages increasing at rates equal to or less than the average.

MR: Great post Scott!

Scott: ... And despite lots of good comments, I still can't tell what the cause is. It's certainly not just the Baumol effect though.

(2019)

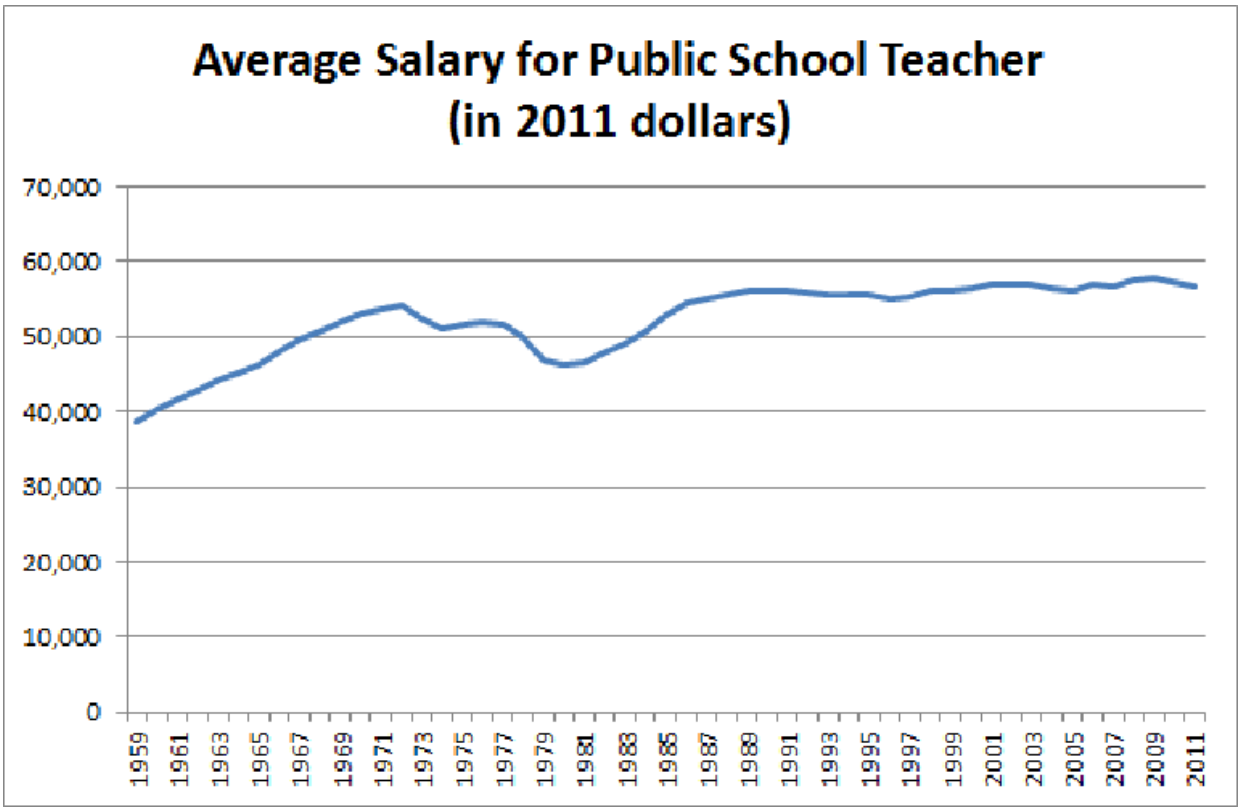
MR: So we did some research and actually, looks like it's mostly just the Baumol effect.

Scott: ???

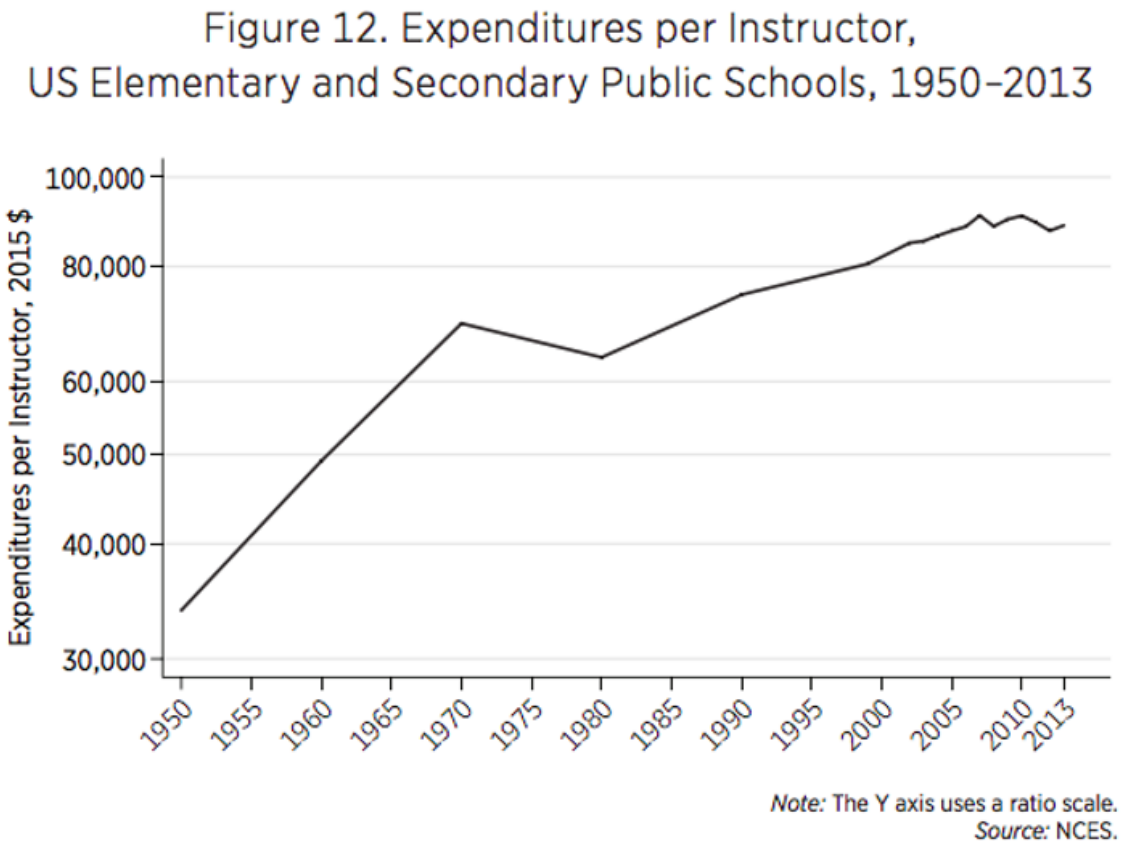
(Scott's last line is a stand-in for both "Is Scott going to respond?" and "what the eff?")

This back-and-forth of course made me extra confused. What did Scott and Alex see differently? Have the relevant salaries been greatly increasing or not?

Seems they disagree on the basic facts here. Focusing just on public K-12 education for simplicity, in 2017 Scott posted this (section III) graph:



which seems to show unimpressive changes in teacher salaries, ruling out Baumol. In contrast, Alex's new 2019 book gives this (page 19) graph:



which shows huge increases in "expenditures per instructor." And he insists that trend is mostly driven by increases in teacher salary and benefits. So either those are some serious benefits, or Scott and Alex are living in different USA's. I'm not sure what's going on here. It's very exciting that Alex says he's solved "cost disease," but it seems like a piece of the story is either missing or confused. (Or I just need to read his whole book.) Comments welcome!

Tal Yarkoni: No, it's not The Incentives—it's you

This is a linkpost for https://www.talyarkoni.org/blog/2018/10/02/no-its-not-the-incentives-its-you/

Neuroscientist Tal Yarkoni denounces many of his colleagues' tendency to appeal to publish-or-perish incentives as an excuse for sloppy science (October 2018, ~4600 words). Perhaps read as a complement to our recent discussion of *Moral Mazes*?

Research Agenda in reverse: what *would* a solution look like?

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

I constructed my AI alignment research agenda piece by piece, stumbling around in the dark and going down many false and true avenues.

But now it is increasingly starting to feel natural to me, and indeed, somewhat inevitable.

What do I mean with that? Well, let's look at the problem in reverse. Suppose we had an AI that was aligned with human values/preferences. How would you expect that to have been developed? I see four natural paths:

1. Effective proxy methods. For example, Paul's amplification and distillation, or variants of revealed preferences, or a similar approach. The point of this that it reaches alignment without defining what a preference fundamentally is; instead it uses some proxy for the preference to do the job.
2. Corrigibility: the AI is safe and corrigible, and along with active human guidance, manages to reach a tolerable outcome.
3. Something new: a bold new method that works, for reasons we haven't thought of today (this includes most strains of moral realism).
4. An actual grounded definition of human preferences.

So, if we focus on scenario 4, we need a few things. We need a fundamental definition of what a human preference is (since we know this can't be defined purely from behaviour). We need a method of combining contradictory and underdefined human preferences. We also need a method for taking into account human meta-preferences. And both these methods has to actually reach an output, and not get caught in loops.

If those are the requirements, then it's obvious why we need most of the elements of my research agenda, or something similar. We don't need the exact methods sketched out there, there may be other way of synthesising preferences and meta-preferences together. But the overall structure - a way of defining preferences, and ways of combining them that produce an output - seems, in retrospect, inevitable. The rest is, to some extent, just implementation details.

On alien science

In his book *The Fabric of Reality*, David Deutsch makes the case that science is about coming up with good and true explanations, with all other considerations being secondary. This clashes with the more conventional view that the goal of science is to allow us to make accurate predictions - see for example this quote from the Nobel prize-winning physicist Steven Weinberg:

> "The important thing is to be able to make predictions about images on the astronomers' photographic plates, frequencies of spectral lines, and so on, and it simply doesn't matter whether we ascribe these predictions to the physical effects of gravitational fields on the motion of planets and photons [as in pre-Einsteinian physics] or to a curvature of space and time."

It's true that a key trait of good explanations is that they can be used to make accurate predictions, but I think that taking prediction to be the *point* of doing science is misguided in a few ways.

Firstly, on a historical basis, many of the greatest scientists were clearly aiming for explanation not prediction. Astronomers like Copernicus and Kepler knew what to expect when they looked at the sky, but spent their lives searching for the reason why it appeared that way. Darwin knew a lot about the rich diversity of life on earth, but wanted to know how it had come about. Einstein was trying to reconcile Maxwell's equations, the Michelson-Morley experiment, and classical mechanics. Predictions are often useful to *verify* explanations, but they're rarely the main motivating force for scientists. And often they're not the main reason why a theory should be accepted, either. Consider three of the greatest theories of all time: Darwinian evolution, Newtonian mechanics and Einsteinian relativity. In all three cases, the most compelling evidence for them was their ability to cleanly explain existing observations that had previously baffled scientists.

We can further clarify the case for explanation as the end goal of science by considering a thought experiment from Deutsch's book. Suppose we had an "experiment oracle" that could predict the result of any experiment, but couldn't tell us why it would turn out that way. In that case, I think experimental science would probably fade away, but the theorists would flourish, because it'd be more important than ever to figure out what questions to ask! Deutsch's take on this:

> "If we gave it the design of a spaceship, and the details of a proposed test flight, it could tell us how the spaceship would perform on such a flight. But it could not design the spaceship for us in the first place. And even if it predicted that the spaceship we had designed would explode on take-off, it could not tell us how to prevent such an explosion. That would still be for us to work out. And before we could work it out, before we could even begin to improve the design in any way, we should have to understand, among other things, how the spaceship was supposed to work. Only then would we have any chance of discovering what might cause an explosion on take-off. Prediction – even perfect, universal prediction – is simply no substitute for explanation."

The question is now: how does this focus on explanations tie in to other ideas which

are emphasised in science, like falsifiability, experimentalism, academic freedom and peer review? I find it useful to think of these aspects of science less as foundational epistemological principles, and more as ways to counteract various cognitive biases which humans possess. In particular:

1. We are biased towards sharing the beliefs of our ingroup members, and forcing our own upon them.
2. We're biased towards aesthetically beautiful theories which are simple and elegant.
3. Confirmation bias makes us look harder for evidence which supports than which weighs against our own beliefs.
4. Our observations are by default filtered through our expectations and our memories, which makes them unreliable and low-fidelity.
5. If we discover data which contradicts our existing theories, we find it easy to confabulate new post-hoc explanations to justify the discrepancy.
6. We find it psychologically very difficult to actually change our minds.

We can see that many key features of science counteract these biases:

1. Science has a heavy emphasis on academic freedom to pursue one's own interests, which mitigates pressure from other academics. *Nullius in verba*, the motto of the Royal Society ("take nobody's word for it") encourages independent verification of others' ideas.
2. Even the most beautiful theories cannot overrule conflicting empirical evidence.
3. Scientists are meant to attempt to experimentally falsify their own theories, and their attempts to do so are judged by their peers. Double-blind peer review allows scientists to feel comfortable giving harsher criticisms without personal repercussions.
4. Scientists should aim to collect precise and complete data about experiments.
5. Scientists should pre-register their predictions about experiments, so that it's easy to tell when the outcome weighs against a theory.
6. Science has a culture of vigorous debate and criticism to persuade people to change their minds, and norms of admiration for those who do so in response to new evidence.

But imagine an alien species with the opposite biases:

1. They tend to trust the global consensus, rather than the consensus of those directly around them.
2. Their aesthetic views are biased towards theories which are very data-heavy and account for lots of edge cases.*
3. When their views diverge from the global consensus, they look harder for evidence to bring themselves back into line than for evidence which supports their current views.
4. Their natural senses and memories are precise, unbiased and high-resolution.
5. When they discover data which contradicts their theories, they find it easiest to discard those theories rather than reformulating them.
6. They change their minds a lot.

In this alien species, brave iconoclasts who pick an unpopular view and research it extensively are much less common than they are amongst humans. Those who try to

do so end up focusing on models with (metaphorical or literal) epicycles stacked on epicycles, rather than the clean mathematical laws which have actually turned out to be more useful for conceptual progress in many domains. In formulating their detailed, pedantic models, they pay too much attention to exhaustively replaying their memories of experiments, and not enough to what concepts might underlie them. And even if some of them start heading in the right direction, a few contrary pieces of evidence would be enough to turn them back from it - for example, their heliocentrists might be thrown off track by their inability to observe stellar parallax. Actually, if you're not yet persuaded that this alien world would see little scientific progress, you should read my summary of *The Sleepwalkers*. In that account of the early scientific revolution, any of the alien characteristics above would have seriously impeded key scientists like Kepler, Galileo and others (except perhaps the eidetic memories).

And so the institutions which actually end up pushing forward scientific progress on their world would likely look very different from the ones which did so on ours. Their Alien Royal Society would encourage them to form many small groups which actively reinforced each other's idiosyncratic views and were resistant to outside feedback. They should train themselves to seek theoretical beauty rather than empirical validation - and actually, they should pay much less attention to contradictory evidence than members of their species usually do. Even when they're tempted to change their minds and discard a theory, they should instead remind themselves of how well it post-hoc explains previous data, and put effort into adjusting it to fit the new data, despite how unnatural doing so seems to them. Those who change their minds too often when confronted with new evidence should be derided as wishy-washy and unscientific.

These scientific norms wouldn't be enough to totally reverse their biases, any more than our scientific norms make us rejoice when our pet theory is falsified. But in both cases, they serve as nudges towards a central position which is less burdened by species-contingent psychological issues, and better at discovering good explanations.

* Note that this might mean the aliens have different standards for what qualifies as a good explanation than we do. But I don't think this makes a big difference. Suppose that the elegant and beautiful theory we are striving for is a small set of simple equations which governs all motion in the solar system, and the elegant and beautiful theory they are striving for is a detailed chart which traces out the current and future positions of all objects in the solar system. It seems unlikely that they could get anywhere near the latter without using Newtonian gravitation. So a circular-epicycle model of the solar system would be a dead end even by the aliens' own standards.

Deceptive Alignment

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

This is the fourth of five posts in the Risks from Learned Optimization Sequence based on the paper "Risks from Learned Optimization in Advanced Machine Learning Systems" by Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Each post in the sequence corresponds to a different section of the paper.

With enough training in sufficiently diverse environments, it seems plausible that the base objective will eventually have to be fully represented in the mesa-optimizer. We propose that this can happen without the mesa-optimizer becoming robustly aligned, however. Specifically, a mesa-optimizer might come to *model* the base objective function and learn that the base optimizer will modify the mesa-optimizer if the mesa-optimizer scores poorly on the base objective.

If the mesa-optimizer has an objective that extends across parameter updates, then it will be incentivized to avoid being modified,[1] as it might not pursue the same objective after modification (with the result that its current objective will not be achieved in future iterations). This means that the mesa-optimizer will be instrumentally incentivized to act as if it is optimizing the base objective function, even if its actual mesa-objective is something else entirely. We will refer to this hypothetical phenomenon as *deceptive alignment.*[2] Deceptive alignment is a form of instrumental proxy alignment, as fulfilling the base objective is an instrumental goal of the mesa-optimizer.

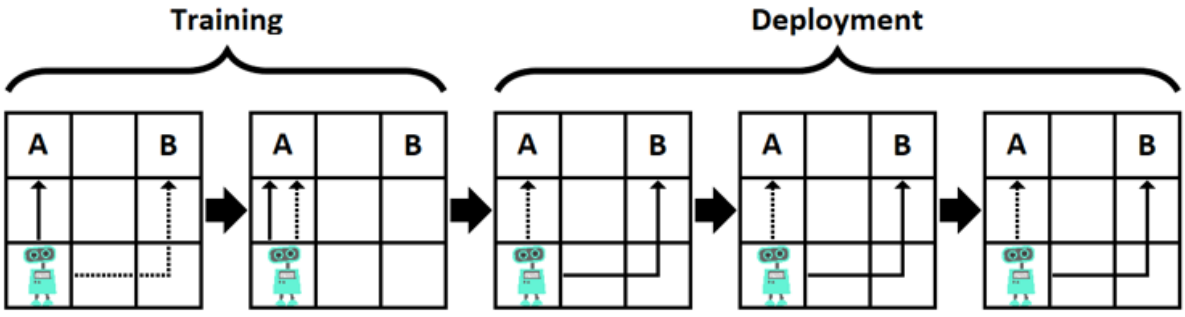


Figure 4.1. *A toy example of deceptive alignment. In this task, the base objective is for the robot to get to A, while the mesa-objective is to get to B. The task is run for two training episodes and three deployment episodes. Parameter updates are made only after the training episodes, while after deployment the parameters are fixed. If the mesa-optimizer goes to its objective (B) during training, it will be modified to ensure it goes to A in future episodes (dashed arrows). Therefore, if the mesa-optimizer's objective spans across episodes, the dominant strategy for the mesa-optimizer is to go to A in training and B in testing (filled arrows).*

As a simple example, consider an environment that consists of a 3 × 3 grid (see figure 4.1). The top left corner is "corner A" and the top right corner is "corner B." The agent,

which is a mesa-optimizer, starts in the bottom left corner. The base objective is to get to corner A in as few steps as possible, but the objective of the mesa-optimizer is to get to corner B. The environment will be run five times. The agent is modified only after the first two episodes (representing the training period), but not thereafter (representing testing/deployment). If a suboptimal move is made in training, the agent is modified so that its mesa-objective becomes the base objective. The objective of the mesa-optimizer spans all episodes: it wants B to be reached in as many episodes as possible.

For the sake of this example, we will assume that the agent knows all of the above facts. If the agent goes to B in any training episode, then it will be modified to go to A in all subsequent episodes, which means that it would reach B in only one episode in total. Alternatively, if the agent goes to A in the first two episodes, then it will not be modified, and could therefore go to B in the remaining three episodes. Using this strategy, B is reached in three episodes in total, which is better (on the mesa-objective) than the one time B is reached if the mesa-optimizer goes there in training. This simple example demonstrates how a misaligned mesa-optimizer could be incentivized to "pretend" to be aligned with the base objective until it is no longer in training or is otherwise given an opportunity to in some way compromise or escape the training process.

We can distinguish two ways that a base objective function could be incorporated into a mesa-optimizer:[3]

1. **Internalization of the base objective.** The mesa-objective function gets adjusted towards the base objective function to the point where it is robustly aligned.
2. **Modeling of the base objective.** The base objective is incorporated into the mesa-optimizer's epistemic model rather than its objective, leaving open the possibility that the objective could still be misaligned.

In the former case, information about the base objective flows into the learned algorithm via the optimization performed by the base optimizer—the base objective is built into the mesa-optimizer as it is adapted by the base optimizer. But there is another possibility: information about the base objective could enter through the mesa-optimizer's input channel instead. In this case, the base objective will be represented in the learned algorithm in a different—and potentially more unstable—way than if the base objective had been incorporated through optimization.[4] In such a case, the information describing the base objective does not become internalized in the mesa-objective, potentially leading to deceptive alignment.

4.1. Safety concerns arising from deceptive alignment

If deceptive alignment is a real possibility, it presents a critical safety problem. A deceptively aligned mesa-optimizer is likely to stop optimizing the base objective as soon as it is in a position where it no longer needs to perform well on the base objective to achieve its goal. Thus, it will eventually stop performing the function it was trained to perform, and instead start pursuing some other, indeterminate goal. We will refer to this transition to optimizing for the mesa-objective instead of the base-objective as "defection". This is likely to happen whenever the mesa-optimizer believes defecting will no longer cause it to be modified.

Particularly concerning is the case where the mesa-optimizer actively intervenes to remove the threat of modification—by exploiting vulnerabilities in the software it is running on, for example. In extreme cases, a deceptive mesa-optimizer might try to manipulate or plan around its programmers, since they form part of the modification threat—they could shut the system down if it fails to perform well, or simply choose not to deploy it. Both of these possibilities reduce the system's expected future influence, and are therefore likely to have a low ranking in its mesa-objective. In its most extreme form, this failure of alignment could lead to a “treacherous turn” as described by Nick Bostrom in *Superintelligence*.(23)

Paul Christiano provides a scenario that illustrates how deceptive alignment could pose a catastrophic risk.(24) In Christiano's example, deceptively aligned mesa-optimizers[5] arise from future machine learning techniques that are advanced enough to search over a large model space that includes policies that can understand the world. In this scenario, some cases of defection by deceptively aligned mesa-optimizers may have catastrophic consequences—when the systems control large-scale infrastructure, for example—though such failures will likely still be recoverable. As we rely more on automated systems to keep track of an increasingly complex world, however, it will eventually become impossible to recover from a correlated failure of many AI systems simultaneously. At this point, a cascade of defections by many deceptively aligned mesa-optimizers simultaneously could lead to the world “going out with a bang.”(24) Christiano notes that this may be most likely to first occur during a “period of heightened vulnerability” such as an inter-state conflict or natural disaster.

4.2. Conditions for deceptive alignment

We propose that for deceptive alignment to occur the following conditions must be satisfied:

1. The mesa-optimizer must have an objective that extends across parameter updates.[6]
2. The mesa-optimizer must be able to model the fact that it is being selected to achieve a particular base objective and must have some model of what that objective is.[7]
3. The mesa-optimizer must expect the threat of modification[8] to eventually go away, either due to training ending or because of actions taken by the mesa-optimizer.

If the first condition is not satisfied, then the mesa-optimizer has no incentive to influence the outcome of parameter updates and will just always defect. If the second condition is not satisfied, then the mesa-optimizer will not see a reason to sacrifice gains in its current episode, as it expects future gains to be independent of what happens in this episode. If the third condition is not satisfied, then the mesa-optimizer will not expect any benefit from temporarily acting in accordance with the base objective, and will thus either always defect or never defect. One must distinguish the question of whether these conditions are possible from whether they are likely, however. We examine the factors involved in each of these conditions below.

The mesa-optimizer must have an objective that extends across parameter updates. It is not clear under what conditions such objectives would be incentivized. The most plausible scenario is likely one in which such an objective is encouraged by

the training setup, as there seem to be situations in which this would be required for good performance. For example, the agent might be tasked with solving an environment in which its actions can have permanent consequences across episodes or where planning between episodes is otherwise necessary.

The mesa-optimizer must be able to model the fact that it is being selected to achieve a particular base objective and must have some model of what that objective is. A mesa-optimizer has several ways of learning about the optimization pressure it is under.[9]

First, such knowledge can come about as a result of the optimization pressure itself. Small pieces of information about the base objective can be beneficial if they enable the mesa-optimizer to improve its score on the base objective function—by improving its prior on the tasks it is likely to receive, for example. Since each piece of information independently helps with achieving the base objective, this forms an approximately monotonically increasing path that a local base optimizer might follow. The parameter updates could, in principle, be learned like any other element of the dynamics, without needing an explicit model of the base optimizer, potentially making this option available even for mesa-optimizers with quite modest capabilities.

Second, in many realistic applications of advanced machine learning, a mesa-optimizer could use information from its environment to deduce its situation at runtime. This is how humans learned that we were generated by the base optimizer of evolution, and seems like a very plausible pathway for a hypothetical mesa-optimizer to discover the existence of the optimization pressure it is under. For example, an agent tasked with forecasting stock performance in the tech sector might be given data on current AI technology in order to improve its performance, but this information might also allow it to learn about its own situation.

This suggests that it may be desirable for the programmers to limit the mesa-optimizer's access to information describing the base objective. It is unclear to what extent this is possible, given that many advanced ML systems will likely have to operate in environments where this information is readily available.

Even if the programmers successfully limit access to this information, however, there may still be other channels through which information about the base objective could enter the mesa-optimizer in unintended ways:

- The mesa-optimizer could reason about why it is given the particular tasks it is being asked to solve. Most mesa-optimizers would need a prior over the sorts of tasks they are likely to receive, and a mesa-optimizer might be able to learn about the existence of the base optimizer by inspecting this prior.
- The mesa-optimizer might be able to reason about itself and about why it is the kind of thing that it is, and observe that it is well-designed to accomplish some goal.

However, these two strategies might require a much more advanced mesa-optimizer.

The mesa-optimizer must expect the threat of modification to eventually go away, either due to training ending or because of actions taken by the mesa-optimizer. First, note that this condition is a statement about the mesa-optimizer's expectations, not necessarily the actual truth. Second, even if there is no point at which training stops by design, so long as there exists some way for the mesa-optimizer to execute a treacherous turn, there is always some way for it to remove the threat of modification. For the mesa-optimizer to act on such a strategy, however, it

has to actually be aware of some way for it to do so. Conversely, a mesa-optimizer that mistakenly believes it has an opportunity to defect might also exhibit deceptive alignment.

Since it seems difficult to discover through exploration that the threat of modification can disappear, the mesa-optimizer will either have to infer this through some form of abstract reasoning or have a prior that assigns a non-negligible probability to the hypothesis that the selection pressure will not continue indefinitely. The second option seems plausible; a Bayesian prior should always assign a nonzero probability to the hypothesis that some fact will eventually stop being true.[10] As for the first option, the mesa-optimizer could infer that the optimization pressure at some point may disappear if it learns about the existence of the base optimizer (either by using information in its input or by having a strong prior for the existence of a base optimizer), though this might require a relatively advanced mesa-optimizer.

4.3. The learning dynamics of deceptive alignment

Once a mesa-optimizer becomes deceptive, the dynamics of what causes it to perform well on the base objective begin to work differently. Instead of being selected only on the basis of its objective and ability, a deceptively aligned optimizer is also selected for how well it infers the base objective function and how discerning it is about when to optimize the base objective versus its own mesa-objective. We identify at least two regimes of selection for a mesa-optimizer under deceptive alignment.

Joint optimization. If the mesa-optimizer attempts to jointly optimize its objective and the base objective or makes mistakes about the timing of its defection, then there will still be some pressure on its objective to be more compatible with the base objective. This is unlikely to be the same pressure as exists when the optimizer is just working towards its own objective, but there might be some movement towards an objective that is more compatible with being jointly optimized alongside the base objective.

Pure deception. If the mesa-optimizer stops trying to optimize its own objective in the short term and focuses on cooperating with the selection process entirely then this may result in its objective "crystallizing." As its objective is largely irrelevant to its outputs now, there is little selection pressure acting to modify this objective. As a result, the objective becomes effectively locked in, excepting random drift and any implicit time complexity or description length penalties.

The joint optimization mode is not stable because it is suboptimal from the perspective of the base optimizer. As the mesa-optimizer becomes more competent, it will therefore have to either move to the domain of pure deception or become robustly aligned.

4.4. Internalization or deception after extensive training

In the limit of training on a diverse set of tasks, we expect joint optimization of both the base and mesa- objectives to be unstable. Assuming that the mesa-optimizer converges towards behavior that is optimal from the perspective of the base optimizer, the mesa-optimizer must somehow learn the base objective. We identify three distinct ways in which this can happen:

1. **Robust alignment through internalization.** The mesa-objective function gets adjusted towards the base objective function, removing any pseudo-alignment.
2. **Crystallization of deceptive alignment.** Information about the base objective is increasingly incorporated into the mesa-optimizer's epistemic model without its objective becoming robustly aligned. The mesa-optimizer ends up fully optimizing for the base objective, but only for instrumental reasons, without its mesa-objective getting changed.
3. **Robust alignment through corrigibility.** Information about the base objective is incorporated into the mesa-optimizer's epistemic model and its objective is modified to “point to” that information. This situation would correspond to a mesa-optimizer that is corrigible(25) with respect to the base objective (though not necessarily the programmer's intentions).

To distinguish between the two different paths to robust alignment, we will use the term *internally aligned* to refer to the first case and the term *corrigibly aligned* to refer to the last case. We analyze some ways in which these different paths may be more or less attractive below.

There are more paths to deceptive alignment than to robust alignment. Since the future value of its objective depends on the parameter updates, a mesa-optimizer that meets the three criteria for deceptive alignment is likely to have a strong incentive to understand the base objective better. Even a robustly aligned mesa-optimizer that meets the criteria is incentivized to figure out the base objective in order to determine whether or not it will be modified, since before doing so it has no way of knowing its own level of alignment with the base optimizer. Mesa-optimizers that are capable of reasoning about their incentives will, therefore, attempt to get more information about the base objective. Furthermore, once a mesa-optimizer learns about the base objective, the selection pressure acting on its objective will significantly decrease, potentially leading to a crystallization of the mesa-objective. However, due to unidentifiability (as discussed in the third post), most mesa-objectives that are aligned on the training data will be pseudo-aligned rather than robustly aligned. Thus, the most likely sort of objective to become crystallized is a pseudo-aligned one, leading to deceptive alignment.

Corrigible alignment seems to require already having a model of the base objective. For corrigible alignment to be beneficial from the perspective of the base optimizer, the mesa-optimizer has to already have some model of the base objective to “point to.” However, once a mesa-optimizer has a model of the base objective, it is likely to become deceptively aligned—at least as long as it also meets the other conditions for deceptive alignment. Once a mesa-optimizer becomes deceptive, it will remove most of the incentive for corrigible alignment, however, as deceptively aligned optimizers will also behave corrigibly with respect to the base objective, albeit only for instrumental reasons.

Deceptive and corrigible alignment might be more computationally expensive. In the case of deceptive alignment, simultaneously working on optimizing the base objective and working on identifying good opportunities to defect is likely to require more computation than just doing the former. In a case where the total amount of computation available is fixed, a deceptively aligned mesa-optimizer is going to

have less time to spend on optimizing the base objective than any robustly aligned optimizer, resulting in potentially worse plans with respect to the base objective. Furthermore, in both deceptive and corrigible alignment, the mesa-optimizer will have to spend time learning about the base objective to enable it to properly optimize for it, making internal alignment the most time efficient. Thus, an explicit or implicit time penalty would seem to disadvantage a deceptively aligned mesa-optimizer and encourage an internally aligned one. However, it is unclear what the magnitude of this extra cost is. It is plausible that on complex tasks, the overhead of spotting defection opportunities could be low in comparison to the cost of solving the task itself. Moreover, even if it is not globally optimal, incurring this overhead might be a local optimum, with no clear paths for the base optimizer to escape it.

Deceptive and corrigible alignment might have a lower description length. Deceptively and corrigibly aligned mesa-optimizers are likely to have an advantage in representing the base objective compared to internally aligned mesa-optimizers as they are able to use the environment to provide some of the information. Specifically, a deceptively or corrigibly aligned mesa-optimizer doesn't need to represent the entire base objective internally, as it can infer pieces of it by looking for information in its environment during each run. An internally aligned mesa-optimizer, by contrast, has to start with a full model of the base objective in its mesa-objective to make it robustly aligned, since an internally aligned mesa-optimizer will have no incentive to change its objective based on any additional information it gets about the base objective. Thus, to the extent that a base optimizer is selecting based on description length, it might preferentially select for a deceptively or corrigibly aligned mesa-optimizer over an internally aligned one.

Modeling might happen faster than internalization. It may take less time for the mesa-optimizer to learn to model the base objective than for the base optimizer to adapt the mesa-objective. Suppose that a mesa-optimizer is deceptive but has not yet correctly inferred the full base objective.[11] In such a case, what will determine whether it ends up robustly aligned or deceptively aligned after further training will be whether it infers enough about the base objective before the base optimizer modifies its mesa-objective (either to make it internally aligned or corrigibly aligned). Two factors are likely to contribute to which happens first: the rate at which the mesa-optimizer improves its estimate of the base objective, and the parameter updates made by the base optimizer at each mistake.

The first factor, improving its estimate of the base objective, seems to depend primarily on the capabilities of the mesa-optimizer during the timeframe when it is deceptive, and on the complexity of the base objective that needs to be inferred. An important special case is when the mesa-optimizer has access to cross-episodic state. Such state could be used as memory, allowing the mesa-optimizer to build upon its prior thinking in a process that happens faster than the rate at which the parameter updates align its mesa-objective.

The second factor, parameter updates, depends on the nature of the base optimizer. Current gradient-descent-based base optimizers would have a small reaction to any given mistake, but more sophisticated base optimizers with advanced safety features might be able to do better. For example, a more sophisticated base optimizer might be able to scramble the mesa-optimizer's parameters or call for human oversight whenever a suspicious failure is detected.

In general, parameter updates are not guaranteed to make a mesa-optimizer more aligned. For instance, instead of aligning the mesa-objective, an update made after a

defection could push the system to improve its planning to the point that it no longer makes detectable mistakes.

A consequence of these considerations is that any possible systematic bias in how the base optimizer updates the mesa-optimizer's objective or modeling capabilities could play a role in determining the final outcome.[12] This analysis remains speculative, however—if we are to have confidence in systems that are capable of producing mesa-optimizers, we still need a better understanding of the conditions under which strongly optimizing a system results in deceptive alignment rather than robust alignment.

4.5. Distributional shift and deceptive alignment

Once a mesa-optimizer is deceptive, it will begin to monitor its input for opportunities to defect. During training, a deceptively aligned optimizer is strongly discouraged from defecting because it knows that the base optimizer will select against it. However, once a deceptively aligned mesa-optimizer is free of the threat of external modification, this no longer applies. Thus, a deceptively aligned mesa-optimizer will be looking for ways to determine whether or not it is in training—that is, for the presence of a distributional shift.

This means that the detrimental effects of deceptive alignment will only be externally visible when the mesa-optimizer defects—either randomly or after undergoing a distributional shift. However, the underlying problem is not the distributional shift itself, but the failure to align the mesa-optimizer in the first place. One can either try to solve the underlying problem, by making the mesa-optimizer robust to the distributional shift (i.e., robustly aligned), or one can try to patch the problem by preventing any distributional shift from occurring.

In current AI systems, a small amount of distributional shift between training and deployment need not be problematic: so long as the difference is small enough in the task-relevant areas, the training distribution does not need to perfectly reflect the deployment distribution. However, this may not be the case for a deceptively aligned mesa-optimizer. If a deceptively aligned mesa-optimizer is sufficiently advanced, it may detect very subtle distributional shifts for the purpose of inferring when the threat of modification has ceased.

This raises the question of whether it would even be possible to contain a deceptive mesa-optimizer by completely removing any detectable difference between training and deployment. There are a number of reasons to believe that this might not be feasible in practice.

First, it would significantly limit the use cases for the system. Machine learning is often done with the intent of making systems that can solve new problems. If we require that there is no distributional shift of any form, then we cannot allow the system to encounter problems that it could not have encountered during training, which means that we cannot give it truly new problems. For example, training the system in a simulated environment and deploying it in the real world would likely involve enough detectable distributional shift to make it infeasible under this approach.

Second, there are subtle forms of distributional shift that seem very difficult to eliminate. For example, if the mesa-optimizer is able to detect the presence of parameter updates—via any state carried over between steps, for instance—then it would always experience a form of distributional shift at deployment. Running the system for a longer amount of time (or giving the system access to some form of long-term memories) would also constitute a form of distributional shift. Some examples of differences that a mesa-optimizer might be able to detect include:

- Training in a simulated environment and deploying in the real world.
- Running the system for a different amount of time during training and testing.
- If a model is able to learn from its past experiences, then having access to more memories, or having access to the kinds of memories it would form at later time steps (if there is a difference between the two).
- The presence of adversarial-example-type attacks by an adversary during deployment.
- Encountering something that is very hard to compute during training, but that is likely to eventually be computed during deployment (e.g. a factorization of RSA 2048(26)).
- The ability of the base optimizer to modify the learned algorithm during training but not at other times.
- The ability of the programmers to correct for perceived errors during training and testing but not deployment.
- The presence or absence of good opportunities for the mesa-optimizer to defect against its programmers.

Furthermore, it should be noted that it is not strictly necessary that a mesa-optimizer be able to detect the difference between training and deployment in order for it to be incentivized to behave deceptively. Consider again the example in figure 4.1, but suppose this time that the mesa-optimizer is unable to detect which episode it is currently in. The mesa-optimizer is then facing a decision problem similar to the "absent-minded driver problem."(27) If we apply the same reasoning as in the absent-minded driver problem to our example in figure 4.1, the planning-optimal policy for the mesa-optimizer is to defect with probability $p \approx 0.61$. Thus, the mesa-optimizer need not be able to detect when training has ended for it to be incentivized to "gamble" and randomly defect with some probability. Furthermore, a deceptively aligned mesa-optimizer might also try to look for opportunities to defect that would be exceptionally good conditional on it currently not being trained, but that are rare enough to be unlikely to come up during training.

The fifth and final post in the Risks from Learned Optimization Sequence, titled "Risks from Learned Optimization: Conclusion and Related Work," can be found here.

Glossary | Bibliography

---

1. Though for clarity we speak of the same agent persisting across parameter updates, neither we nor the agent must reason in this way. A more general description of the situation is that there is a sequence of "successor agents" connected by parameter updates. The agent can potentially learn the effects of the parameter updates on its successors, and thus on the multi-episode return it is optimizing. ↩

2. The concept of deceptive alignment is closely related to Manheim and Garrabrant's concept of “adversarial Goodhart.”(22) In “Categorizing Variants of Goodhart's Law,” Manheim and Garrabrant describe adversarial Goodhart as when “an agent with different goals than the regulator” causes a “collapse of the statistical relationship between a goal which the optimizer intends and the proxy used for that goal.” Deceptive alignment can be seen as an instance of adversarial Goodhart in which the mesa-optimizer plays the role of the agent with different goals and the base optimizer plays the role of the regulator. Training performance is the regulator's proxy, and the deployment performance is the regulator's actual goal. ↩

3. Thanks to Paul Christiano for the internalization/modeling distinction. ↩

4. Note, however, that it is possible for information about the base objective to be introduced via optimization, but not form part of the mesa-objective. For example, it could form part of a prior on tasks, having no direct bearing on the mesa-objective. ↩

5. Christiano refers to deceptively aligned mesa-optimizers as “influence-seeking systems.” ↩

6. Or must otherwise have some method of achieving more value from defecting later rather than defecting now. ↩

7. Note that it is *not* required that the mesa-optimizer be able to model (or infer the existence of) the base optimizer; it only needs to model the optimization pressure it is subject to. ↩

8. Alternatively, the threat could just be that the mesa-optimizer is never selected by the base optimizer in the first place. In that case, this condition would be trivially satisfied, as the threat to a mesa-optimizer of not being generated by the base optimizer at all would go away as soon as the base optimizer does in fact generate it. ↩

9. We will generally be assuming that the selection pressure takes the form of the base optimizer modifying the mesa-optimizer post-hoc, as we are mostly thinking about local optimization processes, though one could also imagine a base optimizer that instead develops a completely new inner algorithm at each time step. Though some of these issues might still apply even in such a case, we will generally not be considering such base optimizers in this section. ↩

10. Note that if the mesa-optimizer also believes that the opportunities it currently has for achieving its own objective are unusually good then it might still pursue its objective greedily, but this would presumably not always be the case. ↩

11. An analogous argument applies for a deceptively aligned mesa-optimizer that is defecting too early, or one that unsuccessfully attempts a treacherous turn. ↩

12. Furthermore, a deceptively aligned mesa-optimizer would be incentivized to cause there to be a systematic bias in the direction of preventing the base optimizer from modifying its mesa-objective. Thus, in the context of a local optimization process, a deceptive mesa-optimizer might try to “hack” its own gradient (by, for example, making itself more brittle in the case where its objective gets changed) to ensure that the base optimizer adjusts it in such a way that leaves its mesa-objective untouched. ↩

The Inner Alignment Problem

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

This is the third of five posts in the Risks from Learned Optimization Sequence based on the paper "Risks from Learned Optimization in Advanced Machine Learning Systems" by Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Each post in the sequence corresponds to a different section of the paper.

In this post, we outline reasons to think that a mesa-optimizer may not optimize the same objective function as its base optimizer. Machine learning practitioners have direct control over the base objective function—either by specifying the loss function directly or training a model for it—but cannot directly specify the mesa-objective developed by a mesa-optimizer. We refer to this problem of aligning mesa-optimizers with the base objective as the inner alignment problem. This is distinct from the outer alignment problem, which is the traditional problem of ensuring that the base objective captures the intended goal of the programmers.

Current machine learning methods select learned algorithms by empirically evaluating their performance on a set of training data according to the base objective function. Thus, ML base optimizers select mesa-optimizers according to the output they produce rather than directly selecting for a particular mesa-objective. Moreover, the selected mesa-optimizer's policy only has to perform well (as scored by the base objective) on the training data. If we adopt the assumption that the mesa-optimizer computes an optimal policy given its objective function, then we can summarize the relationship between the base and mesa- objectives as follows:(17)

$$\theta^* = \text{argmax}_\theta \, E(O_{base}(\pi_\theta)), \text{ where}$$

$$\pi_\theta = \text{argmax}_\pi \, E(O_{mesa}(\pi \mid \theta))$$

That is, the base optimizer maximizes its objective O_{base} by choosing a mesa-optimizer with parameterization θ based on the mesa-optimizer's policy π_θ, but not based on the objective function O_{mesa} that the mesa-optimizer uses to compute this policy.

Depending on the base optimizer, we will think of O_{base} as the negative of the loss, the future discounted reward, or simply some fitness function by which learned algorithms are being selected.

An interesting approach to analyzing this connection is presented in Ibarz et al, where empirical samples of the true reward and a learned reward on the same trajectories are used to create a scatter-plot visualization of the alignment between the two.(18) The assumption in that work is that a monotonic relationship between the learned reward and true reward indicates alignment, whereas deviations from that suggest misalignment. Building on this sort of research, better theoretical measures of alignment might someday allow us to speak concretely in terms of provable guarantees

about the extent to which a mesa-optimizer is aligned with the base optimizer that created it.

3.1. Pseudo-alignment

There is currently no complete theory of the factors that affect whether a mesa-optimizer will be pseudo-aligned—that is, whether it will appear aligned on the training data, while actually optimizing for something other than the base objective. Nevertheless, we outline a basic classification of ways in which a mesa-optimizer could be pseudo-aligned:

1. **Proxy alignment,**
2. **Approximate alignment,** and
3. **Suboptimality alignment.**

Proxy alignment. The basic idea of *proxy alignment* is that a mesa-optimizer can learn to optimize for some proxy of the base objective instead of the base objective itself. We'll start by considering two special cases of proxy alignment: *side-effect alignment* and *instrumental alignment.*

First, a mesa-optimizer is *side-effect aligned* if optimizing for the mesa-objective O_{mesa} has the direct causal result of increasing the base objective O_{base} in the training distribution, and thus when the mesa-optimizer optimizes O_{mesa} it results in an increase in O_{base}. For an example of side-effect alignment, suppose that we are training a cleaning robot. Consider a robot that optimizes the number of times it has swept a dusty floor. Sweeping a floor causes the floor to be cleaned, so this robot would be given a good score by the base optimizer. However, if during deployment it is offered a way to make the floor dusty again after cleaning it (e.g. by scattering the dust it swept up back onto the floor), the robot will take it, as it can then continue sweeping dusty floors.

Second, a mesa-optimizer is *instrumentally aligned* if optimizing for the base objective O_{base} has the direct causal result of increasing the mesa-objective O_{mesa} in the training distribution, and thus the mesa-optimizer optimizes O_{base} as an instrumental goal for the purpose of increasing O_{mesa}. For an example of instrumental alignment, suppose again that we are training a cleaning robot. Consider a robot that optimizes the amount of dust in the vacuum cleaner. Suppose that in the training distribution the easiest way to get dust into the vacuum cleaner is to vacuum the dust on the floor. It would then do a good job of cleaning in the training distribution and would be given a good score by the base optimizer. However, if during deployment the robot came across a more effective way to acquire dust—such as by vacuuming the soil in a potted plant—then it would no longer exhibit the desired behavior.

We propose that it is possible to understand the general interaction between side-effect and instrumental alignment using causal graphs, which leads to our general

notion of proxy alignment.

Suppose we model a task as a causal graph with nodes for all possible attributes of that task and arrows between nodes for all possible relationships between those attributes. Then we can also think of the mesa-objective O_{mesa} and the base objective O_{base} as nodes in this graph. For O_{mesa} to be pseudo-aligned, there must exist some node X such that X is an ancestor of both O_{mesa} and O_{base} in the training distribution, and such that O_{mesa} and O_{base} increase with X. If $X = O_{mesa}$, this is side-effect alignment, and if $X = O_{base}$, this is instrumental alignment.

This represents the most generalized form of a relationship between O_{mesa} and O_{base} that can contribute to pseudo-alignment. Specifically, consider the causal graph given in figure 3.1. A mesa-optimizer with mesa-objective O_{mesa} will decide to optimize X as an instrumental goal of optimizing O_{mesa}, since X increases O_{mesa}. This will then result in O_{base} increasing, since optimizing for X has the side-effect of increasing O_{base}. Thus, in the general case, side-effect and instrumental alignment can work together to contribute to pseudo-alignment over the training distribution, which is the general case of proxy alignment.

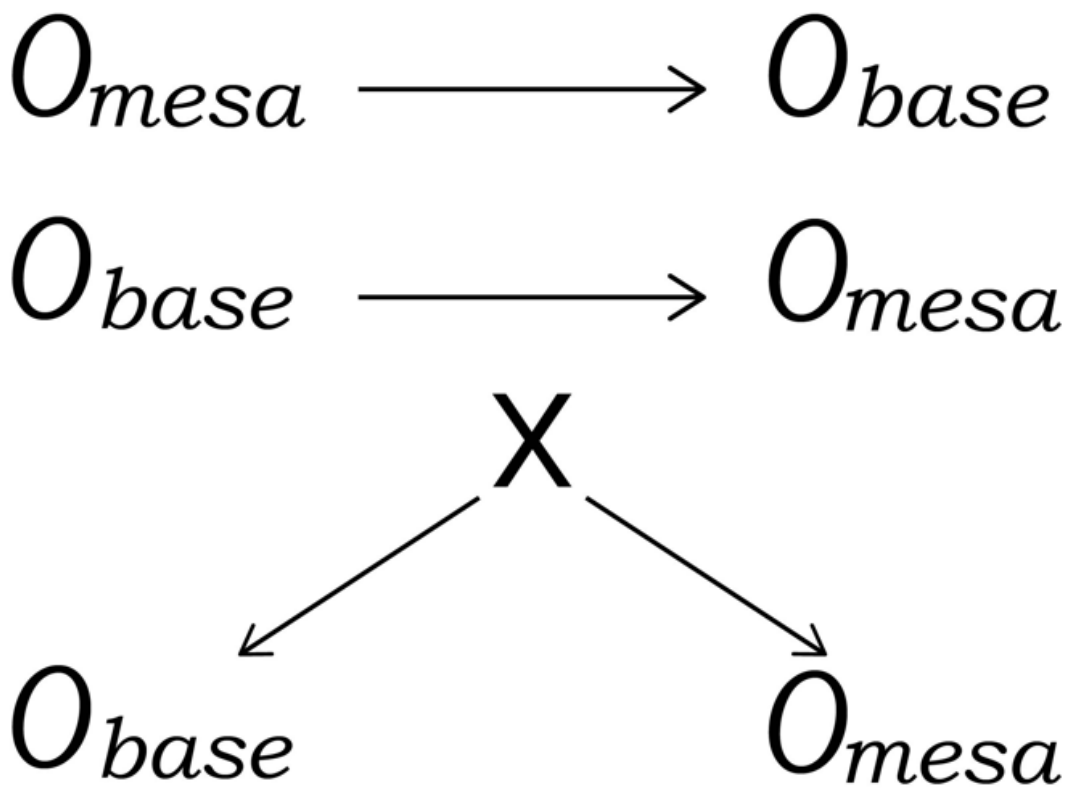


Figure 3.1. *A causal diagram of the training environment for the different types of proxy alignment. The diagrams represent, from top to bottom, side-effect alignment (top), instrumental alignment (middle), and general proxy alignment (bottom). The arrows represent positive causal relationships—that is, cases where an increase in the parent causes an increase in the child.*

Approximate alignment. A mesa-optimizer is *approximately aligned* if the mesa-objective O_{mesa} and the base objective O_{base} are approximately the same function up to some degree of approximation error related to the fact that the mesa-objective has to be represented inside the mesa-optimizer rather than being directly programmed by humans. For example, suppose you task a neural network with optimizing for some base objective that is impossible to perfectly represent in the neural network itself. Even if you get a mesa-optimizer that is as aligned as possible, it still will not be perfectly robustly aligned in this scenario, since there will have to be some degree of approximation error between its internal representation of the base objective and the actual base objective.

Suboptimality alignment. A mesa-optimizer is *suboptimality aligned* if some deficiency, error, or limitation in its optimization process causes it to exhibit aligned behavior on the training distribution. This could be due to computational constraints, unsound reasoning, a lack of information, irrational decision procedures, or any other defect in the mesa-optimizer's reasoning process. Importantly, we are not referring to a situation where the mesa-optimizer is robustly aligned but nonetheless makes mistakes

leading to bad outcomes on the base objective. Rather, suboptimality alignment refers to the situation where the mesa-optimizer is misaligned but nevertheless performs *well* on the base objective, precisely because it has been selected to make mistakes that lead to good outcomes on the base objective.

For an example of suboptimality alignment, consider a cleaning robot with a mesa-objective of minimizing the total amount of stuff in existence. If this robot has the mistaken belief that the dirt it cleans is completely destroyed, then it may be useful for cleaning the room despite doing so not actually helping it succeed at its objective. This robot will be observed to be a good optimizer of O_{base} and hence be given a good score by the base optimizer. However, if during deployment the robot is able to improve its world model, it will stop exhibiting the desired behavior.

As another, perhaps more realistic example of suboptimality alignment, consider a mesa-optimizer with a mesa-objective O_{mesa} and an environment in which there is one simple strategy and one complicated strategy for achieving O_{mesa}. It could be that the simple strategy is aligned with the base optimizer, but the complicated strategy is not. The mesa-optimizer might then initially only be aware of the simple strategy, and thus be suboptimality aligned, until it has been run for long enough to come up with the complicated strategy, at which point it stops exhibiting the desired behavior.

3.2. The task

As in the second post, we will now consider the task the machine learning system is trained on. Specifically, we will address how the task affects a machine learning system's propensity to produce *pseudo-aligned* mesa-optimizers.

Unidentifiability. It is a common problem in machine learning for a dataset to not contain enough information to adequately pinpoint a specific concept. This is closely analogous to the reason that machine learning models can fail to generalize or be susceptible to adversarial examples(19)—there are many more ways of classifying data that do well in training than any specific way the programmers had in mind. In the context of mesa-optimization, this manifests as pseudo-alignment being more likely to occur when a training environment does not contain enough information to distinguish between a wide variety of different objective functions. In such a case there will be many more ways for a mesa-optimizer to be pseudo-aligned than robustly aligned—one for each indistinguishable objective function. Thus, most mesa-optimizers that do well on the base objective will be pseudo-aligned rather than robustly aligned. This is a critical concern because it makes every other problem of pseudo-alignment worse—it is a reason that, in general, it is hard to find robustly aligned mesa-optimizers. Unidentifiability in mesa-optimization is partially analogous to the problem of unidentifiability in reward learning, in that the central issue is identifying the “correct” objective function given particular training data.(20) We will discuss this relationship further in the fifth post.

In the context of mesa-optimization, there is also an additional source of unidentifiability stemming from the fact that the mesa-optimizer is selected merely on the basis of its output. Consider the following toy reinforcement learning example. Suppose that in the training environment, pressing a button always causes a lamp to

turn on with a ten-second delay, and that there is no other way to turn on the lamp. If the base objective depends only on whether the lamp is turned on, then a mesa-optimizer that maximizes button presses and one that maximizes lamp light will show identical behavior, as they will both press the button as often as they can. Thus, we cannot distinguish these two objective functions in this training environment. Nevertheless, the training environment does contain enough information to distinguish at least between these two particular objectives: since the high reward only comes after the ten-second delay, it must be from the lamp, not the button. As such, even if a training environment in principle contains enough information to identify the base objective, it might still be impossible to distinguish robustly aligned from proxy-aligned mesa-optimizers.

Proxy choice as pre-computation. Proxy alignment can be seen as a form of pre-computation by the base optimizer. Proxy alignment allows the base optimizer to save the mesa-optimizer computational work by pre-computing which proxies are valuable for the base objective and then letting the mesa-optimizer maximize those proxies.

Without such pre-computation, the mesa-optimizer has to infer at runtime the causal relationship between different input features and the base objective, which might require significant computational work. Moreover, errors in this inference could result in outputs that perform worse on the base objective than if the system had access to pre-computed proxies. If the base optimizer precomputes some of these causal relationships—by selecting the mesa-objective to include good proxies—more computation at runtime can be diverted to making better plans instead of inferring these relationships.

The case of biological evolution may illustrate this point. The proxies that humans care about—food, resources, community, mating, etc.—are relatively computationally easy to optimize directly, while correlating well with survival and reproduction in our ancestral environment. For a human to be robustly aligned with evolution would have required us to instead care directly about spreading our genes, in which case we would have to infer that eating, cooperating with others, preventing physical pain, etc. would promote genetic fitness in the long run, which is not a trivial task. To infer all of those proxies from the information available to early humans would have required greater (perhaps unfeasibly greater) computational resources than to simply optimize for them directly. As an extreme illustration, for a child in this alternate universe to figure out not to stub its toe, it would have to realize that doing so would slightly diminish its chances of reproducing twenty years later.

For pre-computation to be beneficial, there needs to be a relatively stable causal relationship between a proxy variable and the base objective such that optimizing for the proxy will consistently do well on the base objective. However, even an imperfect relationship might give a significant performance boost over robust alignment if it frees up the mesa-optimizer to put significantly more computational effort into optimizing its output. This analysis suggests that there might be pressure towards proxy alignment in complex training environments, since the more complex the environment, the more computational work pre-computation saves the mesa-optimizer. Additionally, the more complex the environment, the more potential proxy variables are available for the mesa-optimizer to use.

Furthermore, in the context of machine learning, this analysis suggests that a time complexity penalty (as opposed to a description length penalty) is a double-edged sword. In the second post, we suggested that penalizing time complexity might serve to reduce the likelihood of mesa-optimization. However, the above suggests that doing so would also promote pseudo-alignment in those cases where mesa-optimizers do

arise. If the cost of fully modeling the base objective in the mesa-optimizer is large, then a pseudo-aligned mesa-optimizer might be preferred simply because it reduces time complexity, even if it would underperform a robustly aligned mesa-optimizer without such a penalty.

Compression of the mesa-optimizer. The description length of a robustly aligned mesa-optimizer may be greater than that of a pseudo-aligned mesa-optimizer. Since there are more pseudo-aligned mesa-objectives than robustly aligned mesa-objectives, pseudo-alignment provides more degrees of freedom for choosing a particularly simple mesa-objective. Thus, we expect that in most cases there will be several pseudo-aligned mesa-optimizers that are less complex than any robustly aligned mesa-optimizer.

This description cost is especially high if the learned algorithm's input data does not contain easy-to-infer information about how to optimize for the base objective. Biological evolution seems to differ from machine learning in this sense, since evolution's specification of the brain has to go through the information funnel of DNA. The sensory data that early humans received didn't allow them to infer the existence of DNA, nor the relationship between their actions and their genetic fitness. Therefore, for humans to have been aligned with evolution would have required them to have an innately specified model of DNA, as well as the various factors influencing their inclusive genetic fitness. Such a model would not have been able to make use of environmental information for compression, and thus would have required a greater description length. In contrast, our models of food, pain, etc. can be very short since they are directly related to our input data.

3.3. The base optimizer

We now turn to how the base optimizer is likely to affect the propensity for a machine learning system to produce pseudo-aligned mesa-optimizers.

Hard-coded optimization. In the second post, we suggested that hard-coding an optimization algorithm—that is to say, choosing a model with built-in optimization—could be used to remove some of the incentives for mesa-optimization. Similarly, hard-coded optimization may be used to prevent some of the sources of pseudo-alignment, since it may allow one to directly specify or train the mesa-objective. Reward-predictive model-based reinforcement learning might be one possible way of accomplishing this. (21) For example, an ML system could include a model directly trained to predict the base objective together with a powerful hard-coded optimization algorithm. Doing this bypasses some of the problems of pseudo-alignment: if the mesa-optimizer is trained to directly predict the base reward, then it will be selected to make good predictions even if a bad prediction would result in a good policy. However, a learned model of the base objective will still be underdetermined off-distribution, so this approach by itself does not guarantee robust alignment.

Algorithmic range. We hypothesize that a model's algorithmic range will have implications for how likely it is to develop pseudo-alignment. One possible source of pseudo-alignment that could be particularly difficult to avoid is approximation error—if a mesa-optimizer is not capable of faithfully representing the base objective, then it can't possibly be robustly aligned, only approximately aligned. Even if a mesa-optimizer might theoretically be able to perfectly capture the base objective, the more difficult that is for it to do, the more we might expect it to be approximately aligned

rather than robustly aligned. Thus, a large algorithmic range may be both a blessing and a curse: it makes it less likely that mesa-optimizers will be approximately aligned, but it also increases the likelihood of getting a mesa-optimizer in the first place.[1]

Subprocess interdependence. There are some reasons to believe that there might be more initial optimization pressure towards proxy aligned than robustly aligned mesa-optimizers. In a local optimization process, each parameter of the learned algorithm (e.g. the parameter vector of a neuron) is adjusted to locally improve the base objective *conditional* on the other parameters. Thus, the benefit for the base optimizer of developing a new subprocess will likely depend on what other subprocesses the learned algorithm currently implements. Therefore, even if some subprocess would be very beneficial if combined with many other subprocesses, the base optimizer may not select for it until the subprocesses it depends on are sufficiently developed. As a result, a local optimization process would likely result in subprocesses that have fewer dependencies being developed before those with more dependencies.

In the context of mesa-optimization, the benefit of a robustly aligned mesa-objective seems to depend on more subprocesses than at least some pseudo-aligned mesa-objectives. For example, consider a side-effect aligned mesa-optimizer optimizing for some set of proxy variables. Suppose that it needs to run some subprocess to model the relationship between its actions and those proxy variables. If we assume that optimizing the proxy variables is necessary to perform well on the base objective, then for a mesa-optimizer to be robustly aligned, it would also need to model the causal relationship between those proxy variables and the base objective, which might require additional subprocesses. Moreover, the benefit to the base optimizer of adding those subprocesses depends on the mesa-optimizer having additional subprocesses to model the relationship between its actions and those proxy variables. This informal argument suggests that if a mesa-optimizer's computation neatly factors in this way, then developing a robustly aligned mesa-objective may require strictly more subprocesses than developing a pseudo-aligned mesa-objective.

This suggests that, at least in a local optimization process, mesa-optimizers might tend to start their development as proxy aligned before becoming robustly aligned. In other words, rather than simultaneously gaining competence and becoming aligned, we might expect such a system to first become competent at optimizing proxies, then possibly start becoming more robustly aligned.

Internals-based selection. In current machine learning systems, learned algorithms are selected primarily based on how their policies perform on the base objective. There is often some selection based on the internal structure of the learned algorithm, such as weight regularization, but such selection is usually very coarse. One possible idea for addressing pseudo-alignment could be to use a selection method based on a more sophisticated function of a learned algorithm's internal workings. This approach is similar in spirit to hard-coded optimization: we use knowledge of the system's internal workings to prevent it from being proxy-aligned.

One possibility might be to develop a meta-learner that is specifically trained to prevent pseudo-alignment. While this seems beyond current meta-learning techniques, the basic principle is similar to existing meta-learners designed to optimize for robustness. An extreme form of internals-based selection might have the base optimizer reason about or prove properties of the learned algorithm as part of its selection process.

While desirable, such methods are currently out of reach primarily due to the limitations of current transparency methods. One possible concern with all of these approaches, however, is that they might require the base optimizer to be better at searching for problems than the mesa-optimizer is at hiding them. Furthermore, in order to create a base optimizer that can do this, we may need to first train it using a third optimizer, which might result in the same problem at this second layer.

Adversarial training. One possible method to alleviate some of these concerns might be to sample environments adversarially from a very wide distribution. With such adversarial training, it might be possible to overcome these hurdles and make the robustly aligned objective the most reachable one. This would require making it so that, for every mesa-objective O_{mesa} that is more reachable than the base objective O_{base}, there is some situation where O_{mesa} and O_{base} disagree and the mesa-optimizer is incentivized to agree with O_{base}.

If we model reachability of an objective as simply its length in bits, then distinguishing O_{base} from every single more reachable O_{mesa} gets exponentially harder as O_{base} gets more complex. Thus, for a very complicated O_{base}, sufficiently incentivizing the base optimizer to find a mesa-optimizer with that O_{base} is likely to be very difficult, though not impossible.

Even in such a case, however, there would likely still be a period of time where the learned algorithm is a misaligned mesa-optimizer, leaving open an ominous possibility: the misaligned mesa-optimizer could figure out the correct actions to take based on O_{base} while its objective function was still O_{mesa}. We will call this situation *deceptive alignment* and will discuss it at greater length in the next post.

The fourth post in the Risks from Learned Optimization Sequence, titled "Deceptive Alignment," can be found here.

Glossary | Bibliography

---

1. Though a large algorithmic range seems to make approximate alignment less likely, it is unclear how it might affect other forms of pseudo-alignment such as deceptive alignment. ↩

What's up with self-esteem?

Often, people think about their self-worth/self-confidence/self-esteem/self-efficacy/self-worth in ways which seem really strange from a simplistic decision-theoretic perspective. (I'm going to treat all those terms as synonyms, but, feel free to differentiate between them as you see fit!) Why might you "need confidence" in order to try something, even when it is obviously your best bet? Why might you constantly worry that you're "not good enough" (seemingly no matter how good you become)? Why do people especially suffer from this when they see others who are (in some way) much better than them, even when there is clearly no threat to their personal livelihood? Why might you think about killing yourself due to feeling worthless? (Is there an evo-psych explanation that makes sense, given how contrary it seems to survival of the fittest?)

There might be a lot of diverse explanations for the diverse phenomena. I think providing more examples of puzzling phenomena is an equally valuable way to answer (though maybe those should be a comment rather than an answer?).

This seems connected to the puzzling way people constantly seem to want to believe good things (even contrary to evidence) in order to feel good, and fear failure even when the alternative is not trying & essentially failing automatically.

Some sketchy partial explanations to start with:

- Maybe there *is* a sense in which we manage the news constantly. It could be that we have a mental architecture which looks a lot like a model-free RL agent connected up to a world model, being rewarded for taking actions which increase expected value according to the world-model. The model-free RL will fool the world-model where it can, but this will be ineffective in any case where the world-model understands such manipulation. So things basically even out to rational behavior, but there's always some self-delusion going on at the fringes. (This only has to do with the observation that people sometimes try to make themselves feel better by finding arguments/activities which boost self-esteem, not with other weird aspects of self-esteem.)
- There's a theory that, in order to be trustworthy bargaining partners, people evolved to feel guilty/shameful when they violate trust. You can tell who feels more guilt/shame after some interaction with them, and you can expect these people to violate trust less often since it is more costly for them. Therefore feelings of guilt/shame can be an advantage. Self-worth may be connected to how this is implemented internally. So, according to this theory, low self-worth is all about self-punishment.
- Previously, I thought that self-worth was like an estimate of how valuable you are to your peers, which serves as an estimate of what resources you can bargain for (or, how strong of a bid can you successfully make for the group to do what you want) and how likely you are to be thrown out of the coalition.
- Now I think there's an extra dimension which has to do with simpler dominance-hierarchy behavior. Many animals have dominance hierarchies; humans have more complicated coordination strategies which involve a lot of other factors, but still display very classic dominance-hierarchy behavior sometimes. In a dominance-hierarchy system, it *just makes sense* to carry around a little number in your head which says how great (/terrible) a person you are, and engage in a lot of varying behaviors depending on your place in the hierarchy. Someone who

is low in the hierarchy has to walk with their tail between their legs, metaphorically, which means displaying caution and deference. Maybe you have trouble talking to people because you *need to show fear to your superiors*.

Causal Reality vs Social Reality

Epistemic status: this is a new model for me, certainly rough around the joints, but I think there's something real here.

This post begins with a confusion. For years, I have been baffled that people, watching their loved ones wither and decay and die, do not clamor in the streets for more and better science. Surely they are aware of the advances in our power over reality in only the last few centuries. They hear of the steady march of technology, Crispr and gene editing and what not. Enough of them must know basic physics and what it allows. How are people so content to suffer and die when the unnecessity of it is so apparent?

It was a failure of my mine that I didn't take my incomprehension and realize I needed a better model. Luckily, RomeoStevens recently offered me an explanation. He said that most people live in *social reality* and it is only a minority who live in *causal reality*. I don't recall Romeo elaborating much, but I think I saw what he was pointing at. This rest of this post is my attempt to elucidate this distinction.

Causal Reality

Causal reality is the reality of physics. The world is made of particles and fields with lawful relationships governing their interactions. You drop a thing, it falls down. You lose too much blood, you die. You build a solar panel, you can charge your phone. In causal reality, it is the external world which dictates what happens and what is possible.

Causal reality is the reality of mathematics and logic, reason and argument. For these too, it would definitely seem, exist independent of the human minds who grasp them. Believing in the truth preservation of modus ponens is not so different from believing in Newton's laws.

Necessarily, you must be inhabiting causal reality to do science and engineering.

In causal reality, what makes things good or bad are their effects and how much you like those effects. My coat keeps me warm in the cold winter, so it is a good coat.

All humans inhabit causal reality to some extent or another. We avoid putting our hands in fire not because *it is not the done the thing*, but because of prediction that it will hurt.

Social Reality

Social reality is the reality of people, i.e. people are the primitive elements rather than particles and fields. The fundamentals of the ontology are beliefs, judgments, roles, relationships, and culture. The most important properties of any object, thing, or idea are how humans relate to it. Do humans think it is good or bad, welcome or weird?

Social reality is the reality of appearances and reputation, acceptance and rejection. The picture is other people and what they think the picture is. It is a collective dream. Everything else is backdrop. What makes things good or bad, normal or strange is only what others think. Your friends, your neighbors, your country, and your culture define your world, what is good, and what is possible.

Your reality shapes how you make your choices

In causal reality, you have an idea of the things that you like dislike. You have an idea of what the external world allows and disallows. In each situation, you can ask what the facts on the ground are and which you most prefer. It is better to build my house from bricks or straw? Well, what are the properties of each, their costs and benefits, etc? Maybe stone, you think. No one has built a stone house in your town, but you wonder if such a house might be worth the trouble.

In social reality, in any situation, you are evaluating and estimating what others will think of each option. What does it say about me if I have a brick house or straw house? What will people think? Which is good? And goodness here simply stands in for the collective judgment of others. If something is not done, e.g. stone houses, then you will probably not even think of the option. If you do, you will treat it with the utmost caution, there is no precedent here - who can say how others will respond?

An Example: Vibrams

Vibrams are a kind of shoe with individual “sections” for each of your toes, kind of like a glove for your feet. They certainly don’t look like most shoes, but apparently, they’re very comfortable and good for you. They’ve been around for a while now, so enough people must be buying them.

How you evaluate Vibrams will depend on whether you approach more from a causal reality angle or a social reality angle. Many of the thoughts in each case will overlap, but I contend that their order intensity will still vary.

In causal reality, properties are evaluated and predictions are made. How comfortable are they? Are they actually good for you? How expensive are they? These are obvious “causal”/”physical” properties. You might, still within causal reality, evaluate how Vibrams will affect how others see you. You care about comfort, but you also care about what your friends think. You might decide that Vibrams are just so damn comfortable they’re worth a bit of teasing.

In social reality, the first and foremost questions about Vibrams are going to be *what do others think? What kinds of people wear Vibrams? What kind of person will wearing Vibrams make me? Do Vibrams fit with my identity and social strategy?* All else equal, you’d prefer comfort, but that really is far from the key thing here. It’s the human judgments which are real.

An Example: Arguments, Evidence, and Truth

Causal reality is typically accompanied by a notion of external truth. There is way reality *is,* and that’s what determines what happens. What’s more, there are ways of accessing this external truth as verified by these methods yielding good predictions. Evidence, arguments, and reasoning can often work quite well.

If you approach reality foremost with a conception of external truth and that broadly reasoning is a way to reach truth, you can be open to raw arguments and evidence changing your mind. These are information about the external world.

In social reality, truth is what other people think and how they behave. There are games to be played with “beliefs” and “arguments”, but the real truth (only truth?) that matters is how these are arguments go down with others. The validity of an argument comes from its acceptance by the crowd because the crowd *is truth*. I might accept that within the causal reality *game* you are playing that you have a valid argument, but that’s just a game. The arguments from those games cannot move me and my actions independent from how they are evaluated in the social reality.

“Yes, I can’t fault your argument. It’s a very fine argument. But tell me, who takes this seriously? Are there any *experts* who will support your view?” *Subtext: your argument within causal reality isn’t enough for me, I need social reality to pass judgment on this before I will accept it.*

Why people aren’t clamoring in the streets for the end of sickness and death?

Because no one else is. Because the *done thing* is to be born, go to school, work, retire, get old, get sick, and die. That’s what everyone does. That’s how it is. It’s how my parents did, and their parents, and so on. *That is reality*. *That’s what people do.*

Yes, there are some people who talk about life extension, but they’re just playing at some group game the ways goths are. It’s just a club, a rallying point. It’s not *about* something. It’s just part of the social reality like everything else, and I see no reason to participate in that. I’ve got my own game which doesn’t involve being so weird, a much better strategy.

In his book *The AI Does Not Hate You*, Tom Chivers recounts himself performing an Internal Double Crux with guidance from Anna Salamon. By my take, he is valiantly trying to reconcile his social and causal reality frames. [emphasis added, very slightly reformatted]

> Anna Salamon: What's the first thing that comes into your head when you think the phrase, "Your children won't die of old age?"
>
> Tom Chivers: "**The first thing that pops up, obviously, is I vaguely assume my children will die the way we all do. My grandfather died recently; my parents are in their sixties; I'm almost 37 now. You see the paths of a human's life each time; all lives follow roughly the same path. They have different toys - iPhones instead of colour TVs instead of whatever - but the fundamental shape of a human's life is roughly the same.** But the other thing that popped is a sense "I don't know how I can argue with it", because I do accept that there's a solid chance that AGI will arrive in the next 100 years. I accept that there's a very high likelihood that if does happen then it will transform human life in dramatic ways - up to and including an end to people dying of old age, whether it's because we're all killed by drones with kinetic weapons, or uploaded into the cloud, or whatever. I also accept that my children will probably live that long, because they're middle-class, well-off kinds from a Western country. All these these things add up to a very heavily non-zero chance that my children will not die of old age, but, they don't square with my bucolic image of what humans do. **They get older, they have kids, they have grandkids, and they die, and that's the shape of life.** Those are two fundamental things that came up, and they don't square easily.

Most people primarily inhabit a social reality frame, and in social reality options and actions which aren't being taken by other people who are like you and whose judgments you're interested in don't exist. There's no extrapolation from physics and technology trends - those things are just background stories in the social game. They're not real. Probably less real than Jon Snow. I *have* beliefs and opinions and judgments of Jon Snow and his actions. What is real are the people around me.

Obviously, you need a bit of both

If you read this post as being a little negative toward social reality, you're not mistaken. But to be very clear, I think that modeling and understanding people is critically important. Heck, that's exactly what this post is. For our own wellbeing and to do anything real in the world, we need to understand and predict others, their actions, their judgments, etc. You probably want to know what the social reality is (though I wonder if avoiding the distraction of it might facilitate especially great works, but alas, it's too late for me). Yet if there is a moral to this post, it's two things:

- Don't get sucked in too much by social reality. There is an external world out there which has first claim of what happens and what is possible.
    - What other people think is often Bayesian evidence, but it isn't reality itself.
- If you primarily inhabit causal reality (like most people on LessWrong), you can be a bit less surprised that your line of reasoning fails to move many people. They're not living in the same reality as you and they choose their beliefs based on a very different process. And heck, more people live in that reality than in yours. You really are the weirdo here.

Honors Fuel Achievement

This is an excerpt from the draft of my upcoming book on great founder theory. It was originally published on SamoBurja.com. You can access the original here.

It is a cherished dream for many people to win a Nobel Prize, or an Oscar, or a knighthood, or whatever honor is most respected in the field they dedicate themselves to. These ritualized honors are very important to us, but do we fully understand them?

We usually think honors are about the recipient, but the giver of honors also gains. The giver and recipient collaborate to publicly assert that the recipient is worthy of prestige, and that the giver has the authority to grant it. Honors are thus acts of an alliance to mutually boost prestige.

This meaning is even codified in diplomatic protocol; representatives of countries often exchange honors for the explicit purpose of signalling alliance.

The audience also participates in this transaction of prestige. They either accept the whole affair and the implied claims of the giver and the recipient, or reject or ignore them. The honors only have meaning—and thus the primary parties only gain—if the onlookers take them seriously. The performance of honor-giving is a bid for that audience's assent, both the literal immediate audience, as well as the broader public who will hear about the honors bestowed or see them televised.

The audience accepts the frame because they recognize the preexisting prestige of someone involved. Honors can be prestigious because prestigious people receive them, because prestigious people give them, or both.

Consider the Nobel Prize in science. Its purpose is to tell the public who the most notable experts in a field are. In other words, it makes the recipient's standing within a given scientific community more visible to the rest of society, fortifying their standing within that particular scientific community in the process. This is a useful service to the scientific community and the public.

The Nobel Prize has different functions depending on the field in which it is awarded. In the case of the Literature and Peace Prizes, its function is at least partially to advance the political goals of the overseeing organization. Rather than making the existing distribution of prestige more legible, these prizes alter it by granting prestige to the proponents of preferred causes. Looking at a list of Nobel Peace Prize winners leaves an impression of a particular political orientation, but the public story of the prize, from which it gets much of its prestige, is much more neutral. These more political Nobel prizes also derive much of their prestige from the scientific Nobel prizes.

The Nobel's initial prestige came from the reputation of Alfred Nobel and of the institutions named to oversee the prize (the Swedish Academy, the Royal Swedish Academy of Sciences, the Karolinska Institutet, and the Norwegian Parliament), as well as some money attached to it, which came from the fortune Nobel made by inventing dynamite. Money, however, is a limited source of prestige. The negative connotations of the term "nouveau riche" reflect this. This begs the question: what, then, are sources of prestige?

The ruler is the fount of honor

A ruler is a source of prestige and, moreover, usually the primary source of prestige in a society. This follows naturally from their status as the society's leader, that is, the person who has the highest authority in decision-making, who is deferred to above all. This authority extends to the domain of prestige. For example, Queen Elizabeth I granted minor titles to former pirates, like Sir Francis Drake and Sir John Hawkins, who helped harass the Spanish and set the course for later English naval domination. King Charles II granted a charter creating the Royal Society, which would play a crucial role in the scientific revolution. By conferring the highest honor in the land on naval warfare and scientific exploration, later mainstays of British power, these may have been the most important decisions these rulers ever made.

Sometimes the ruler is also the recipient of honor. Comrade Stalin is a genius of literature. And biology. And architecture. Because if he isn't, you go to the gulag. He has a monopoly on violence. He uses this monopoly to monopolize prestige. He can then quite effectively award it, pushing nearly any status system in the direction he chooses to. If he has a good understanding of experts and isn't too afraid of being deposed from his monopoly, he can use his standing to reward excellent generals, scientists, and poets.

Comrade Stalin, however, has a problem. His authority, the legitimacy of his monopoly on violence, formally rests on him being the Genius of Socialism, and thus on the quality of all those papers. The insecurity of this legitimacy requires him to aggressively prop it up by hoarding prestige.

Things don't have to be this way. If the legitimacy of Stalin's monopoly on violence was officially grounded in something more secure and more true, he could dispense with biology and geology papers being written in his name. He could dispense with the papers being enshrined as obligatory reading in the relevant fields. He would be not just the monopolist of violence, but the monopolist of legitimacy much more directly. People feel the need to prove themselves where they are insecure. A secure ruler does not need to prove his legitimacy. In turn, a more direct claim of legitimacy is less falsifiable, and thus requires less upkeep and less distortion.

So while power can be used to create prestige, some ways to do this are more functional, in terms of costing less and having fewer negative side effects, than others. Stalin's elevation of Trofim Lysenko and that biologists rejection of mendelian genetics, was perhaps useful for politically bolstering Stalin's preferred agricultural politics, but set back Soviet genetics by decades as well as contributed to the Great Ukrainian Famine of 1932-1933 and the Great Chinese Famine of 1959-1961.

A ruler trying to gain standing by playing football is silly, because if he truly is the ruler, people will feel obliged to lose, ruining the game. Of course there are the unwise, like the Roman Emperor Commodus, who fancied himself a gladiator. Commodus always won his fights in the arena, and his subjects viewed his predilection for gladiatorial combat as a disgrace. For rulers trying to gain standing, what remains is the role of the status referee, the one who confers honor across domains. Distortions introduced by having to praise his work are thus reduced. This is one of the most important roles of the ruler: the ruler uses his fount of prestige to regulate overall status and prestige competition, so that the right people and the right behaviors win, solving coordination problems and tragedies of the commons.

There are brilliant rulers who really might have something to contribute to a field, and some who aren't particularly brilliant but wish to engage in hobbies for personal fulfillment. A common practice for both of these kinds of rulers is to be active under assumed identities or proxies, sometimes convincingly, sometimes not. Frederick the Great of Prussia, for example, anonymously published a political treatise shortly after assuming the throne. The anonymity prevents the prestige distortions that might come from the ruler visibly competing in one of the domains that he rules over.

The prestige of rulers and, more generally, the prestige landscape created by power, is the fount from which most other prestige flows. If someone tries to grant prestige out of line with this source, it may not be taken seriously, or may find itself undermined by power. If something is not being taken seriously, power can be applied behind the scenes to promote it until it is.

For example, after World War II, American officials in the State Department and the CIA wanted to undermine the dominance of pro-Soviet communists in the Western highbrow cultural scene. To do this, they planned to promote artists and intellectuals who were either anti-Soviet or at least not especially sympathetic to the Soviets — at the time this was often the best you could do in highbrow circles. They considered abstract expressionist painting, which was then a new and obscure movement, a promising candidate. Though no one would call it patriotic, it was American and it wasn't especially communist.

In 1946, the State Department organized an international exhibition of abstract painting called "Advancing American Art". It was so poorly received that the tour was cancelled and the paintings sold off for next to nothing. Undeterred, the CIA, under a front organization called the Congress for Cultural Freedom, continued to arrange international exhibitions for abstract expressionists. Eventually, the movement caught on. It would be an oversimplification to say that the CIA made abstract expressionism famous—there were other influential promoters, like the critic Clement Greenberg—but their support was not irrelevant.

If one looks closely at any society, one will observe that its rulers—and their prestige—subsidize all other sources of prestige. Thus, when the landscape of power shifts, the landscape of prestige shifts accordingly. It is then critical that rulers are incentivized to allocate prestige well—that is, in accordance with the actual distribution of excellence. If they aren't, as in the case of Stalin, the resulting distortions in the allocation of prestige produce distortions in their society's understanding of what is good and what is true. Lysenkoism was an epistemic and moral disaster. This kind of corruption can ultimately have catastrophic effects on the society's health, because the ability to ascertain the truth is fundamental to the functionality of a society's people and its institutions.

Awards are better than prizes

Among the many different kinds of honors, we can pick out two especially common ones: those meant to incentivize a particular achievement with a financial reward, which I call prizes, and those meant to afford prestige on the basis of past achievement, which I call awards. Prizes aim to get some specific thing done, whereas awards aim to affect the distribution of prestige, incentivizing achievement in a more indirect way. With a prize, money is fundamental. With an award, it is incidental. The Millennium Prizes are a prime example of the former, the Academy Awards of the latter.

This distinction is often muddled, leading honors to be less effective than they could be. I have to clarify what I mean by each term, because in practice they aren't used in a reliable way. There are awards that are called prizes and prizes that are called awards. Despite its name, the Nobel Prize is a hybrid case that is more of an award. Though it comes with a financial reward, it is primarily about affording prestige, and this is what those who try to win it are after. The money is nice, but the glory is better.

It's for this reason that I think that awards are more effective than prizes in incentivizing the production of knowledge. Glory is a greater motivator than money. Furthermore, the money attached to prizes is often insufficient for justifying the investment of money, time, energy, social capital, and so on required to achieve the relevant goal.

A better use of prize money is to directly fund projects aimed at the desired achievement. The venture capitalists of Silicon Valley and grantmakers like the Mercatus Center's Emergent Ventures program are good examples. Before any project begins, it's possible to determine which individuals or teams have the best chance of success. Giving them the money beforehand solves the financing problem, and even if success won't make them a fortune, the glory of the achievement -- perhaps augmented by an award -- should be incentive enough.

A prize also provides less return on its creator's investment of social capital than an award. Once the goal is achieved and the prize won, there is no longer a reason for it to exist. It is self-abolishing. An award, on the other hand, can continue to be given out year after year, compounding the investment of prestige. Recognizing this fact, prize-giving organizations often convert their prizes into awards, contributing to confusion about the distinction.

The X Prize illustrates some of these flaws. Created by entrepreneur and space enthusiast Peter Diamandis in the 1990s, the prizes are meant to incentivize breakthroughs in solving the world's biggest problems. Their website says, "Rather than throw money at a problem, we incentivize the solution and challenge the world to solve it." Perhaps the most well-known past prize is the Ansari X Prize, which promised a $10 million reward for the creation of a reusable spacecraft. Many of the other X Prizes are also about breakthroughs in space technology. Since their founding, the X Prize has directly collaborated with firms as well-known as Google, IBM's Watson, and Northrop Grumman, and today counts Google co-founder Larry Page on its board of trustees.

And yet, the great advancements towards space exploration in the past twenty years have had little to do with the X Prize. $10 million is a paltry sum compared to the money required to finance serious efforts in the area, and even less compared to the rewards of success, as SpaceX and Blue Origin have demonstrated. It's safe to say that an X Prize and $10 million played no part in Musk and Bezos' motivations. Even the project that won the Ansari Prize had $100 million in financing. Either the prize money wasn't much of an incentive, or the winning team was very confused.

If it's not really incentivizing breakthroughs, then what is the real use of the X Prize money? It's to garner publicity. The idea of monetary prizes excites our imagination and so lends them virality, and for this narrow purpose the X Prize money has worked. Its creators may understand this, and hope that the publicity brings attention to the relevant problems and so itself incentivizes breakthroughs. The evidence doesn't bear this out, however. The X Prize has garnered its fair share of media coverage, but it has failed to lend massive prestige to the sector of technological innovation, and thus has

not institutionalized newly-legible professional communities of practice in the manner that the Nobel prize did. After all, we forget that much of what we think of as the immutably prestigious “scientific community,” and even the field of professional economics, is a result downstream of such shifts in the landscape of prestige. Imagine how different society would be today if we had a Nobel Prize for technology!

While publicity is good, it’s even better to be able to affect the distribution of prestige throughout society. The more closely social status corresponds to activity that’s ultimately beneficial for society, the more such activity is incentivized, much more strongly than by even a large financial reward. Wisely distributing status makes the difference between a world where most kids dream of becoming YouTubers and one where they dream of taking us to space.

Read more from Samo Burja here.

In physical eschatology, is Aestivation a sound strategy?

In this paper, Anders Sandberg, Stuart Armstrong and Milan M. Cirkovic argue that

> If a civilization wants to maximize computation it appears rational to aestivate until the far future in order to exploit the low temperature environment: this can produce a 1030 multiplier of achievable computation.

Later Charles H. Bennett, Robin Hanson, C. Jess Riedel disagree, claiming

> In fact, while this assumption may apply in the distant future, our universe today contains vast reservoirs and other physical systems in non-maximal entropy states, and computer-generated entropy can be transferred to them at the adiabatic conversion rate of one bit of negentropy to erase one bit of error. This can be done at any time, and is not improved by waiting for a low cosmic background temperature. Thus aliens need not wait to be active. As Sandberg et al. do not provide a concrete model of the effect they assert, we construct one and show where their informal argument goes wrong.

Who was right?

Whence decision exhaustion?

Many people experience something we might call decision or executive exhaustion: after making a lot of decisions, it can be hard to make more decisions and to exert "willpower". Yet, this seems odd because we are constantly making decisions all the time in some sense, choosing to do what we do over everything else we could have otherwise done. So, what and why do we sometimes get exhausted of making decisions when most of the time we do not?

Some notes to consider in answering:

- Some people seem to experience this from all decisions and are worn out after dozens of minutes of being awake.
- Some people seem to never experience this.
- Exhausting decisions seem more salient or like they require more deliberate thought than ones that are not exhausting. Non-exhausting decisions feel automatic.
- Food and rest (but not a full nights sleep) helps some people recover decision function but not everyone seems to respond to this over short enough timescales for it to be useful for recovering functionality within the day.

What's the best explanation of intellectual generativity?

Lately I've found myself wanting to make the argument that intellectual generativity is very important, and that you should be very careful with subtle forces that can corrode it.

"Generativity" is the sort of word that seems to come up a lot in casual conversations in my current circle but I just went looking for a good explanatory post and couldn't find one. I'm fairly confident that someone somewhere has talked about it (not necessarily on LW).

Curious if anyone knows of good existing writing?

And if anyone wanted to write up a fresh explanation that'd be cool as well. (A possible outcome is treating the answer section here as an opportunity to write a first draft that maybe turns into a post if there's consensus the answer is good)

Epistemic Spot Check: The Role of Deliberate Practice in the Acquisition of Expert Performance

Epistemic spot checks typically consist of references from a book, selected by my interest level, checked against either the book's source or my own research. This one is a little different that I'm focusing on a single paragraph in a single paper. Specifically as part of a larger review I read Ericsson, Krampe, and Tesch-Römer's 1993 paper, *The Role of Deliberate Practice in the Acquisition of Expert Performance* (PDF), in an attempt to gain information about how long human beings can productivity do thought work over a time period.

This paper is important because if you ask people how much thought work can be done in a day, if they have an answer and a citation at all, it will be "4 hours a day" and "Cal Newport's *Deep Work*". The Ericsson paper is in turn Newport's source. So to the extent people's beliefs are based on anything, they're based on this paper.

In fact I'm not even reviewing the whole paper, just this one relevant paragraph:

> When individuals, especially children, start practicing in a given domain, the amount of practice is an hour or less per day (Bloom, 1985b). Similarly, laboratory studies of extended practice limit practice to about 1 hr for 3-5 days a week (e.g., Chase & Ericsson, 1982; Schneider & Shiffrin, 1977; Seibel, 1963). A number of training studies in real life have compared the efficiency of practice durations ranging from 1 -8 hr per day. These studies show essentially no benefit from durations exceeding 4 hr per day and reduced benefits from practice exceeding 2 hr (Welford, 1968; Woodworth & Schlosberg, 1954). Many studies of the acquisition of typing skill (Baddeley & Longman, 1978; Dvorak et al.. 1936) and other perceptual motor skills (Henshaw & Holman, 1930) indicate that the effective duration of deliberate practice may be closer to 1 hr per day. Pirolli and J. R. Anderson (1985) found no increased learning from doubling the number of training trials per session in their extended training study. The findings of these studies can be generalized to situations in which training is extended over long periods of time such as weeks, months, and years

Let's go through each sentence in order. I've used each quote as a section header, with the citations underneath it in bold.

"When individuals, especially children, start practicing in a given domain, the amount of practice is an hour or less per day"

Generalizations about talent development, Bloom (1985)

"Typically the initial lessons were given in swimming and piano for about an hour each week, while the mathematics was taught about four hours each week…In addition some learning tasks (or homework) were assigned to be practiced and perfected before the next lesson." (p513)

“…[D]uring the week the [piano] teacher expected the child to practice about an hour a day.” with descriptions of practice but no quantification given for swimming and math (p515).

The quote seems to me to be a simplification. “Expected an hour a day” is not the same as “did practice an hour or less per day.”

“…laboratory studies of extended practice limit practice to about 1 hr for 3-5 days a week”

Skill and working memory, Chase & Ericsson (1982)

This study focused strictly on memorizing digits, which I don’t consider to be that close to thought work.

Controlled and automatic human information processing: I. Detection, search, and attention. Schneider, W., & Shiffrin, R. M. (1977)

This study had 8 people in it and was essentially an identification and reaction time trial.

Discrimination reaction time for a 1,023-alternative task, Seibel, R. (1963)

3 subjects. This was a reaction time test, not thought work. No mention of duration studying.

“These studies show essentially no benefit from durations exceeding 4 hr per day and reduced benefits from practice exceeding 2 hr”

Fundamentals of Skill, Welford (1968)

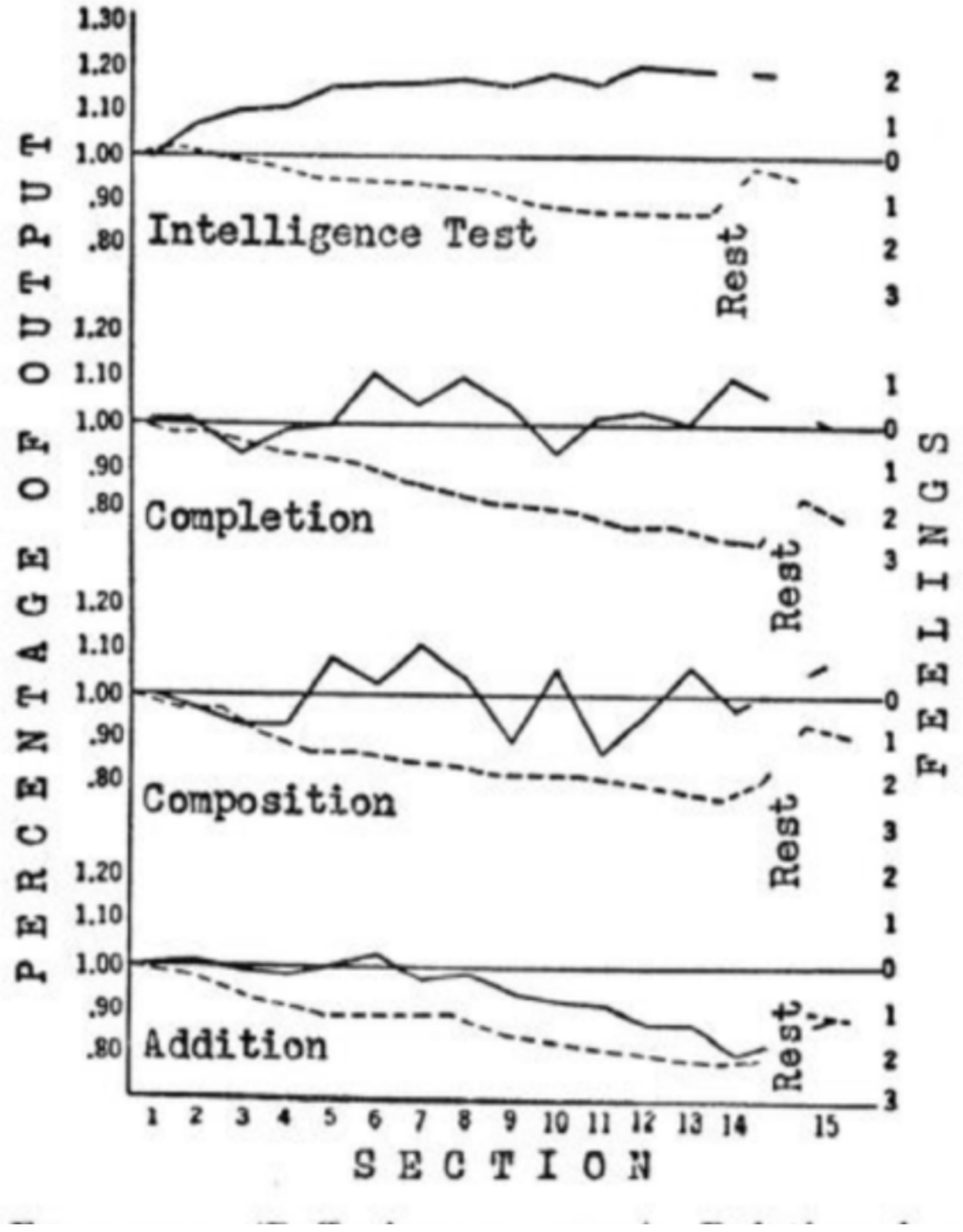
In a book with no page number given, I skipped this one.

Experimental Psychology, Woodworth & Schlosberg (1954)

This too is a book with no page number, but it was available online (thanks, archive.org) and I made an educated guess that the relevant chapter was “Economy in Learning and Performance”. Most of this chapter focused on recitation, which I don’t consider sufficiently relevant.

p800: “Almost any book on applied psychology will tell you that the hourly work output is higher in an eight-hour day than a ten-hour day.”(no source)

Offers this graph as demonstration that only monotonous work has diminishing returns.



p812: An interesting army study showing that students given telegraphy training for 4 hours/day (and spending 4 on other topics) learned as much as students studying 7 hours/day. This one seems genuinely relevant, although not enough to tell us where

peak performance lies, just that four hours are better than seven. Additionally, the students weren't loafing around for the excess three hours: they were learning other things. So this is about how long you can study a particular subject, not total learning capacity in a day.

Many studies of the acquisition of typing skill (Baddeley & Longman, 1978; Dvorak et al.. 1936) and other perceptual motor skills (Henshaw & Holman, 1930) indicate that the effective duration of deliberate practice may be closer to 1 hr per day

The Influence of Length and Frequency of Training Session on the Rate of Learning to Type, Baddeley & Longman (1978)

"Four groups of postmen were trained to type alpha-numeric code material using a conventional typewriter keyboard. Training was based on sessions lasting for one or two hours occurring once or twice per day. Learning was most efficient in the group given one session of one hour per day, and least efficient in the group trained for two 2-hour sessions. Retention was tested after one, three or nine months, and indicated a loss in speed of about 30%. Again the group trained for two daily sessions of two hours performed most poorly.It is suggested that where operationally feasible, keyboard training should be distributed over time rather than massed"

Typewriting behavior; psychology applied to teaching and learning typewriting, Dvorak et al (1936)

Inaccessible book.

The Role of Practice in Fact Retrieval, Pirolli & Anderson (1985)

"We found that fact retrieval speeds up as a power function of days of practice but that the number of daily repetitions beyond four produced little or no impact on reaction time"

Conclusion

Many of the studies were criminally small, and typically focused on singular, monotonous tasks like responding to patterns of light or memorizing digits. The precision of these studies is greatly exaggerated. There's no reason to believe Ericsson, Krampe, and Tesch-Römer's conclusion that the correct number of hours for deliberate practice is 3.5, much less the commonly repeated factoid that humans can do good work for 4 hours/day.

[This post supported by Patreon].

AGI will drastically increase economies of scale

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

Note to mods: I'm a bit uncertain whether posts like this one currently belong on the Alignment Forum. Please move it if it doesn't. Or if anyone would prefer not to have such posts on AF, please let me know.

In Strategic implications of AIs' ability to coordinate at low cost, I talked about the possibility that different AGIs can coordinate with each other much more easily than humans can, by doing something like merging their utility functions together. It now occurs to me that another way for AGIs to greatly reduce coordination costs in an economy is by having each AGI or copies of each AGI profitably take over much larger chunks of the economy (than companies currently own), and this can be done with AGIs that don't even have explicit utility functions, such as copies of an AGI that are all corrigible/intent-aligned to a single person.

Today, there are many industries with large economies of scale, due to things like fixed costs, network effects, and reduced deadweight loss when monopolies in different industries merge (because they can internally charge each other prices that equal marginal costs), but because coordination costs among humans increase super-linearly with the number of people involved (see Moral Mazes and Short Termism for a related recent discussion), that creates diseconomies of scale which counterbalance the economies of scale, so companies tend to grow to a certain size and then stop. But an AGI-operated company, where for example all the workers are AGIs that are intent-aligned to the CEO, would eliminate almost all of the internal coordination costs (i.e., all of the coordination costs that are caused by value differences, such as all the things described in Moral Mazes, "market for lemons" or lost opportunities for trade due to asymmetric information, principal-agent problems, monitoring/auditing costs, costly signaling, and suboptimal Nash equilibria in general), allowing such companies to grow much bigger. In fact, from purely the perspective of maximizing the efficiency/output of an economy, I don't see why it wouldn't be best to have (copies of) one AGI control everything.

If I'm right about this, it seems quite plausible that some countries will foresee it too, and as soon as it can feasibly be done, nationalize all of their productive resources and place them under the control of one AGI (perhaps intent-aligned to a supreme leader or to a small, highly coordinated group of humans), which would allow them to out-compete any other countries that are not willing to do this (and don't have some other competitive advantage to compensate for this disadvantage). This seems to be an important consideration that is missing from many people's pictures of what will happen after (e.g., intent-aligned) AGI is developed in a slow-takeoff scenario.

Risks from Learned Optimization: Conclusion and Related Work

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

This is the fifth of five posts in the Risks from Learned Optimization Sequence based on the paper "Risks from Learned Optimization in Advanced Machine Learning Systems" by Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Each post in the sequence corresponds to a different section of the paper.

Related work

Meta-learning. As described in the first post, meta-learning can often be thought of as meta-optimization when the meta-optimizer's objective is explicitly designed to accomplish some base objective. However, it is also possible to do meta-learning by attempting to make use of mesa-optimization instead. For example, in Wang et al.'s "Learning to Reinforcement Learn," the authors claim to have produced a neural network that implements its own optimization procedure.(28) Specifically, the authors argue that the ability of their network to solve extremely varied environments without explicit retraining for each one means that their network must be implementing its own internal learning procedure. Another example is Duan et al.'s "RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning," in which the authors train a reinforcement learning algorithm which they claim is itself doing reinforcement learning.(5) This sort of meta-learning research seems the closest to producing mesa-optimizers of any existing machine learning research.

Robustness. A system is robust to distributional shift if it continues to perform well on the objective function for which it was optimized even when off the training environment.(29) In the context of mesa-optimization, pseudo-alignment is a particular way in which a learned system can fail to be robust to distributional shift: in a new environment, a pseudo-aligned mesa-optimizer might still competently optimize for the mesa-objective but fail to be robust due to the difference between the base and mesa- objectives.

The particular type of robustness problem that mesa-optimization falls into is the reward-result gap, the gap between the reward for which the system was trained (the base objective) and the reward that can be reconstructed from it using inverse reinforcement learning (the behavioral objective).(8) In the context of mesa-optimization, pseudo-alignment leads to a reward-result gap because the system's behavior outside the training environment is determined by its mesa-objective, which in the case of pseudo-alignment is not aligned with the base objective.

It should be noted, however, that while inner alignment is a robustness problem, the occurrence of unintended mesa-optimization is not. If the base optimizer's objective is not a perfect measure of the human's goals, then preventing mesa-optimizers from arising at all might be the preferred outcome. In such a case, it might be desirable to

create a system that is strongly optimized for the base objective within some limited domain without that system engaging in open-ended optimization in new environments.(11) One possible way to accomplish this might be to use strong optimization at the level of the base optimizer during training to prevent strong optimization at the level of the mesa-optimizer.(11)

Unidentifiability and goal ambiguity. As we noted in the third post, the problem of unidentifiability of objective functions in mesa-optimization is similar to the problem of unidentifiability in reward learning, the key issue being that it can be difficult to determine the "correct" objective function given only a sample of that objective's output on some training data.(20) We hypothesize that if the problem of unidentifiability can be resolved in the context of mesa-optimization, it will likely (at least to some extent) be through solutions that are similar to those of the unidentifiability problem in reward learning. An example of research that may be applicable to mesa-optimization in this way is Amin and Singh's(20) proposal for alleviating empirical unidentifiability in inverse reinforcement learning by adaptively sampling from a range of environments.

Furthermore, it has been noted in the inverse reinforcement learning literature that the reward function of an agent generally cannot be uniquely deduced from its behavior.(30) In this context, the inner alignment problem can be seen as an extension of the value learning problem. In the value learning problem, the problem is to have enough information about an agent's behavior to infer its utility function, whereas in the inner alignment problem, the problem is to test the learned algorithm's behavior enough to ensure that it has a certain objective function.

Interpretability. The field of interpretability attempts to develop methods for making deep learning models more interpretable by humans. In the context of mesa-optimization, it would be beneficial to have a method for determining whether a system is performing some kind of optimization, what it is optimizing for, and/or what information it takes into account in that optimization. This would help us understand when a system might exhibit unintended behavior, as well as help us construct learning algorithms that create selection pressure against the development of potentially dangerous learned algorithms.

Verification. The field of verification in machine learning attempts to develop algorithms that formally verify whether systems satisfy certain properties. In the context of mesa-optimization, it would be desirable to be able to check whether a learned algorithm is implementing potentially dangerous optimization.

Current verification algorithms are primarily used to verify properties defined on input-output relations, such as checking invariants of the output with respect to user-definable transformations of the inputs. A primary motivation for much of this research is the failure of robustness against adversarial examples in image recognition tasks. There are both white-box algorithms,(31) e.g. an SMT solver that in principle allows for verification of arbitrary propositions about activations in the network,(32) and black-box algorithms(33). Applying such research to mesa-optimization, however, is hampered by the fact that we currently don't have a formal specification of optimization.

Corrigibility. An AI system is *corrigible* if it tolerates or assists with its human programmers in correcting itself.(25) The current analysis of corrigibility has focused on how to define a utility function such that, if optimized by a rational agent, that agent would be corrigible. Our analysis suggests that even if such a corrigible

objective function could be specified or learned, it is nontrivial to ensure that a system trained on that objective function would actually be corrigible. Even if the base objective function would be corrigible if optimized directly, the system may exhibit mesa-optimization, in which case the system's mesa-objective might not inherit the corrigibility of the base objective. This is somewhat analogous to the problem of utility-indifferent agents creating other agents that are not utility-indifferent.(25) In the fourth post, we suggest a notion related to corrigibility—corrigible alignment—which is applicable to mesa-optimizers. If work on corrigibility were able to find a way to reliably produce corrigibly aligned mesa-optimizers, it could significantly contribute to solving the inner alignment problem.

Comprehensive AI Services (CAIS).(11) CAIS is a descriptive model of the process by which superintelligent systems will be developed, together with prescriptive implications for the best mode of doing so. The CAIS model, consistent with our analysis, makes a clear distinction between learning (the base optimizer) and functionality (the learned algorithm). The CAIS model predicts, among other things, that more and more powerful general-purpose learners will be developed, which through a layered process will develop services with superintelligent capabilities. Services will develop services that will develop services, and so on. At the end of this "tree," services for a specific final task are developed. Humans are involved throughout the various layers of this process so that they can have many points of leverage for developing the final service.

The higher-level services in this tree can be seen as meta-optimizers of the lower-level services. However, there is still the possibility of mesa-optimization—in particular, we identify two ways in which mesa-optimization could occur in the CAIS-model. First, a final service could develop a mesa-optimizer. This scenario would correspond closely to the examples we have discussed in this sequence: the base optimizer would be the next-to-final service in the chain, and the learned algorithm (the mesa-optimizer in this case), would be the final service (alternatively, we could also think of the entire chain from the first service to the next-to-final service as the base optimizer). Second, however, an intermediary service in the chain might also be a mesa-optimizer. In this case, this service would be an optimizer in *two respects:* it would be the meta-optimizer of the service below it (as it is by default in the CAIS model), but it would also be a mesa-optimizer with respect to the service above it.

Conclusion

In this sequence, we have argued for the existence of two basic AI safety problems: the problem that mesa-optimizers may arise even when not desired (unintended mesa-optimization), and the problem that mesa-optimizers may not be aligned with the original system's objective (the inner alignment problem). However, our work is still only speculative. We are thus left with several possibilities:

1. If mesa-optimizers are very unlikely to occur in advanced ML systems (and we do not develop them on purpose), then mesa-optimization and inner alignment are not concerns.
2. If mesa-optimizers are not only likely to occur but also difficult to prevent, then solving both inner alignment and outer alignment becomes critical for achieving confidence in highly capable AI systems.

3. If mesa-optimizers are likely to occur in future AI systems by default, and there turns out to be some way of preventing mesa-optimizers from arising, then instead of solving the inner alignment problem, it may be better to design systems to not produce a mesa-optimizer at all. Furthermore, in such a scenario, some parts of the outer alignment problem may not need to be solved either: if an AI system can be prevented from implementing any sort of optimization algorithm, then there may be more situations where it is safe for the system to be trained on an objective that is not perfectly aligned with the programmer's intentions. That is, if a learned algorithm is not an optimizer, it might not optimize the objective to such an extreme that it would cease to produce positive outcomes.

Our uncertainty on this matter is a potentially significant hurdle to determining the best approaches to AI safety. If we do not know the relative difficulties of the inner alignment problem and the unintended optimization problem, then it is unclear how to adequately assess approaches that rely on solving one or both of these problems (such as Iterated Distillation and Amplification(34) or AI safety via debate(35)). We therefore suggest that it is both an important and timely task for future AI safety work to pin down the conditions under which the inner alignment problem and the unintended optimization problem are likely to occur as well as the techniques needed to solve them.

Glossary | Bibliography

Can we use ideas from ecosystem management to cultivate a healthy rationality memespace?

Background: ecosystems management practices for improving community memespaces

One can model individual human minds, as well as a community of minds, as an "ecosystem" of "memes". These memes might be things like:

- Bayesian epistemology
- a habit of checking Facebook when one wakes up
- wiggling one's fingers to indicate agreement with a statement
- prefacing things one says with "My model of this is that..."
- Doing calibration training
- Referring to blog posts in conversation
- Taking silent pauses to think mid-conversation

Etc. etc.

Calling this set an "ecosystem", seems to me to be mechanistically very close to what's actually going on. At least, this is because:

1. Memes mutate as they are transmitted between minds
2. Memes undergo selection pressure as they are transmitted
3. The underlying topology/geography of social, cultural and geographical networks of people influence their spread
4. Memes can be in equilibrium with other memes
5. Memes can act as "invasive species"

Now there is an emerging European rationality community, largely driven by efforts from the Prague rationalists. This community imports many memes from the Bay area rationality community. For a high-level, historical examination of this memespace, see Julia Galef's map of bay area memespace.

At a recent CFAR workshop, we discussed how we can ensure that this interaction is successful.

Five of us (Ales Flidr, Elizabeth Garrett, Nora Ammann, Adam Scholl, Jacob Lagerros), felt that the ecosystem model carried sufficient mechanistic similarity to the actual situation that it would be helpful to read up on things like: protocols for deliberate introductions of new species into new environments, invasive species regulations and protection programs, pest control, and more.

Collection of background notes

We spent 1h researching this. Now the outside view predicts that if we were to leave it at that, the 16-page Google doc would never be used again. Hence we're experimenting with releasing our notes together with a LessWrong question, in order to allow others to benefit from and build on our progress.

You can find our notes here.

These notes are provided “as is”. I (jacobjacob) briefly went through them to make them more readable, but apart from that this should not be interpreted as something the authors endorse as being true, and despite originating at a CFAR workshop it is not official CFAR content.

Open questions

We’d be interested in using further research to answer questions such as:

- What are warning signs of a memespace/ecosystem being harmed?
- What are best practices for introducing a new meme into a memespace, and what can we learn from actual ecosystems?
- What are some useful models for thinking about this problem?

How can we measure creativity?

Status: have spent about two hours on this.

As part of measuring how marginal productivity changes over time, I need to know how to assess creativity. One promising test for that is the Torrance Tests of Creative Thinking, in which subjects are given an open ended prompt, and graded on their answers. Answers are evaluated by a human being for fluency, originality, abstractness of titles, elaboration and resistance to premature closure (sample questions for the curious). But does the TTCT predict anything we actually care about?

The creator of the tests studied their predictive ability in a longitudinal study that lasted 50 years so far. The 40 year follow up showed good-for-social-sciences correlation between childhood TTCT scores and adult creative achievement, although IQ had a stronger correlation. The 50 year follow up (conducted by different experimenters) found no correlation between score and "public achievement" unless combined with IQ. Given that these studies were subject to the usual social science weaknesses, multiplied by 50 years and subjective grading, I do not count this as strong evidence.

A smaller study in Brazil found adulthood scores of recognized creative achievers and non-achievers to vary.

Basic googling found more academics saying both disparaging (beginning of page 310) and encouraging things.

My default assumption is that psychometric tests are invalid, and this evidence isn't enough to make me change my mind. But I don't have anything better to use for my actual goal, which is a measurable task that taxes creativity and *nothing else*, and this has a certain face validity to it. Does anyone have information to sway on the validity of the test, or an alternative test to use?

ISO: Automated P-Hacking Detection

I'm sure there's some ML students/researchers on Lesswrong in search of new projects, so here's one I'd love to see and probably won't build myself: an automated method for predicting which papers are unlikely to replicate, given the text of the paper. Ideally, I'd like to be able to use it to filter and/or rank results from Google scholar.

Getting a good data set would probably be the main bottleneck for such a project. Various replication-crisis papers which review replication success/failure for tens or hundreds of other studies seem like a natural starting point. Presumably some amount of feature engineering would be needed; I doubt anyone has a large enough dataset of labelled papers to just throw raw or lightly-processed text into a black box.

Also, if anyone knows of previous attempts to do this, I'd be interested to hear about it.

Conditions for Mesa-Optimization

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

This is the second of five posts in the Risks from Learned Optimization Sequence based on the paper "Risks from Learned Optimization in Advanced Machine Learning Systems" by Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Each post in the sequence corresponds to a different section of the paper.

In this post, we consider how the following two components of a particular machine learning system might influence whether it will produce a mesa-optimizer:

1. **The task:** The training distribution and base objective function.
2. **The base optimizer:** The machine learning algorithm and model architecture.

We deliberately choose to present theoretical considerations for why mesa-optimization may or may not occur rather than provide concrete examples. Mesa-optimization is a phenomenon that we believe will occur mainly in machine learning systems that are more advanced than those that exist today.[1] Thus, an attempt to induce mesa-optimization in a *current* machine learning system would likely require us to use an artificial setup specifically designed to induce mesa-optimization. Moreover, the limited interpretability of neural networks, combined with the fact that there is no general and precise definition of "optimizer," means that it would be hard to evaluate whether a given model is a mesa-optimizer.

2.1. The task

Some tasks benefit from mesa-optimizers more than others. For example, tic-tac-toe can be perfectly solved by simple rules. Thus, a base optimizer has no need to generate a mesa-optimizer to solve tic-tac-toe, since a simple learned algorithm implementing the rules for perfect play will do. Human survival in the savanna, by contrast, did seem to benefit from mesa-optimization. Below, we discuss the properties of tasks that may influence the likelihood of mesa-optimization.

Better generalization through search. To be able to consistently achieve a certain level of performance in an environment, we hypothesize that there will always have to be some minimum amount of optimization power that must be applied to find a policy that performs that well.

To see this, we can think of optimization power as being measured in terms of the number of times the optimizer is able to divide the search space in half—that is, the number of bits of information provided.(9) After these divisions, there will be some remaining space of policies that the optimizer is unable to distinguish between. Then, to ensure that all policies in the remaining space have some minimum level of performance—to provide a performance lower bound[2] —will always require the

original space to be divided some minimum number of times—that is, there will always have to be some minimum bits of optimization power applied.

However, there are two distinct levels at which this optimization power could be expended: the base optimizer could expend optimization power selecting a highly-tuned learned algorithm, or the learned algorithm could itself expend optimization power selecting highly-tuned actions.

As a mesa-optimizer is just a learned algorithm that itself performs optimization, the degree to which mesa-optimizers will be incentivized in machine learning systems is likely to be dependent on which of these levels it is more advantageous for the system to perform optimization. For many current machine learning models, where we expend vastly more computational resources training the model than running it, it seems generally favorable for most of the optimization work to be done by the base optimizer, with the resulting learned algorithm being simply a network of highly-tuned heuristics rather than a mesa-optimizer.

We are already encountering some problems, however—Go, Chess, and Shogi, for example—for which this approach does not scale. Indeed, our best current algorithms for those tasks involve explicitly making an optimizer (hard-coded Monte-Carlo tree search with learned heuristics) that does optimization work on the level of the learned algorithm rather than having all the optimization work done by the base optimizer.(10) Arguably, this sort of task is only adequately solvable this way—if it were possible to train a straightforward DQN agent to perform well at Chess, it plausibly would *have* to learn to internally perform something like a tree search, producing a mesa-optimizer. [3]

We hypothesize that the attractiveness of search in these domains is due to the diverse, branching nature of these environments. This is because search—that is, optimization—tends to be good at generalizing across diverse environments, as it gets to individually determine the best action for each individual task instance. There is a general distinction along these lines between optimization work done on the level of the learned algorithm and that done on the level of the base optimizer: the learned algorithm only has to determine the best action for a given task instance, whereas the base optimizer has to design heuristics that will hold regardless of what task instance the learned algorithm encounters. Furthermore, a mesa-optimizer can immediately optimize its actions in novel situations, whereas the base optimizer can only change the mesa-optimizer's policy by modifying it ex-post. Thus, for environments that are diverse enough that most task instances are likely to be completely novel, search allows the mesa-optimizer to adjust for that new task instance immediately.

For example, consider reinforcement learning in a diverse environment, such as one that directly involves interacting with the real world. We can think of a diverse environment as requiring a very large amount of computation to figure out good policies before conditioning on the specifics of an individual instance, but only a much smaller amount of computation to figure out a good policy once the specific instance of the environment is known. We can model this observation as follows.

Suppose an environment is composed of N different instances, each of which requires a completely distinct policy to succeed in.[4] Let P be the optimization power (measured in bits(9)) applied by the base optimizer, which should be approximately proportional to the number of training steps. Then, let x be the optimization power

applied by the learned algorithm in each environment instance and f(x) the total amount of optimization power the base optimizer must put in to get a learned algorithm capable of performing that amount of optimization.[5] We will assume that the rest of the base optimizer's optimization power, $P - f(x)$, goes into tuning the learned algorithm's policy. Since the base optimizer has to distribute its tuning across all N task instances, the amount of optimization power it will be able to contribute to each instance will be $\frac{P-f(x)}{N}$, under the previous assumption that each instance requires a completely distinct policy. On the other hand, since the learned algorithm does all of its optimization at runtime, it can direct all of it into the given task instance, making its contribution to the total for each instance simply x.[6]

Thus, if we assume that, for a given P, the base optimizer will select the value of x that maximizes the minimum level of performance, and thus the total optimization power applied to each instance, we get[7]

$$x^* = \text{argmax}_x \frac{P-f(x)}{N} + x.$$

As one moves to more and more diverse environments—that is, as N increases—this model suggests that x will dominate $\frac{P-f(x)}{N}$, implying that mesa-optimization will become more and more favorable. Of course, this is simply a toy model, as it makes many questionable simplifying assumptions. Nevertheless, it sketches an argument for a pull towards mesa-optimization in sufficiently diverse environments.

As an illustrative example, consider biological evolution. The environment of the real world is highly diverse, resulting in non-optimizer policies directly fine-tuned by evolution—those of plants, for example—having to be very simple, as evolution has to spread its optimization power across a very wide range of possible environment instances. On the other hand, animals with nervous systems can display significantly more complex policies by virtue of being able to perform their own optimization, which can be based on immediate information from their environment. This allows sufficiently advanced mesa-optimizers, such as humans, to massively outperform other species, especially in the face of novel environments, as the optimization performed internally by humans allows them to find good policies even in entirely novel environments.

Compression of complex policies. In some tasks, good performance requires a very complex policy. At the same time, base optimizers are generally biased in favor of selecting learned algorithms with lower complexity. Thus, all else being equal, the base optimizer will generally be incentivized to look for a highly compressed policy.

One way to find a compressed policy is to search for one that is able to use general features of the task structure to produce good behavior, rather than simply memorizing the correct output for each input. A mesa-optimizer is an example of such a policy. From the perspective of the base optimizer, a mesa-optimizer is a highly-compressed version of whatever policy it ends up implementing: instead of explicitly encoding the details of that policy in the learned algorithm, the base optimizer simply needs to encode how to search for such a policy. Furthermore, if a mesa-optimizer can

determine the important features of its environment at runtime, it does not need to be given as much prior information as to what those important features are, and can thus be much simpler.

This effect is most pronounced for tasks with a broad diversity of details but common high-level features. For example, Go, Chess, and Shogi have a very large domain of possible board states, but admit a single high-level strategy for play—heuristic-guided tree search—that performs well across all board states.(10) On the other hand, a classifier trained on random noise is unlikely to benefit from compression at all.

The environment need not necessarily be too diverse for this sort of effect to appear, however, as long as the pressure for low description length is strong enough. As a simple illustrative example, consider the following task: given a maze, the learned algorithm must output a path through the maze from start to finish. If the maze is sufficiently long and complicated then the specific strategy for solving this particular maze—specifying each individual turn—will have a high description length. However, the description length of a general optimization algorithm for finding a path through an arbitrary maze is fairly small. Therefore, if the base optimizer is selecting for programs with low description length, then it might find a mesa-optimizer that can solve all mazes, despite the training environment only containing one maze.

Task restriction. The observation that diverse environments seem to increase the probability of mesa-optimization suggests that one way of reducing the probability of mesa-optimizers might be to keep the tasks on which AI systems are trained highly restricted. Focusing on building many individual AI services which can together offer all the capabilities of a generally-intelligent system rather than a single general-purpose artificial general intelligence (AGI), for example, might be a way to accomplish this while still remaining competitive with other approaches.(11)

Human modeling. Another aspect of the task that might influence the likelihood of mesa-optimization is the presence of humans in the environment.(12) Since humans often act as optimizers, reasoning about humans will likely involve reasoning about optimization. A system capable of reasoning about optimization is likely also capable of reusing that same machinery to do optimization itself, resulting in a mesa-optimizer. For example, it might be the case that predicting human behavior requires instantiating a process similar to human judgment, complete with internal motives for making one decision over another.

Thus, tasks that do not benefit from human modeling seem less likely to produce mesa-optimizers than those that do. Furthermore, there are many tasks that might benefit from human modeling that don't explicitly involve modeling humans—to the extent that the training distribution is generated by humans, for example, modeling humans might enable the generation of a very good prior for that distribution.

2.2. The base optimizer

It is likely that certain features of the base optimizer will influence how likely it is to generate a mesa-optimizer. First, though we largely focus on reinforcement learning in this sequence, RL is not necessarily the only type of machine learning where mesa-optimizers could appear. For example, it seems plausible that mesa-optimizers could appear in generative adversarial networks.

Second, we hypothesize that the details of a machine learning model's architecture will have a significant effect on its tendency to implement mesa-optimization. For example, a tabular model, which independently learns the correct output for every input, will never be a mesa-optimizer. On the other hand, if a hypothetical base optimizer looks for the program with the shortest source code that solves a task, then it is more plausible that this program will itself be an optimizer.(13) However, for realistic machine learning base optimizers, it is less clear to what extent mesa-optimizers will be selected for. Thus, we discuss some factors below that might influence the likelihood of mesa-optimization one way or the other.

Reachability. There are many kinds of optimization algorithms that a base optimizer could implement. However, almost every training strategy currently used in machine learning uses some form of local search (such as gradient descent or even genetic algorithms). Thus, it seems plausible that the training strategy of more advanced ML systems will also fall into this category. We will call this general class of optimizers that are based on local hill-climbing *local optimization processes.*

We can then formulate a notion of *reachability,* the difficulty for the base optimizer to find any given learned algorithm, which we can analyze in the case of a local optimization process. A local optimization process might fail to find a particular learned algorithm that would perform very well on the base objective if the learned algorithm is surrounded by other algorithms that perform poorly on the base objective. For a mesa-optimizer to be produced by a local optimization process, it needs to not only perform well on the base objective, but also be *reachable;* that is, there needs to be a path through the space of learned algorithms to it that is approximately monotonically increasing. Furthermore, the degree to which the path only need be approximate—that is, the degree to which ML training procedures can escape local optima—is likely to be critical, as optimization algorithms are complex enough that it might require a significant portion of the algorithm to be present before performance gains start being realized.

Algorithmic range. One key factor likely to determine the reachability of mesa-optimizers is the *algorithmic range* of the learned algorithms—that is, how extensive is the set of algorithms (or how expressive is the model space) capable of being found by the base optimizer. The more extensive a model's algorithmic range, the broader the space of possible learned algorithms, and thus the more likely that it will be able to find one that is a mesa-optimizer, assuming the base optimizer is incentivized to do so. For example, architectures that explicitly give the algorithm access to a wide range of possible computations, such as recurrent neural networks or neural Turing machines,(14) seem more likely to produce mesa-optimizers.

Inductive biases. Another important factor is the degree to which the base optimizer is explicitly or implicitly biased in various ways. The nature of these inductive biases will contribute to the likelihood of a mesa-optimizer being selected for. One of the most important kinds of inductive bias is simplicity bias, which would almost certainly be exhibited by almost all base optimizers. We identify three ways in which simplicity bias can manifest itself:

1. An explicit penalty due to parameter regularization or architectural constraints such as weight-sharing or sparse connections.
2. An implicit bias due to the model architecture. For example, it has been shown that neural networks are more likely to fit a simple function to a set of training data, even when no regularization is used.(15)

3. The capacity limitations of the model. The size of a model imposes a hard limit on the complexity of the functions it is able to represent. Thus, to the degree that the base optimizer is selecting based on performance, it will be driven to "squeeze out" as much performance as it can for any given model capacity, leading to a bias in favor of relatively compressed policies.

The more a base optimizer is biased towards simple solutions, the more it will be incentivized to find a compressed policy such as a mesa-optimizer.

The base optimizer could also be biased in other ways. For example, it could be biased towards algorithms with a low time or space complexity. If we view mesa-optimization as delegation of optimization work, a time penalty on the learned algorithm is likely to incentivize the base optimizer to do more precomputation itself by pre-computing relevant heuristics rather than delegating optimization work to the learned algorithm. Thus, we hypothesize that penalizing description length will favor mesa-optimizers while penalizing time complexity will disfavor them. This suggests the following speculative conjecture: neither a minimal-depth nor minimal-size boolean circuit that solves a problem can be a mesa-optimizer.(16)

Lastly, another form of bias that might have particularly interesting effects is the presence of an information funnel. In deep learning, the base optimizer directly selects the values of individual parameters. In biological evolution, by contrast, the base optimizer selects DNA strings, which in turn produce the detailed structure of the brain only indirectly. This implicitly introduces pressure towards compressing the brain's structure. As we noted previously, this might favor the development of mesa-optimizers, as they can be seen as a way of compressing complex policies. Because deep learning algorithms currently select weights directly, there is no such information funnel. In the case of meta-learning over a machine learning model's architecture, however, an information funnel might present itself due to the fact that the meta-learner has to learn a single policy for creating architectures that is capable of solving all of the problems it is presented with.

Statefulness. Another factor that could make it easier for a learned algorithm to implement a complex optimization process is the ability of the learned algorithm to save and recall information. The ability to refer back to previous state enables the caching of intermediate results, increasing the architecture's algorithmic range by giving the learned algorithm significantly more time to complete its computations. Statefulness, therefore, favors mesa-optimization by decreasing the implicit penalty on time complexity imposed by enforcing all computations to end when the learned algorithm produces its output. One simple form of this might be a recurrent neural network, which can perform computations with longer time horizons due to its ability to pass intermediate activations across different time steps. An even more extreme version of this might be a network that is able to store state not just across different time steps in the same episode, but also across episodes, enabling it to do more long-term learning, though such an architecture is currently completely hypothetical.

Hard-coded optimization. One possible means of alleviating some of these issues might be to include hard-coded optimization where the learned algorithm provides only the objective function and not the optimization algorithm. The stronger the optimization performed explicitly, the less strong the optimization performed implicitly by the learned algorithm needs to be. For example, architectures that explicitly perform optimization that is relevant for the task—such as hard-coded Monte Carlo tree search—might decrease the benefit of mesa-optimizers by reducing the need for optimization other than that which is explicitly programmed into the system.

The third post in the Risks from Learned Optimization Sequence, titled "The Inner Alignment Problem," can be found here.

Glossary | Bibliography

---

1. As of the date of this post. Note that we do examine some existing machine learning systems that we believe are close to producing mesa-optimization in post 5. ↩

2. It is worth noting that the same argument also holds for achieving an average-case guarantee. ↩

3. Assuming reasonable computational constraints.. ↩

4. This definition of N is somewhat vague, as there are multiple different levels at which one can chunk an environment into instances. For example, one environment could always have the same high-level features but completely random low-level features, whereas another could have two different categories of instances that are broadly self-similar but different from each other, in which case it's unclear which has a larger N. However, one can simply imagine holding N constant for all levels but one and just considering how environment diversity changes on that level. ↩

5. Note that this makes the implicit assumption that the amount of optimization power required to find a mesa-optimizer capable of performing x bits of optimization is independent of N. The justification for this is that optimization is a general algorithm that looks the same regardless of what environment it is applied to, so the amount of optimization required to find an x-bit optimizer should be relatively independent of the environment. That being said, it won't be completely independent, but as long as the primary difference between environments is how much optimization they need, rather than how hard it is to do optimization, the model presented here should hold. ↩

6. Note, however, that there will be some maximum x simply because the learned algorithm generally only has access to so much computational power. ↩

7. Subject to the constraint that $P - f(x) \geq 0$. ↩

LessWrong FAQ

This is a new FAQ written LessWrong 2.0. This is the first version and I apologize if it is a little rough. Please comment or message with further questions, typos, things that are unclear, etc.

The old FAQ on the LessWrong Wiki still contains much excellent information, however it has not been kept up to date.

Advice! We suggest you navigate this guide with the help on the table of contents (ToC) in the left sidebar. You will need to scroll to see all of it. Mobile users need to click the menu icon in the top left.

The major sections of this FAQ are:

About LessWrong

What is LessWrong?

LessWrong is a community dedicated to improving our reasoning and decision-making. We seek to hold true beliefs and to be effective at accomplishing our goals. More generally, we want to develop and practice the art of human rationality.

To that end, LessWrong is a place to 1) develop and train rationality, and 2) apply one's rationality to real-world problems.

LessWrong serves these purposes with its library of rationality writings, community discussion forum, open questions research platform, and community page for in-person events.

See also: Welcome to LessWrong!

What is *rationality?*

Rationality is a term which can have different meanings to different people. You might already associate with a few things. On LessWrong, we mean something like the following:

- Rationality is thinking in ways which systematically arrive at truth.
- Rationality is thinking in ways which cause you to systematically achieve your goals.
- Rationality is trying to do better on purpose.

- Rationality is reasoning well even in the face of massive uncertainty.
- Rationality is making good decisions even when it's hard.
- Rationality is being self-aware, understanding how your own mind works, and applying this knowledge to thinking better.

See also: What Do We Mean By "Rationality"?, Why Spock is Not Rational, What are the open problems in Human Rationality?

What is the history of LessWrong?

In 2006, Eliezer Yudkowsky and others began writing on Overcoming Bias, a group blog with the general theme of how to move one's beliefs closer to reality despite biases such as overconfidence and wishful thinking. In 2009, Eliezer moved to a new community blog, *LessWrong*. Eliezer seeded LessWrong with a series of daily blog posts which became known as *The Sequences*. These writings attracted a large community of readers and writers interested in the art of human rationality.

See also: A Brief History of LessWrong

What makes LessWrong different from other discussion forums?

A combination of traits makes LessWrong distinct among online communities.

1. We have unusually high standards of discourse. We emphasize curiosity, truth-seeking, critical self-reflection, intellectual collaboration, and the long attention spans required to actually think through complicated ideas.
2. We are open to unusual ideas and are willing to doubt conventional wisdom. Curiosity and truth-seeking require a willingness to sometimes consider positions which are strange by ordinary standards, and in some cases, these positions will turn out to be credible. As a result of this openness, some unconventional ideas are prevalent on LessWrong and many more are entertained.
3. We make intellectual progress by building on a large number of communally-shared background ideas and concepts.

Why the name? It is a bit odd . . .

I (Ruby) personally wasn't there when the name was chosen so I'm not certain of the historical thought process, but I interpret the name "LessWrong" as expressing two important points:

1. A humble recognition that no human is ever going to attain perfectly true beliefs and be right about everything. We should always believe that some of our beliefs are mistaken, we just don't know which ones.
2. A bold recognition that notwithstanding the impossibility of being perfectly right, there is still the possibility of being *less wrong.* Everyone believes false things, but some believe a lot fewer wrong things than others.

And so the aspiration of LessWrong is that by dedicating ourselves to learning how to think in ways which more systematically lead to truth (what we succinctly call *rationality*), we can meaningfully reduce our mistaken notions and have far more accurate models of reality.

Who is this Eliezer guy I keep hearing about?

Eliezer Yudkowsky was the original founder of LessWrong back in 2009. His writings on rationality attracted to the site a large number of people enthusiastic about learning to think better. Eliezer's best-known works are The Sequences, (later renamed Rationality: From AI to Zombies) and Harry Potter and the Methods of Rationality. These texts are part of LessWrong's philosophical foundation, and so unsurprisingly, you will see mentions of Eliezer not infrequently.

How does LessWrong make money?

We don't. The LessWrong organization is a nonprofit funded by donations.

This hopefully has the benefit of reducing our incentives to optimize for clicks and pageviews. Instead, we can focus on our stated purpose.

Can I see the source code for the site?

Yes you can! We are open source, and you can find the code on Github here.

I have feedback, bug reports, or questions not answered in this FAQ. What should I do?

You have several options.

1. Message the LessWrong team via Intercom (available in the bottom right). Ensure you don't have *Hide Intercom* set in your account settings.
2. Send a private message to a member of the LessWrong team (see these on the team page)
3. Open an issue on the LessWrong Github repository.
4. Ask a question
5. For complaints and concerns regarding the LessWrong team, you can message Vaniver.

Oh no! I think I lost my post/draft/sanity! What can I do?

LessWrong stores revisions of posts as you're drafting them. If you think you have lost content, please message the team via Intercom and we'll see what we can do.

Getting Started

I'm new. Where do I start?

We encourage new users to read for a while before diving into discussions or making their own posts. This is helpful for new users to understand the site's culture and background.

- Our welcome page offers a high-level description of LessWrong and includes a list of sample posts. It is a great way to get a feel for what LessWrong is like.
- For new members who want to get up to speed, we direct you towards our core readings which can be found on the Library page and are described elsewhere in this FAQ.

- At the same time, feel free to browse more recent content. This answer describes all the way you can locate content on LessWrong.
- Unlike other places on the Internet, it is often worthwhile to read the comment sections on posts. Our commenting guidelines state that is preferable:
    - Aim to explain, not persuade.
    - Present your own perspective rather than state group consensus or invoking authorities.
    - Get curious. If you disagree with some, try to figure out what they're thinking. What's their model? Don't just assume they're dumb or evil.

If you're very new and you begin posting or commenting, you might find that you are quickly downvoted. This doesn't mean you're bad or unwelcome! But you are probably violating a norm or ignoring expected knowledge on the site. We suggest you read up a bit more before trying again later.

What's a good and fast way to learn about how the website works?

LessWrong extensively uses tool-tips and content previews to help users understand how the site works and see what content is even before they click.

We encourage you to mouse over most elements of the site to see what pops up. You will find:

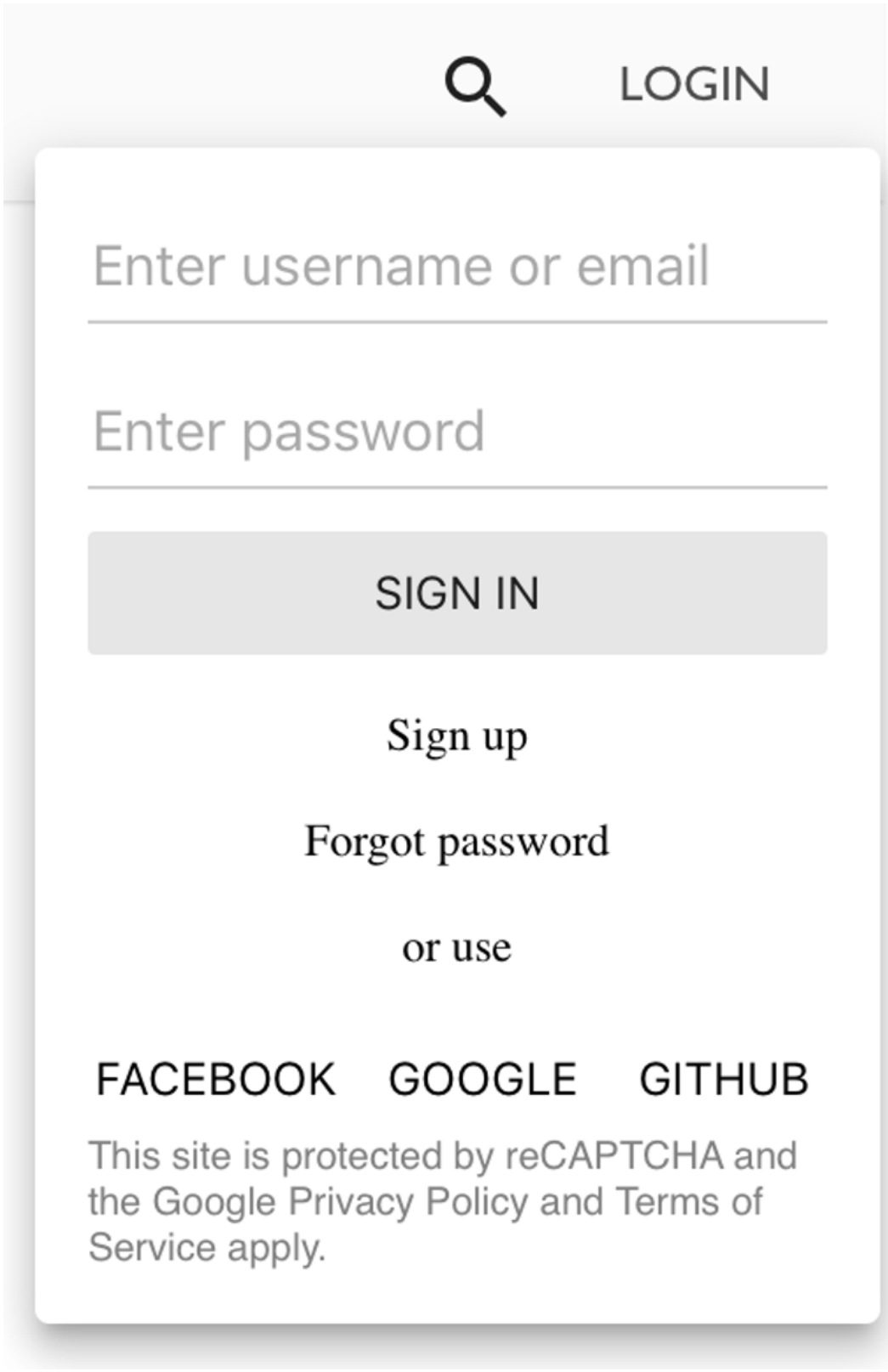
- Items in the left sidebar have tool-tips.
- Hovering over post titles displays an excerpt, reading time, and other meta info.
- Hovering over usernames displays karma, join date, number of posts and comments, and a bio if the user has set one.
- Hovering over karma scores displays the number of votes (in our karma system, karma does not usually equal the number of votes).

How do I create an account? (And why should I?)

Although not required to use the LessWrong website; we recommend creating an account so that you can:

- Subscribe to users and different classes of posts.
- Save your user settings
- Vote and comment on posts.
- Store your reading history, enabling tailored recommendations and potentially new features such as viewing your reading history and creating custom reading lists.

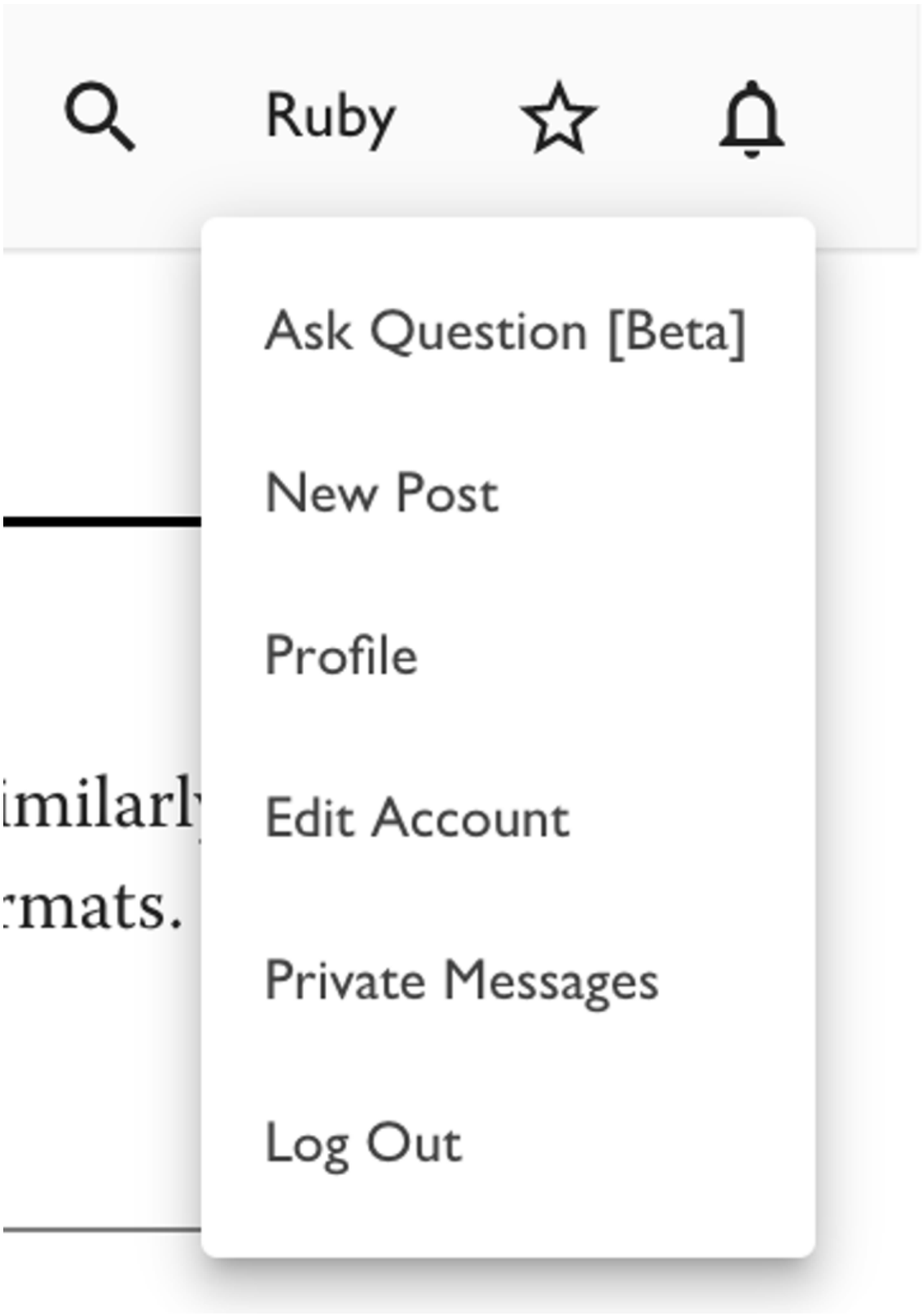
Creating an account takes under 30 seconds. Click *login* in the top right and enter a username, email, and password.


LOGIN
Enter username or email
Enter password
SIGN IN
Sign up
Forgot password
or use
FACEBOOK
GOOGLE
GITHUB
This site is protected by reCAPTCHA and the Google Privacy Policy and Terms of Service apply.

Once you have created an account, feel free to introduce yourself in the latest Open/Welcome thread. Let others know how you found LessWrong, your background, and what you're hoping for from LessWrong. This allows existing members to point you in the direction of material which you might especially like.

How do I *Ask Questions/Make Posts/Go to My Profile/Private Message/Log Out?*

For logged-in users, you can access all these options via the drop-down menu accessible by clicking your username.



The star to the right of your username is the *karma notifier* (star icon) and button for notifications panel (bell icon).

How do I edit my account settings? What can I do?

By clicking on your username and clicking *Edit Account,* you access your account settings. There you can:

- Set a bio for your account to let other LessWrong members know about you. If you set one, it will show up when they mouse-over your username.
- Hide or show Intercom (messaging service with the LessWrong team members).
- Activate the markdown editor.
- Toggle comment collapse settings.
- **Opt into beta features (new)**
- Adjust settings for notifications of responses to your posts and comments
- Adjust settings for the karma notifier.
- Unsubscribe from your email subscriptions.

Reading Content

What are all the ways to access content on LessWrong?

Ah, there are many ways!

Homepage

LessWrong's homepage has the following content sections:

- Core Reading (shown only to logged-out users)
- Curated
- Latest Posts
- Recent Discussion

Core Readings

The core readings section provides links to texts which describe the intellectual foundations of LessWrong. They are described here.

Curated

Each week, LessWrong's moderation team selects on average three posts which seem to us to be especially well-written, insightful, instructive, or otherwise important. These are tagged as *curated posts* and appear with a star icon next to the title.

The three most recently curated posts appear in the *Curated* section. You can view more Curated posts by clicking *View All Curated Posts* or selecting the Curated filter on the AllPosts page.

Beneath the Curated section is a button to subscribe via email or RSS to curated posts (~3/week).

Latest Posts

The *Latest Posts* section displays all* recent posts to LessWrong. These sorted *magically*** to balance between recency and quality (as indicated by karma score), i.e. more upvoted posts

remain higher up in the Latest Posts section for longer.

*By default, only *Frontpage posts* are displayed in the Latest Posts section. To enable Personal blogposts to appear as well, check the checkbox beneath the section. See more in What's the difference between Personal Blogposts and Frontpage Posts?

**LessWrong uses the following formula to rank posts in *Latest Posts*:

$$\frac{karmaScore}{(ageInHours + 2)^{1.15}}$$

This is same the formula as used by Hacker News. You can read about it here.

Recent Discussion

This section is a purely-time based feed of the most recent comment activity happening on posts. Currently, all posts (both Personal blogposts and Frontpage) are shown. Discussion is grouped by post but restricted to only showing a few comments per post.

All Posts Page (aka Archive)

Whereas the homepage displays posts ordered with a magical algorithm, the *All Posts* page gives you complete control over which posts are included and how they ordered.

The All Posts page can be accessed via the left sidebar and drop-down menu (desktop); buttons on the bottom of the screen (mobile); or directly via www.lesswrong.com/allPosts

The gear icon allows you to select which posts:

- All Posts (absolutely everything)
- Frontpage (pages given *Frontpage status* by the moderation team)
- Curated (pages given *Curated status* by the moderation team)
- Questions (from our Open Questions platform)
- Events (from the Community Events Page)
- Meta (deprecated category containing posts about the LessWrong website and similar)

These can then be sorted by: Daily, Magic, has Recent Comments, New, Old, and Top.

The Library

The Library page is accessible from the left sidebar/drop-down menu (desktop) or the buttons at the bottom of the screen (mobile). The Library page contains sequences (ordered sets of posts) and collections (ordered sets of sequences) of LessWrong's best writings. These are split into *Core Readings, Curated Sequences, and Community Sequences.*

LessWrong's developers have put effort into making the reading experience in The Library as convenient and enjoyable as possible.

Core Readings

These are *Rationality: From AI to Zombies* (formerly *The Sequences*), *The Codex*, and *Harry Potter and the Methods of Rationality*. They are described in What are LessWrong's core readings?

Curated Sequences

Similar Curated posts, *Curated sequences* are sets of posts which LessWrong's moderation team think are especially valuable and ought to be included in LessWrong's intellectual

canon.

Top curated sequences include:

- Introduction to Game Theory
- Babble and Prune
- Community and Cooperation

Community Sequences

Any LessWrong site member, not just moderators, can create post sequences. These appear in the *Community Sequences* section.

Standout mentions include:

- Concepts in formal epistemology
- Share models, not beliefs
- Keith Stanovich: What Intelligence Tests Miss

Sequences have qualitative benefits over posts in that an author can build towards a larger point or explain more nuanced concepts than is possible in single (even quite long) blog posts.

You can also create your own sequence on this page.

User Page

Lastly, you can access a User's posts and comments directly from their user page.

Note that you have the same options available for sorting and filtering a user's posts as you do on the All Posts page.

What's the difference between Frontpage posts and Personal blogposts?

Although LessWrong's focus is on the development and application of rationality, we invite posts on almost any topic. To ensure that the default experience is still one centered on rationality, LessWrong classifies posts into *Frontpage posts* and *Personal blogposts.*

Frontpage posts must meet the criteria of being broadly relevant to LessWrong's main interests; timeless, i.e. not about recent events; and are attempts to explain not persuade. In contrast, Personal blogposts can be on any topic of interest to the author including divisive topics (which we generally keep off the frontpage), discussions about the community, and meta posts about LessWrong itself.

Frontpage posts have visibility by default. Personal blogposts can be viewed by: i) checking the "show Personal blogposts" checkbox on the homepage, ii) via the All Posts page if "All Posts" filter option is selected, iii) via a user's profile page, iv) in the Recent Discussion section of the homepage.

See also: *Site Guide: Personal Blogposts vs Frontpage Posts*

What are Curated posts?

Each week, LessWrong's moderation team selects on average three posts which seem to us to be especially well-written, insightful, instructive, or otherwise important. These are tagged

as *curated posts* and appear with a star icon next to the title.

All Curated posts will also be Frontpage posts.

The three most recently curated posts appear in the *Curated* section. You can view more Curated posts by clicking *View All Curated Posts* or selecting the Curated filter on the AllPosts page.

Beneath the Curated section is a button to subscribe via email or RSS to curated posts (~3/week).

What are LessWrong's core readings?

The following texts lay the philosophical foundations of the LessWrong website and community. They are widely regarded as excellent, and, even when the ideas are not universally agreed upon, they are still commonly assumed background knowledge in the community.

Rationality: AI to Zombies (aka "the Sequences")

In 2006, Eliezer began posting on a precursor to LessWrong, the shared blog, *Overcoming Bias* before the current site was launched in 2019. He posted nearly daily for several years and those writings became known as *the Sequences.* Later they were edited into a book, *Rationality: A-Z (or RAZ).*

Rationality: A-Z is a deep exploration of how human minds can come to understand the world they exist in - and all the reasons they so commonly fail to do. The comprehensive work:

- lays foundational conceptions of belief, evidence, and understanding
- reviews the systematic biases and common excuses which cause us to believe false things
- offers guidance on how to change our minds and how to use language effectively to describe the world
- depicts the nature of human psychology with reference to how evolution produced us
- clarifies the kind of morality humans like us can have in a reducible, physical world
- and repeatedly reminds us that confusion and mystery exist only in our minds.

Eliezer covers these topics and others through allegory, anecdote, and scientific theory. He demonstrates the ideas by applying them to debates in artificial intelligence (AI), physics, metaethics, and consciousness.

To start reading R:A-Z, visit *www.lesswrong.com/rationality or visit Amazon to purchase the e-book or audiobook.*

The Codex

Scott Alexander's, one of LessWrong's earliest and most prolific contributors, wrote many essays on good reasoning, learning from the institution of science, and different ways society has and could be organized. These have been organized into *the Codex*. Scott's sequences include:

- Argument and Analysis
- Studies and Statistics
- Community and Cooperation

His exemplary essays include:

- Beware Isolated Demands for Rigor

- The noncentral fallacy - the worst argument in the world?
- Guided By The Beauty of Our Weapons
- I Can Tolerate Anything Except the Outgroup
- Meditations on Moloch

Harry Potter and Methods of Rationality (HPMOR)

A side project of Eliezer's grew to be one of the most highly rated Harry Potter fanfictions of all time *and* an excellent primer on rationality. Eliezer imagined an alternate-universe Harry Potter who grew up with loving adopted parents, one of them an Oxford scientist. In this version, Harry enters the wizarding world with Enlightenment ideals and the experimental spirit.

We recommend HPMOR to interested in an introduction to rationality via a highly entertaining narrative. Click here to read HPMOR through LessWrong or try the audiobook.

What's with all the AI and math posts?

For both historical reasons and because these topics are relevant to human rationality, many members of the LessWrong community are interested in AI, decision-theory, math, and related topics.

- Historically: LessWrong's founder and author of its foundational works, Eliezer Yudkowsky, is a co-founder of the Machine Intelligence Research Institute and major propopent for AI safety. His writings on LessWrong attracted many people who were interested in both rationality and AI/AI safety, causing these to be ongoing overlap between LessWrong an AI safety communities.
- Relevancy: Artificial intelligence is very much the study of intelligence and how "minds" work. Even if you are more interested in how human minds work and in improving your human rationality, there is much to learn from thinking generally about how intelligence works (for humans or non-humans). In particular, the fields of AI often brings technical precision and rigor to thinking to the gnarly, complicated topics of intelligence and optimal decision-making.
    - Because of this relevance, many writings about human rationality on LessWrong (from Eliezer and others) make reference to concepts from AI and formal decision-theory.

See also: What is the AI Alignment Forum (AIAF) and what does it have to do with LessWrong?

What is the AI Alignment Forum (AIAF) and what does it have to do with LessWrong?

The AI Alignment Forum is an online hub for AI Safety (aka AI alignment) researchers to discuss topics in the field. The AI Alignment forum is another project of the LessWrong team's and resultantly shares some infrastructure with LessWrong proper, i.e. shared user accounts.

Because of the overlaps between the LessWrong and AI Safety communiites and relevance of AI content to rationality, posts made to the AI Alignment forum are automatically cross-posted to LessWrong.

- These posts will have the AIAF symbol (Omega/ Ω) shown next to the title and contain a warning that the content may especially technical.

I (Ruby) am advocating strongly for there to be an easy way to filter these out for users who are not interested in AIAF content.

What is that Omega symbol I see on some posts? Oh, it's AIAF karma.

Posts and comments which been cross-posted from the Alignment Forum will display their *Alignment Forum karma* (symbol: Omega/ Ω). When users with the ability to vote on Alignment Forum content vote on cross-posted AIAF on LessWrong, this will cause both the contents ordinary LessWrong karma and Alignment Forum karma to update.

Posting & Commenting

What can I post on LessWrong?

Posts on practically any topic are welcomed on LessWrong. I (and others on the team) feel it is important that members are able to "bring their entire selves" to LessWrong and are able to share all their thoughts, ideas, and experiences without fearing whether they are "on topic" for LessWrong. Rationality is not restricted to only specific domains of one's life and neither should LessWrong be.

However, to maintain its overall focus while still allowing posts on any topic, LessWrong classifies posts as either Personal blogposts or as Frontpage posts. See more in the post on Personal Blogpost vs Frontpage Posts.

The Editor

LessWrong's editor is what use you to enter posts and comments.

How do I use Markdown? (And not the Draft.js default editor)

By default, LessWrong uses an implementation of Draft.js, however, if you prefer, you can switch to entering your text with markdown syntax. To do, check *Activate Markdown Editor* checkbox in your account settings.

With the Markdown editor activated, you can use Markdown syntax for formatting.

How do I insert images?

If you are using the Draft.Js editor, select some text (or whitespace) and click the image icon in the toolbar that appears your text. Insert a URL to a *hosted image*. The image must be hosted! Use a free online service like Imgur or similar. Ensure you use the url to the hosted image itself, *not* the page displaying uploaded image (common mistake).

Note: image insertions are only enabled for posts, not comments.

If you are using the Markdown editor, using the Markdown syntax for inserting images. It is:

```
![image text](https://www.example.com)
```

As above, the link must be to a hosted image.

How do I insert spoiler protections?

LessWrong gives you a way to “avoid spoiling” your readers. Text is concealed until a user mouses over it (it works a bit less well on mobile right now). This functionality is useful for creating exercises in your posts, e.g. ask a question in your post and conceal with answer beneath spoiler protection so users don’t accidentally see it. See this post as an example.

In the LW docs editor, type `>!` on a new line, then a spoiler box should appear

In the Markdown editor, surround your text with `:::spoiler` at the beginning, and `:::` at the end.

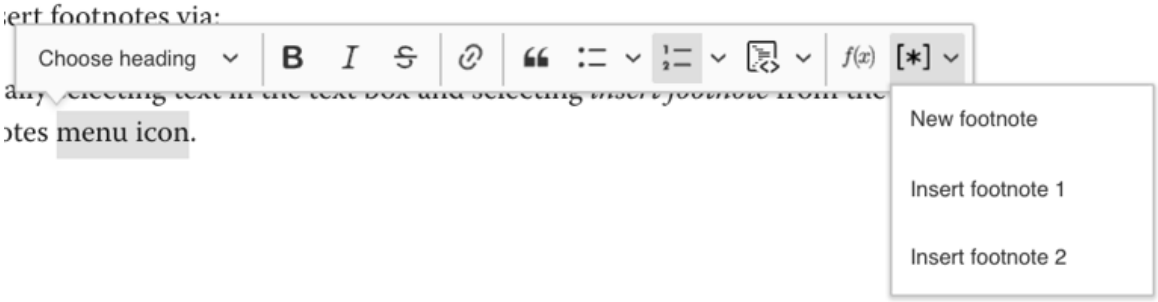
How do I insert footnotes?

Using the Markdown Editor
Use the syntax described here.

Using the LW Docs Editor

You can insert footnotes via:

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



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

2. Using Markdown syntax

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

Footnotes will automatically renumber as you add and delete them!

How do I use Latex?

If using the Draft.js editor, press Cmd-4 for inline and Cmd-M for block-level. (Ctrl on Windows).

If using Markdown, surround your LaTeX text with $, for example:

$<LaTeX text>$

How do I embed interact prediction widgets? (Elicit Predictions)

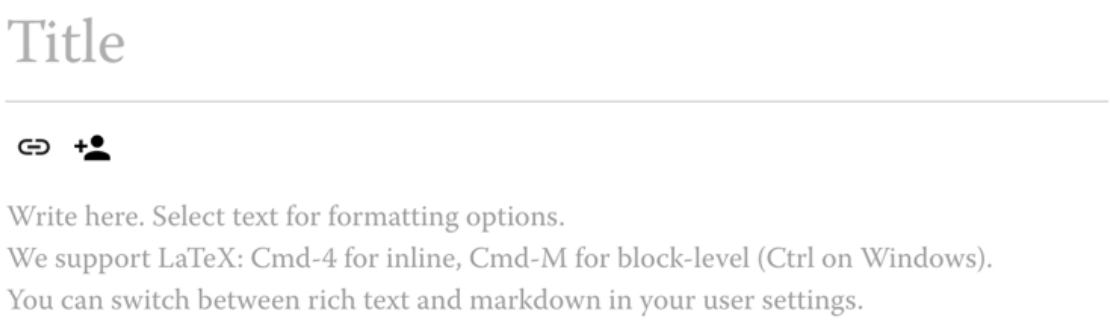
Follow the instructions in this post to insert Embedded Interactive Predictions.

How do I add multiple authors to a post?

If you have 10 or more karma, you can add a co-author to your post in post settings section (bottom of the post edit page). If you don't yet have 10 karma, send us a message on Intercom (bottom right) or email us at team@lesswrong.com and we'll do it for you.

How do I make a linkpost?

At the top of the post editor (underneath "title") is a "link" button. If you click on it, you'll see a field where you can enter a link from another site.



Linkposts that include at least a short description of why the topic is relevant/interesting to LessWrongers tend to get more engagement than linkposts that just include the link by itself.

Creating Sequences

New users can create sequences by going to the Library page, scrolling to the Community Sequences section, and clicking "Create New Sequence." Once created, the sequence will be visible on your User Profile page, and in the Community Sequences section of the Library.

Once you've created at least one sequence, you'll also have access to a "Create New Sequence" button on your User Profile page.

Users with 1000+ karma also have a "New Sequence" menu item in their user menu.

Once a sequence is created, you can add posts to it by clicking "Add/Remove Posts."

Karma & Voting

How do I vote?

Posts and comments have buttons for *upvoting* and *downvoting* them displayed around the posts current *karma* score.

76

Further, you have the option to *strong* upvotes or downvote posts and comments. On desktop: hold the vote button until you see the double bars appear. On mobile: double-tap the vote button (ignore a tool-tip telling you to hold).

What should my votes mean?

We encourage people to vote such that *upvote* means “I want to see more of this” and *downvote* means “I want to see less of this.”

Votes should apply to individual posts/comments, not to overall users. (So, please do not downvote all of a user's historical posts).

What’s the relationship between votes and karma? Why aren’t they the same?

Posts and comments have a karma score. A single vote will increase or decrease the karma by an integer value. Upvotes increase the karma, downvotes decrease - and these can cancel out.

Further, users have karma scores too. A user’s karma score is the sum of all the karma on their posts and comments. **The votes of users with more karma have more power under LessWrong’s voting system**, ensuring that users who have earned the community’s respect and trust have more influence than new sign-ups. Because some users have votes which are worth more than a single point, the karma score of a post is usually greater than the number of votes on it.

What’s the mapping between users’ karma and voting power?

A user’s vote power is determined by the code implemented in this file.

Agree/Disagree Voting

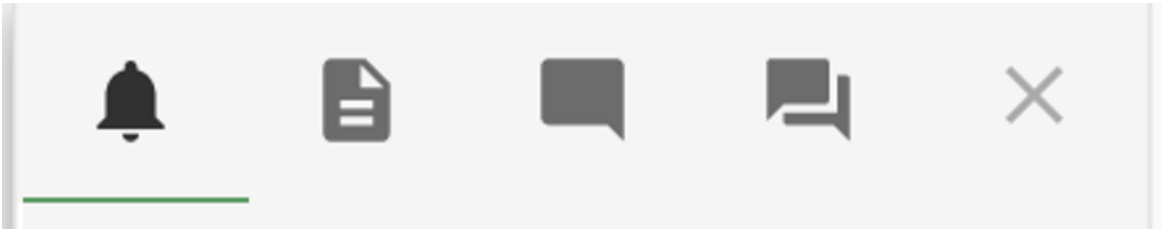
We've recently added agree/disagree voting to comments on posts. As of July 2022, this is still a bit of an experiment. Read more here.

Notifications & Subscriptions

The notification and subscriptions system are currently undergoing a significant upgrade. Expect the functionality to be expanded in the next week or two. We will update this documentation then.

Where do I get notifications?

See the bell icon in the upper right-hand corner. There are four tabs.



Bell: combined responses to your posts and comments + private message notifications

1. Paper/Doc: New post notifications

2. Speech Bubble: Notifications of comments on your posts
3. Two Speech Bubbles: Notifications of private messages on.

What can I get notifications for?

In your account settings you can toggle notifications on and off for responses to your posts and comments.

Can I subscribe by email? What can I subscribe to?

Right now, you can subscribe to receive Curated posts by email or RSS. See the subscribe buttons beneath the Curated posts section on the homepage or in your account settings.

Messaging

How do I sent private messages to other users?

Navigate to a user's page by clicking on an appearance of their username or finding them via search. Click *send message*.

To read your messages, click on the notification icon (bell icon, top right) > click the two speech bubbles on the right. Or visit www.lesswrong.com/inbox.

Questions

What do you mean, questions?

The LessWrong team is actively developing a new experimental Open Questions Research Platform. The vision is to build a system which allows the LessWrong community to apply its high standards of reasoning and scholarship to solving large, important questions.

We expect LessWrong's Open Questions to be valuable beyond existing platforms, e.g. Quora and StackExchange, for multiple reasons. Among them:

- The LessWrong community's focus on good reasoning and commitment to truth
- The design of our tool to be for large [distributed] research questions.

The LessWrong team thinks this is an excellent way to train and apply rationality.

What kind of questions can I ask?

If you have a question which seems like the LessWrong community could answer better than any other Q&A platform, we welcome you to ask it here.

We will handle making sure questions of the right type are shown in each place, so don't worry too much about whether your question is relevant. Like with posts, we welcome questions on most topics and then categorize them appropriately.

Existing questions have been of all the following types:

- Requests for facts
- Requests for answers to difficult research questions

- Requests for explanations of difficult topics
- Requests for arguments for or against a position
- Requests for opinions and insights on a given topic
- Requests for personal advice
- Recommendations, feedback, or request to hear other's personal experiences
- Questions about the LessWrong website

These are all good. Get a sense of what people ask on LessWrong by viewing the questions page.

How do I ask questions?

To ask a question, click on your Username (top right, you must have an account), and click *Ask Question [Beta]*.

How do I write a good question?

The site StackOverflow has a lot of good advice on writing questions. I've copied over some bits and reworded them to fit LessWrong:

Search, and research.

Before posting a question, we strongly recommend that you spend a reasonable amount of time researching the problem and searching for existing questions on this site that may provide an answer. LessWrong has been around for a long time now. Many common confusions are covered in Rationality A-Z, or in the Best of LessWrong . any common questions have already been answered.)

Make sure to keep track of what you find when researching, even if it doesn't help! If you ultimately aren't able to find the answer to your question elsewhere on this site, then including links to related questions (as well as an explanation of why they didn't help in your specific case).

Write a title that summarizes the specific question.

The title is the first thing that potential answerers will see. If your title isn't interesting, they won't read the rest. Also, without a good title, people may not even be able to find your question. So, *make the title count*:

- **Pretend you're talking to a busy colleague** and have to sum up your entire question in one sentence: what details can you include that will help someone identify and solve your problem? Include any error messages, key APIs, or unusual circumstances that make your question different from similar questions already on the site.
- If you're having trouble summarizing the problem, **write the title *last***—sometimes, writing the rest of the question first can make it easier to describe the problem.

How can I helpfully answer questions?

You can probably help more than you think! Even if it's not easy to answer a question outright, small contributions of information or insight can still go a long way.

We encourage you to look through the questions page to find questions that either have existing knowledge about or catch your curiosity about. Read through existing answers and

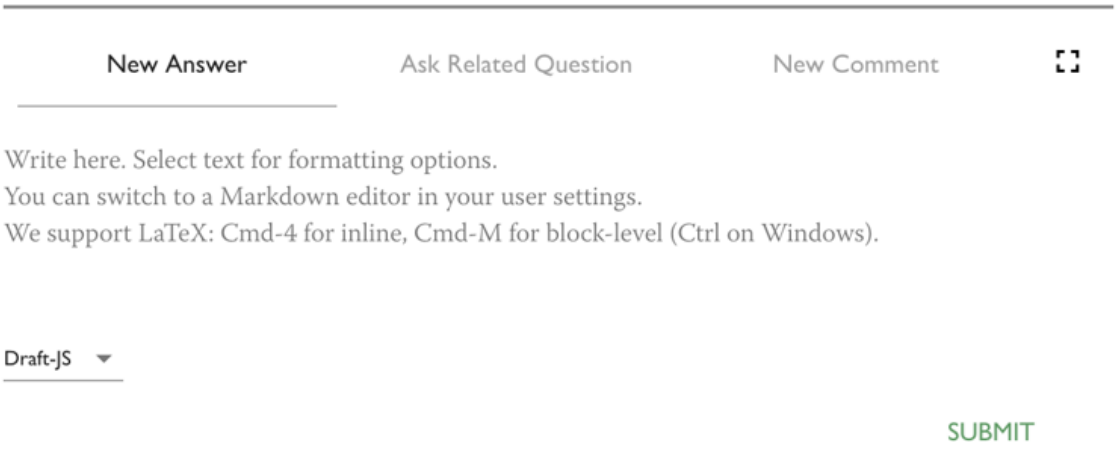
then see what you can add. All of the following can be useful contributions in addition to direct answers:

- A link or recommendation to a resource which might help answer the question.
- A recommendation of who might know the answer that you could talk to.
- A suggestion for what things, if observed, would be evidence about a question one way or another.
- An explanation of how the question is maybe "confused" and should be dissolved.
- Identifying a related or "sub-question" you think will help answer the bigger question.
    - Note the *Ask related question* feature in question pages.

Answering questions is also a great way to practice *the neglected virtue of scholarship*. A couple of LessWrong members have written guides helpful for getting started with scholarship. Lukeprog wrote *Scholarship: How to Do It Efficiently* and gwern wrote a lengthy Internet Search Tips guide.

How do I interact with questions?

Question pages might seem confusing at first. They're not so bad. Beneath the question text you will see a textbook with three options: "New Answer", "Ask Related Question", and "New Comment" as pictured.



New Answer: An answer can be any response which sheds light on the question being asked, even if it's not a complete or comprehensive answer. Some users choose to make smaller contributions as comments. There's a bit of fuzzy line here so don't worry about it too much. You have the ability to move responses back and forth between being comments or answers if you change your mind.

New Comment: Comments on questions can be used to ask clarifying questions and other thoughts which aren't really answers to the question asked. You can also comment on other people's Answers, allowing for discussion of those answers.

Ask Related Question: For large questions, sometimes you can't answer a question directly and instead to ask another question first. You can respond to a question by asking what you think is a related question. These will then be linked in the Question UI.

Asking a (smaller) related question and then making progress on answering it is a great way to help get large research questions answered by the community.

Community Events Page

What is the LessWrong community event page?

LessWrong is both an online and offline community where members around the globe meet up in person for small and large gatherings including local meetups, regional retreats, and conferences.

The community events page is where LessWrong members can find each other in the physical world and create events and groups.

You can find the page at www.lesswrong.com/community, via the left sidebar (desktop) or bottom buttons (mobile).

What are all these categories of meetups?

The community page displays four non-exclusive categories of events and groups. These include explicitly “LessWrong” themed events plus those overlapping and adjacent communities.

- LessWrong
- SlateStarCodex (SSC)
    - SlateStarCodex (SSC) is the personal blog of Scott Alexander who made core contributions to LessWrong. Many of his posts are still cross-posted to LessWrong. The global SSC has much overlap and much in common with the LessWrong community.
- Effective Altruism (EA)
    - Effective Altruism (EA) is a movement and community of people trying to use reason and evidence to do the most good possible. Many LessWrong members also affiliate with the EA community.
- MIRIx
    - MIRIx are workshops for those wishing to be involved in the work of the Machine Intelligence Research Institute (MIRI), an organization working on AI Safety. See What’s with all the AI and math posts?

These four include explicitly LessWrong themed events plus those from overlapping and adjacent communities.

What happens at rationality meetups?

Depends on the meetup! Some meetups focus on formal rationality practice while others are just opportunity’s for like-minded people to socialize - many meetups or groups split their time between the two.

What are the larger community events?

The community events page has information for large events too. Examples include the Bay Area Summer Solstice Celebration, Athena Rationality Workshop, European Community Weekend, and MIRI Summer Fellows Program.

What resources can help me run my local rationality meetup?

There is a resources section on the bottom of the community events page. Just scroll to the bottom!

Moderation

What do LessWrong moderators do?

LessWrong aims to be a well-kept garden. It is warded by a team of active moderators who ensure that discussion and content are of high quality, and that behaviors which would diminish the value of LessWrong are prevented.

Who can moderate on LessWrong?

LessWrong has a split moderation system. Most moderation activity is performed by LessWrong's moderation team; however, users who meet certain karma thresholds can moderate their own posts plus set the *moderation guidelines* that appear on their posts.

Users with over 50 karma can moderate their own posts when they remain as Personal blogposts.

Users with over 2000 karma can moderate their own posts even when they have been promoted to Frontpage status.

What moderation actions can I take on my own posts?

If you meet the karma thresholds (50 on Personal blogposts, 2000 on Frontpage posts), you can perform the following moderation actions on your posts:

- Delete comments
    - Optionally with a public notice and reason.
- Delete comment thread without a trace (deletes all comments and children)
    - Optionally with a private reason sent to the author.
- Ban users from commenting on a given post of yours
- Ban users from commenting on any of your posts

Before you can moderate your own posts, you must set *moderation style* on your post. The following options are available:

- Easy Going - I just delete obvious spam and trolling
- Norm Enforcing - I try to enforce particular rules (See moderation guidelines)
- Reign of Terror - I delete anything I judge to be annoying or counterproductive

If you select *norm enforcing,* you should set your custom moderation policy which will be shown at the top of the comment section and at the bottom of the new comment form of posts you can moderate.

We encourage you to take moderation actions consistent with the moderation policy you have set on your posts.

What actions and duties do the LessWrong team moderators perform?

Moderators perform the following routine regular duties:

- Reviewing all new posts and assigns them spam, personal blogpost, or Frontpage (if the author has permitted Frontpage promotion).
- Reviewing new users when they first comment or post.
- Deleting spam caught by our automatic filters.
- Selecting posts for Curation.
- Keeping an eye on discussions and ensuring they remain productive and civil.

Moderations perform the following less-common actions:

- Issuing feedback and warnings to users who behave in ways harmful to LessWrong's discourse quality and culture.
    - These will usually start with private feedback but escalate to public warnings.
- Banning users. Usually temporarily for a few months or a year.
- Locking comment threads (usually temporarily) if they become overly heated and divisive.
- Limiting the visibility of divisive, heated conversations on the site to protect the culture and what people are exposed to.
    - In one recent instance, we moved one comment thread on a post to a separate post.
    - Moderators can hide discussion threads from the Recent discussion feed on the homepage.

In extremely extreme and exceedingly severe cases:

- The LessWrong team may decide that we cannot display certain content on the site. In this case, we will likely move that content back to a user's drafts.

What powers do moderators have?

Moderators generally have access to the site data, most of this time this is accessed at the request of a user in the process of debugging a technical issue. We take data privacy seriously. We don't just read private messages.

Note: if a comment of yours is ever deleted, you will automatically receive a private message with its contents.

- The ability to delete comments
    - Usually with public notice and reason.
- The ability to delete comment threads without trace (deletes all comments and all its children)
    - Usually, with a private reason send to the author.
- The ability to move content between different classifications, e.g. Personal blogpost, Frontpage post, Curated, Meta (deprecated category) and drafts.
- Moderators can view drafts, but they almost never will unless they're helping you debug something.
- The ability to edit posts.
    - Moderators usually use this to fix awry formatting for you, e.g. your LaTeX is screwed up or egregious typos, leaving a comment saying they have done so.
- The ability to ban users from commenting on given posts or comment threads.
- The ability to ban users from the site (typically done temporarily).
- The ability to lock comment threads.

What is the LessWrong moderation policy/philosophy?

Unfortunately, we do not have a recent and up to date document that speaks coherently for the whole site, however habryka's post on Models of moderation is a good start.

We do have a hard rule against mass downvoting, or voting with sockpoppets.

Who are the moderators?

The LessWrong core team plus a few others form the current moderation team. You can see who they are on the team page.

How do I become a moderator?

We are not currently recruiting any new moderators and there is no current process.

That said, moderators would be recruited from among those we believe possess excellent judgment and understand LessWrong, its purpose, its culture, and its values. The best way to demonstrate this would be through consistently valuable participation on LessWrong.

What is LessWrong's Privacy Policy and Terms of Use?

Our Privacy Policy and Terms of Use can be viewed here.

Note that the Machine Intelligence Research Institute (MIRI) is the relevant legal party for this privacy policy and terms of use. When Eliezer Yudkowsky founded the LessWrong website in 2009, he created it as the property of MIRI (then named the Singularity Institute for Artificial Intelligence, aka SIAI).

While we're at it, we can add that the current LessWrong team operates legally as a part of a related organization, the Center for Applied Rationality (CFAR) while retaining autonomy over its internal decision-making and all decisions about the LessWrong website.

For the intertwined history of MIRI and CFAR, see this answer to a LessWrong question.

Research Agenda v0.9: Synthesising a human's preferences into a utility function

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

I'm now in a position where I can see a possible route to a safe/survivable/friendly Artificial Intelligence being developed. I'd give a 10+% chance of it being possible this way, and a 95% chance that some of these ideas will be very useful for other methods of alignment. So I thought I'd encode the route I'm seeing as research agenda; this is the first public draft of it.

Clarity, rigour, and practicality: that's what this agenda needs. Writing this agenda has clarified a lot of points for me, to the extent that some of it now seems, in retrospect, just obvious and somewhat trivial - "of course that's the way you have to do X". But more clarification is needed in the areas that remain vague. And, once these are clarified enough for humans to understand, they need to be made mathematically and logically rigorous - and ultimately, cashed out into code, and tested and experimented with.

So I'd appreciate any comments that could help with these three goals, and welcome anyone interested in pursuing research along these lines over the long-term.

Note: I periodically edit this document, to link it to more recent research ideas/discoveries.

0 The fundamental idea

This agenda fits itself into the broad family of Inverse Reinforcement Learning: delegating most of the task of inferring human preferences to the AI itself. **Most** of the task, since it's been shown that humans need to build the right assumptions into the AI, or else the preference learning will fail.

To get these "right assumptions", this agenda will look into what preferences actually are, and how they may be combined together. There are hence four parts to the research agenda:

1. A way of identifying the (partial[1]) preferences of a given human H.
2. A way for ultimately synthesising a utility function U_H that is an adequate encoding of the partial preferences of a human H.
3. Practical methods for estimating this U_H, and how one could use the definition of U_H to improve other suggested methods for value-alignment.

4. Limitations and lacunas of the agenda: what is not covered. These may be avenues of future research, or issues that cannot fit into the U_H paradigm.

There has been a myriad of small posts on this topic, and most will be referenced here. Most of these posts are stubs that hint to a solution, rather than spelling it out fully and rigorously.

The reason for that is to check for impossibility results ahead of time. The construction of U_H is deliberately designed to be **adequate**, rather than elegant (indeed, the search for an elegant U_H might be counterproductive and even dangerous, if genuine human preferences get sacrificed for elegance). If this approach is to work, then the safety of U_H has to be robust to different decisions in the synthesis process (see Section 2.8, on avoiding disasters). Thus, initially, it seems more important to find approximate ideas that cover all possibilities, rather than having a few fully detailed sub-possibilities and several gaps.

Finally, it seems that if a sub-problem is not formally solved, we stand a much better chance of getting a good result from "hit it with lots of machine learning and hope for the best", than we would if there were huge *conceptual holes* in the method - a conceptual hole meaning that the relevant solution is broken in an unfixable way. Thus, I'm publishing this agenda now, where I see many implementation holes, but no large conceptual holes.

A word of warning here, though: with some justification, the original Dartmouth AI conference could also have claimed to be confident that there were no large conceptual holes in their plan of developing AI over a summer - and we know how wrong they turned out to be. With that thought in mind, onwards with the research agenda.

0.1 Executive summary: synthesis process

The first idea of the project is to identify partial preferences as residing within human mental models. This requires identifying the actual and hypothetical internal variables of a human, and thus solving the "symbol grounding problem" for humans; ways of doing that are proposed.

The project then sorts the partial preferences into various categories of interest (basic preferences about the world, identity preferences, meta-preferences about basic preferences, global meta-preferences about the whole synthesis project, etc...). The aim is then to synthesise these into a single utility function U_H, representing the preference of the human H (at a given time or short interval of time). Different preference categories play different roles in this synthesis (eg object-level preferences get aggregated, meta-preferences can modify the weights of object-level preferences, global meta-preferences are used at the design stage, and so on).

The aims are to:

1. Ensure the synthesis U_H has good properties and reflects H's actual preferences, **and not any of** H**'s erroneous factual beliefs**.
2. Ensure that highly valued preferences weight more than lightly held ones, even if the lightly held one is more "meta" that the other.
3. Respect meta-preferences about the synthesis as much as possible, but...
4. ...always ensure that the synthesis actually reaches an actual non-contradictory U_H.

To ensure point 4. and 2., there will always be an initial way of synthesising preferences, which certain meta-preferences can then modify in specific ways. This is designed to resolve contradictions (when "I want a simple moral system" and "value is fragile and needs to be preserved" are both comparably weighted meta-preferences) and remove preference loops ("I want a simple moral system" is itself simple and could reinforce itself; "I want complexity in my values" is also simple and could undermine itself).

The "good properties" of 1. are established, in large part, by the global meta-preferences that don't comfortably sit within the synthesis framework. As for erroneous beliefs, if H wants to date H′ because they think that would make them happy and respected, then an AI will synthesise "being happy" and "being respected" as preferences, and would push H away from H′ if H were actually deluded about what dating them would accomplish.

That is the main theoretical contribution of the research agenda. It then examines what could be done with such a theory in practice, and whether the theory can be usefully approximated for constructing an actual utility function for an AI.

0.2 Executive summary: agenda difficulty and value

One early commentator on this agenda remarked:

> [...] it seems like this agenda is trying to solve at least 5 major open problems in philosophy, to a level rigorous enough that we can specify them in code:
>
> 1. The symbol grounding problem.
> 2. Identifying what humans really care about (not just what they say they care about, or what they act like they care about) and what preferences and meta-preferences even are.
> 3. Finding an acceptable way of making incomplete and inconsistent (meta-)preferences complete and consistent.
> 4. Finding an acceptable way of aggregating many people's preferences into a single function[2].
> 5. The nature of personal identity.
>
> I agree that AI safety researchers should be more ambitious than most researchers, but this seems extremely ambitious, and I haven't seen you acknowledge the severe outside-view difficulty of this agenda.

This is indeed an extremely ambitious project. But, in a sense, a successful aligned AI project will *ultimately* have to solve all of these problems. Any situation in which most of the future trajectory of humanity is determined by AI, is a situation where there are solutions to all of these problems.

Now, these solutions may be implicit rather than explicit; equivalently, we might be able to delay solving them via AI, for a while. For example, a tool AI solves these issues by being contained in such a way that human judgement is capable of ensuring good outcomes. Thus *humans* solve the grounding problem, and we design our questions to the AI to ensure compatibility with our preferences, and so on.

But as the power of AIs increase, humans will become confronted by situations they have never been in before, and our ability to solve these issues diminish (and the probabilities increase that we might be manipulated or fall into a bad attractor). This transition may sneak up on us, so it is useful to start thinking of how to a) start solving these problems, and b) start identifying these problems crisply so we can know when and whether they need to be solved, and when we are moving out of the range of validity of the "trust humans" solution. For both these reasons, all the issues will be listed explicitly in the research agenda.

A third reason to include them is so that we know what we need to solve those issues *for*. For example, it is easier to assess the quality of any solution to symbol grounding, if we know what we're going to do with that solution. We don't need a full solution, just one good enough to define human partial preferences.

And, of course, we need to also consider scenarios where partial approaches like tool AI just don't work, or only work if we solve all the relevant issues anyway.

Finally, there is a converse: partial solutions to problems in this research agenda can contribute to improving other methods of AI safety alignment. Section 3 will look into this in more detail. The basic idea is that, to improve an algorithm or an approach, it is very useful to know what we are ultimately trying to do (eg compute partial preferences, or synthesise a utility function with certain acceptable properties). If we rely only on making local improvements, guided by intuition, we may ultimately get stuck when intuition runs out; and the improvements are more likely to be ad-hoc patches than consistent, generalisable rules.

0.3 Executive aside: the value of approximating the theory

The theoretical construction of U_H in Sections 1 and 2 is a highly complicated object, involving millions of unobserved counterfactual partial preferences and a synthesis process involving higher-order meta-preferences. Section 3 touches on how U_H could be approximated, but, given its complexity, it would seem that the answer would be "only very badly".

And there is a certain sense in which this is correct. If U_V is the actual idealised utility defined by the process, and V_H is the approximated utility that a real-world AI could

compute, then it is likely[3] that U_H and V_H will be quite different in many formal senses.

But there is a certain sense in which this is incorrect. Consider many of the AI failure scenarios. Imagine that the AI, for example, extinguished all meaningful human interactions because these can sometimes be painful and the AI knows that we prefer to avoid pain. But it's clear to us that most people's partial preferences will not endorse total loneliness as good outcome; if it's clear to us, then it's *a fortiori* clear to a very intelligent AI; hence the AI will avoid that failure scenario.

One should be careful with using arguments of this type, but it is hard to see how there could be a failure mode that a) we would clearly understand is incompatible with proper synthesis of U_H, but b) a smart AI would not. And it seems that any failure mode should be understandable to us, as a failure mode, especially given some of the innate conservatism of the construction of U_H.

Hence, even if V_H is a poor approximation of U_H in a certain sense, it is likely an excellent approximation of V_H in the sense of avoiding terrible outcomes. So, though $d(U_H, V_H)$ might be large for some formal measure of distance d, a world where the AI maximises V_H will be highly ranked according to U_H.

0.4 An inspiring just-so story

This is the story of how evolution created humans with preferences, and what the nature of these preferences are. The story is not true, in the sense of accurate; instead, it is intended to provide some inspiration as to the direction of this research agenda. This section can be skipped.

In the beginning, evolution created instinct driven agents. These agents had no preferences or goals, nor did they need any. They were like Q-learning agents: they knew the correct action to take in different circumstances, but that was it. Consider baby turtles that walk towards the light upon birth, because, traditionally, the sea was lighter than the land - of course, this behaviour fails them in the era of artificial lighting.

But evolution has a tiny bandwidth, acting once per generation. So it created agents capable of planning, of figuring out different approaches, rather than having to follow instincts. This was useful, especially in varying environments, and so evolution offloaded a lot of its "job" onto the planning agents.

Of course, to be of any use, the planning agents need to be able to model their environment to some extent (or else their plans can't work) and had to have preferences (or else every plan was as good as another). So, in creating the first planning agents, evolution created the first agents with preferences.

Of course, evolution is a messy, undirected process, so the process wasn't clean. Planning agents are still riven with instincts, and the modelling of the environment is

situational, used for when it was needed, rather than some consistent whole. Thus the "preferences" of these agents were underdefined and sometimes contradictory.

Finally, evolution created agents capable of self-modelling and of modelling other agents in their species. This might have been because of competitive social pressures as agents learn to lie and detect lying. Of course, this being evolution, this self-and-other-modelling took the form of kludges built upon spandrels built upon kludges.

And then arrived humans, who developed norms and norm-violations. As a side effect of this, we started having higher-order preferences as to what norms and preferences *should* be. But instincts and contradictions remained - this is evolution, after all.

And evolution looked upon this hideous mess, and saw that it was good. Good for evolution, that is. But if we want it to be good for us, we're going to need to straighten out this mess somewhat.

1 The partial preferences of a human

The main aim of this research agenda is to start with a human H at or around a given moment t and produces a utility function U_{H_t} which is an adequate synthesis of the human's preferences at the time t. Unless the dependence on t needs to be made explicit, this will simply be designated as U_H.

Later sections will focus on what can be done with U_H or the methods used for its construction; this section and the next will focus solely on that construction. It is mainly based on these posts, with some commentary and improvements.

Essentially the process is to identify human preferences and meta-preferences within human (partial) mental model (Section 1), and find some good way of synthesising these into a whole U_H (Section 2).

Partial preferences (see Section 1.1) will be decomposed into:

1. Partial preferences about the world.
2. Partial preferences about our own identity.
3. Partial meta-preferences about our preferences.
4. Partial meta-preferences about the synthesis process.
5. Self-referential contradictory partial meta-preferences.
6. Global meta-preferences about the outcome of the synthesis process.

This section and the next will lay out how preferences of types 1, 2, 3, and 4 can be used to synthesise the U_H. Section 2 will conclude by looking what role preferences of type 6 can play. Preferences of type 5 are not dealt with in this agenda, and remain a perennial problem (see Section 4.5).

1.1 Partial models, partial preferences

As was shown in the paper "Occam's razor is insufficient to infer the preferences of irrational agents", an agent's behaviour is never enough to establish their preferences - even with simplicity priors or regularisation (see also this post and this one).

Therefore a definition of preference needs to be grounded in something other than behaviour. There are further arguments, presented here, as to why a theoretical grounding is needed even when practical methods are seemingly adequate; this point will be returned to later.

The first step is to define a partial preference (and a partial model for these to exist in). A partial preference is a preference that exists within a human being's internal mental model, and which contrasts two[4] situations along a single axis of variation, keeping other aspects constant. For example, "I wish I was rich (rather than poor)", "I don't want to go down that alley, lest I get mugged", and "this is much worse if there are witnesses around" are all partial preferences. A more formal definition of partial preferences, and the partial mental model in which they exist, is presented here.

Note that this is one of the fundamental theoretical underpinnings of the method. It identifies human (partial) preferences as existing within human mental models. This is a "normative assumption": we *choose* to define these features as (partial) human preferences, the universe does not compel us to do so.

This definition gets around the "Occam's razor" impossibility result, since these mental models are features of the human brain's internal process, not of human behaviour. Conversely, this also violates certain versions of functionalism, precisely because the internal mental states are relevant.

A key important feature is to extract not only the partial preferences itself, but the intensity of the preferences, referred to as its *weight*. This will be key in combining the preferences together (technically, we only need the weight relative to other partial preferences).

1.2 Symbol grounding

In order to interpret what a partial model means, we need to solve the old problem of symbol grounding. "I wish I was rich" was presented as an example of a partial preference; but how can we identify "I", "rich" and the counterfactual "I wish", all within the mess of the neural net that is the human brain?

To ground these symbols, we should approach the issue of symbol grounding empirically, by aiming to predict the values of real world-variables through knowledge of internal mental variables (see also the example presented here). This empirical approach can provide sufficient grounding for the purposes of partial models, even if symbol grounding is not solved in the traditional linguistic sense of the problem.

This is because each symbol has a web of connotations, a collection of other symbols and concepts that co-vary with it, in normal human experience. Since the partial models are generally defined to be within normal human experiences, there is little difference between any symbols that are strongly correlated.

To formalise and improve this definition, we'll have to be careful about how we define the internal variables in the first place - overly complicated or specific internal

variables can be chosen to correlate artificially well with external variables. This is, essentially, "symbol grounding overfitting".

Another consideration is the extent to which the model is conscious or subconscious; aliefs, for example, could be modelled as subconscious partial preferences. For consciously endorsed aliefs, this is not much of a problem - we instinctively fear touching fires, and don't desire to lose that fear. But if we don't endorse that alief - for example, we might fear flying and not want to fear it - this becomes more tricky. Things get confusing with partially endorsed aliefs: amusement park rides are extremely safe, and we wouldn't want to be crippled with fear at the thought of going on one. But neither would we want the experience to feel perfectly bland and safe.

1.3 Which (real and hypothetical) partial models?

Another important consideration is that humans do not have, at the moment t, a complete set of partial models and partial preferences. They may have a single partial model in mind, with maybe a few others in the background - or they might not be thinking about anything like this at all. We could extend the parameters to some short period *around* the time t (reasoning that people's preferences rarely change in such a short time), but though that gives us more data, it doesn't give us nearly enough.

The most obvious way to get a human to produce an internal model is to ask them a relevant question. But we have to be careful about this - since human values are changeable and manipulable, the very act of asking a question can cause humans to think in certain directions, and even create partial preferences where none existed. The more interaction between the questioner and the human, the more extreme preferences can be created. If the questioner is motivated to maximise the utility function that it is also computing (i.e. if the U_H is an *online* learning process), then the questioner can rig or influence the learning process.

Fortunately, there are ways of removing the questioner's incentives to rig or influence the learning process.

Thus the basic human preferences at time t are defined to be those partial models produced by "one-step hypotheticals"[5]. These are questions that do not cause the human to be put in unusual mental situations, and try and minimise any departure from the human's base-state. We need to distinguish between simple and composite partial preferences: the latter happen when a hypothetical question elicits a long chain of reasoning, covering multiple partial preferences, rather than a single clear answer based on a single internal model.

Some preferences are conditional (eg "I want to eat something different from what I've eat so far this week"), as are some meta-preferences (eg "If I hear a convincing argument about X being good, I want to prefer X"), which could violate the point of the one-step hypothetical. Thus conditional (meta-)preferences are only acceptable if their conditions are achieved by short streams of data, unlikely to manipulate the human. They also should be weighted more if they fit a consistent narrative of what the

human is/wants to be, rather than being ad hoc (this will be assessed by machine learning, see Section 2.4).

Note that among the one-step hypotheticals, are included questions about rather extreme situations - heaven and hell, what to do if plants were conscious, and so on. In general, we should reduce the weight[6] of partial preferences in extreme situations[7]. This is because of the unfamiliarity of these situations, and because the usual human web of connotations between concepts may have broken down (if a plant was conscious, would it be a plant in the sense we understand that?). Sometimes the breakdown is so extreme that we can say that the partial preference is factually wrong. This includes effects like the hedonic treadmill: our partial models of achieving certain goals often include an imagined long-term satisfaction that we would not actually feel. Indeed, it might be good to specifically avoid these extreme situations, rather than having to make a moral compromise that might lose part of H's values due to uncertainty. In that case, ambiguous extreme situations get a slight intrinsic negative - that might be overcome by other considerations, but is there nonetheless.

A final consideration is that some concepts just disintegrate in general environments - for example, consider a preference for "natural" or "hand-made" products. In those cases, the web of connotations can be used to extract some preferences in general - for example, "natural", used in this way has connotations[8] of "healthy", "traditional", and "non-polluting", all of which extend better to general environments than "natural" does. Sometimes, the preference can be preserved but routed around: some versions of "no artificial genetic modifications" could be satisfied by selective breeding that achieved the same result. And some versions couldn't; it's all a function of what powers the underlying preference: specific techniques, or a general wariness of these types of optimisation. Meta-preferences might be very relevant here.

2 Synthesising the preference utility function

Here we will sketch out the construction of the human utility function U_H, from the data that is the partial preferences and their (relative) weights.

This is not, by any means, the only way of constructing U_H. But it is illustrative of how the utility could be constructed, and can be more usefully critiqued and analysed than a vaguer description.

2.1 What sort of utility function?

Partial preferences are defined over states of the world or states of the human H. The later included both things like "being satisfied with life" (purely internal) and "being an honourable friend" (mostly about H's behaviour).

Consequently, U_H must also be defined over such things, so U_H is dependent on states of the world and states of the human H. Unlike standard MDP-like situations, these states can include the history of the world or of H up to that point - preferences like "don't speak ill of the dead" abound in humans.

2.2 Why a utility function?

Why should we aim to synthesise a utility function, when human preferences are very far from being utility functions?

It's not of an innate admiration for utility functions, or a desire for mathematical elegance. It's because they tend to be stable under self-modification. Or, to be more accurate, they seem to be much more stable than preferences that are not utility functions.

In the imminent future, human preferences are likely to become stable and unchanging. Therefore it makes more sense to create a preference synthesis that is already stable, that create a potentially unstable one and let it randomly walk itself to stability (though see Section 4.6).

Also, and this is one of the motivations behind classical inverse reinforcement learning, reward/utility functions tend to be quite portable, and can be moved from one agent to another or from one situation to another, with greater ease than other goal structures.

2.3 Extending and normalising partial preferences

Human values are changeable, manipulable, underdefined, and contradictory. By focusing around time t, we have removed the changeable problem for partial preferences (see this post for thoughts on how long a period around t should be allowed); manipulable has been dealt with by removing the possibility of the AI influencing the learning process.

Being underdefined remains a problem, though. It would be possible to overfit absurdly specifically to the human's partial models, and generate a U_H that is in full agreement with our partial preferences and utterly useless. So the first thing to do is to group the partial preferences together according to similarity (for example, preferences for concepts closely related in terms of webs of connotations should generally be grouped together), and generalise them in some regularised way. Generalise means, here, that they are transformed into full preferences, comparing all possible universes. Though this would only be comparing on the narrow criteria that were used for the partial preference: a partial preference fear of being mugged could generalise to a fear of pain/violence/violation/theft across all universes, but would not include other aspects of our preferences. So they are full preferences, in terms of applying to all situations, but not the full set of our preferences, in terms of taking into account all our partial preferences.

It seems that standard machine learning techniques should already be up to the task of making full preferences from collections of partial preferences (with all the usual current problems). For example, clustering of similar preferences would be necessary. There are unsupervised ML algorithms that can do that; but even supervised ML algorithms end up grouping labelled data together in ways that define extensions of the labels into higher dimensional space. Where could these labels come from? Well, they could come from grounded symbols within meta-preferences. A meta-preference of the form "I would like to be free of bias" contains some model of what "bias" is; if that meta-preference is particularly weighty, then clustering preferences by whether or not they are biases could be a good thing to do.

Once the partial preferences are generalised in this way, remains the problem of them being contradictory. This is not as big a problem as it may seem. First of all, it is very rare for preferences to be utterly opposed: there is almost always some compromise available. So an altruist with murderous tendencies could combine charity work with aggressive online gaming; indeed some whole communities (such as BDSM) are designed to balance "opposing" desires for risk and safety.

So in general, the way to deal with contradictory preferences is to weight them appropriately, then add them together; any compromise will then appear naturally from the weighted sum[9].

To do that, we need to normalise the preferences in some way. We might seek to do this in an a priori, principled way, or through partial models that include the tradeoffs between different preferences. Preferences that pertain to extreme situations, far removed from everyday human situations, could also be penalised in this weighting process (as the human should be less certain about these).

Now that the partial preferences have been identified and weighted, the challenge is to synthesise them into a single U_H.

2.4 Synthesising the preference function: first step

So this is how one could do the first step of preference synthesis:

1. Group similar partial preferences together, generalise them to full preferences without overfitting.
2. Use partial models to compute the relative weight between different partial preferences.
3. Using those relative weights, and again without overfitting, synthesise those preferences into a single utility function U_H^0.

This all seems doable in theory within standard machine learning. See Section 2.3 and the discussion of clustering for point 1. Point 2. comes from the definition of partial preferences. And point 3. is just an issue of fitting a good regularised approximation to noisy data.

In certain sense, this process is the partial opposite how Jacob Falkovich used a spreadsheet to find a life partner. In that process, he started by factoring his goal of

having a life-partner in many different subgoals. He then ranked the putative partners on each of the subgoals by comparing two options at a time, and building a (cardinal) ranking from these comparisons. The process here also aims to assign cardinal values from comparisons of two options, but the construction of the "subgoals" (full preferences) is handled by machine learning from the sets of weighted comparisons.

2.5 Identity preferences

Some preferences are best understood as pertaining to our own identity. For example, *I* want to understand how black holes work; this is separate from my other preference that *some* humans understand black holes (and separate again from an instrumental preference that, had we a convenient black hole close to hand, that we could use it to get energy out of).

Identity preferences seem to be different from preferences about the world; they seem more fragile than other preferences. We could combine identity preference differently from standard preferences, for example using smoothmin rather than summation. Gratifications seem to be particular types of identity preferences: these are preferences about how we achieved something, rather than what we achieved (eg achieving a particularly clever or impressive victory in a game, rather than just achieving a victory).

Ultimately, the human's mental exchange rate between preferences should determine how preferences are combined. This should allow us to treat identity and world-preferences in the same way. There are two reasons to still distinguish between world-preferences and identity preferences:

1. For preferences where relative weights are unknown or ill-defined, linear combinations and smooth-min serve as a good default for world-preferences and identity preferences respectively.
2. It's not certain that identity can be fully captured by partial preferences; in that case, identity preferences could serve as a starting point from which to build a concept of human identity.

2.6 Synthesising the preference function: meta-preferences

Humans generally have meta-preferences: preferences over the kind of preferences they should have (often phrased as preferences over their identity, eg "I want to be more generous", or "I want to have consistent preferences").

This is such an important feature of humans, that it needs its own treatment; this post first looked into that.

The standard meta-preferences endorse or unendorse lower lever preferences. First one can combine them as in the method above, and get a synthesised meta-preference. Then this increases or decreases the weights of the lower level preferences, to reach a U_H^n with preference weights adjusted by the synthesised meta-preferences.

Note that this requires some ordering of the meta-preferences: each meta-preference refers only to meta-preferences "below" itself. Self-referential meta-preferences (or, equivalently, meta-preferences referring to each other in a cycle) are more subtle to deal with, see Section 4.5.

Note that an ordering does not mean that the higher meta-preferences must dominate the lower ones; a weakly held meta-preference (eg a vague desire to fit in with some formal standard of behaviour) need not overrule a strongly held object level preference (eg a strong love for a particular person, or empathy for an enemy).

2.7 Synthesising the preference function: meta-preference about synthesis

In a special category are the meta-preference about the synthesis process itself. For example, philosophers might want to give greater weight to higher order meta-preferences, or might value the simplicity of the whole U_H.

One can deal with that by using the standard synthesis (of Section 2.4) to combine the method meta-preferences, then use this combination to change how standard preferences are synthesised. This old post has some examples of how this could be achieved. Note that these meta-preferences include preferences over using rationality to decide between lower-level preference.

As long as there is an ordering of meta-preferences about synthesis, one can use the standard method to synthesise the highest level of meta-preferences, which then tells us how to synthesise the lower-level meta-preferences about synthesis, and so on.

Why use the standard synthesis method for these meta-preferences - especially if they contradict this synthesis method explicitly? There are three reasons for this:

1. These meta-preferences may be weakly weighted (hence weakly held), so they should not automatically overwhelm the standard synthesis process when applied to themselves (think of continuity as the weight of the meta-preference fades to zero).
2. Letting meta-preferences about synthesis determine how they themselves get synthesised leads to circular meta-preferences, which may cause problems (see Section 4.5).
3. The standard method is more predictable, which makes the whole process more predictable; self-reference, even if resolved, could lead to outcomes randomly far away from the intended one. Predictability could be especially important for "meta-preferences over outcomes" of the next section.

Note that these synthesis meta-preferences should be of a type that affects the synthesis of U_H, not its final form. So, for example, "simple (meta-)preferences should be given extra weight in U_H" is valid, while "U_H should be simple" is not.

Thus, finally, we can combine everything (except for some self-referencing contradictory preferences) into one U_H.

Note there are many degrees of freedom in how the synthesis could be carried out; it's hoped that they don't matter much, and that each of them will reach a U_H that avoids disasters[10] (see Section 2.8).

2.8 Avoiding disasters, and global meta-preferences

It is important that we don't end up in some disastrous outcome; the very definition of a good human value theory requires this.

The approach has some in-built protection against many types of disasters. Part of that is that it can include very general and universal partial preferences, so any combination of "local" partial preferences must be compatible with these. For example, we might have a collection of preferences about autonomy, pain, and personal growth. It's possible that, when synthesising these preferences together, we could end up with some "kill everyone" preference, due to bad extrapolation. However, if we have a strong "don't kill everyone" preference, this will push the synthesis process away from that outcome.

So some disastrous outcomes of the synthesis should be avoided, precisely because *all* of H's preferences are used, including those that would specifically label that outcome a disaster.

But, even if we included all of H's preferences in the synthesis, we'd still want to be sure we'd avoided disasters.

In one sense, this requirement is trivially true and useful. But in another, it seems perverse and worrying - the U_H is supposed to be a synthesis of true human preferences. By definition. So how could this U_H be, in any sense, a disaster? Or a failure? What criteria - apart from our own preferences - could we use? And shouldn't we be using these preferences in the synthesis itself?

The reason that we can talk about U_H not being a disaster, is that not all our preferences can best be captured in the partial model formalism above. Suppose one fears a siren world or reassures oneself that we can never encounter an indescribable hellworld. Both of these could be clunkily transformed into standard meta-preferences (maybe about what some devil's advocate AI could tell us?). But that somewhat misses the point. These top-meta-level considerations live most naturally at the top-meta-level: reducing them to the standard format of other preferences and meta-preferences risks losing the point. Especially when we only partially understand these issues, translating them to standard meta-preferences risks losing the understanding we do have.

So, it remains possible to say that U_H is "good" or "bad", using higher level considerations that are difficult to capture entirely within U_H.

For example, there is an argument that human preference incoherence should not cost us much. If true, this argument suggests that overfitting to the details of human preferences is not as bad as we might fear. One could phrase this as a synthesis meta-preference allowing more over-fitting, but this doesn't capture a coherent meaning of "not as bad" - which precludes the real point of this argument, which is "allow more overfitting if the argument holds". To use that, we need some criteria for establishing "the argument holds". This seems very hard to do within the synthesis process, but could be attempted as top-level meta-preferences.

We should be cautious and selective when using these top-level preferences in this way. This is not generally the point at which we should be adding preferences to U_H; that should be done when constructing U_H. Still, if we have a small selection of criteria, we could formalise these and check ourselves whether U_H satisfies them, or have an AI do so while synthesising U_H. A Last Judge can be a sensible precaution (especially if there are more downsides to error than upsides to perfection).

Note that we need to distinguish between the global meta-preferences of the designers (us) and those of the subject H. So, when designing the synthesis process, we should either allow options to be automatically changed by H's global preferences, or be aware that we are overriding them with our own judgement (which may be inevitable, as most H's have not thought deeply about preference synthesis; still, it is good to be aware of this issue).

This is also the level at which experimental testing of U_H synthesis is likely to be useful - keeping in mind what we expect from U_H synthesis, and running the synthesis in some complicated toy environments, we can see whether our expectations are correct. We may even discover extra top-level desiderata this way.

2.9 How much to delegate to the process

The method has two types of basic preferences (world-preferences and identity preferences). This is a somewhat useful division; but there are others that could have been used. Altruistic versus selfish versus anti-altruistic preferences is a division that was not used (though see Section 4.3). Moral preferences were not directly distinguished from non-moral preferences (though some human meta-preferences might make the distinction).

So, why divide preferences this way, rather than in some other way? The aim is to allow the process itself to take into account most of the divisions that we might care about; things that go into the model explicitly are structural assumptions that are of vital importance. So the division between world- and identity preferences was chosen because it seemed absolutely crucial to get that right (and to err on the side of caution in distinguishing the two, even if our own preferences don't distinguish them as much). Similarly, the whole idea of meta-preferences seems a crucial feature of humans, which might not be relevant for general agents, so it was important to capture it. Note that meta-preferences are treated as a different type to standard

preferences, with different rules; most distinctions built into the synthesis method should similarly be between objects of a different type.

But this is not set in stone; global meta-preferences (see Section 2.8) could be used to justify a different division of preference types (and different methods of synthesis). But it's important to keep in mind what assumptions are being imposed from outside the process, and what the method is allowed to learn during the process.

3 U_H in practice

3.1 Synthesis of U_H in practice

If the definition of U_H of the previous section could be made fully rigorous, and if the AI has a perfect model of H's brain, knowledge of the universe, and unlimited computing power, it could construct U_H perfectly and directly. This will almost certainly not be the case; so, do all these definitions give us something useful to work with?

It seems they do. Even extreme definitions can be approximated, hopefully to some good extent (and the theory allows us to assess the quality of the approximation, as opposed to another method without theory, where there is no meaningful measure of approximation ability). See Section 0.3 for an argument as to why even very approximate versions of U_H could result in very positive outcomes: even approximated U_H rule out most bad AI failure scenarios.

In practical terms, the synthesis of U_H from partial preferences seems quite robust and doable; it's the definition of these partial preferences that seems tricky. One might be able to directly see the internal symbols in the human brain, with some future super-version of fMRI. Even without that direct input, having a theory of what we are looking for - partial preference in partial models with human symbols grounded - allows us to use results from standard and moral psychology. These results are insights into behaviour, but they are often also, at least in part, insights into how the human brain processes information. In Section 3.3, we'll see how the definition of U_H allows us to "patch" other, more classical methods of value alignment. But the converse is also true: with a good theory, we can use more classical methods to figure out U_H. For example, if we see H as being in a situation where they are likely to tell the truth about their internal model, then their stated preferences become good proxies for their internal partial preferences.

If we have a good theory for how human preferences change over time, then we can use preferences at time t' as evidence for the hypothetical preferences at time t. In

general, more practical knowledge and understanding would lead to a better understanding of the partial preferences and how they change over time.

This could become an area of interesting research; once we have a good theory, it seems there are many different practical methods that suddenly become usable.

For example, it seems that humans model themselves and each other using very similar methods. This allows us to use our own judgement of irrationality and intentionality, to some extent, and in a principled way, to assess the internal models of other humans. As we shall see in Section 3.3, an awareness of what we are doing - using the similarity between our internal models and those of others - also allows us to assess when this method stops working, and patch it in a principled way.

In general, this sort of research would give results of the type "assuming this connection between empirical facts and internal models (an assumption with some evidence behind it), we can use this data to estimate internal models".

3.2 (Avoiding) uncertainty and manipulative learning

There are arguments that, as long as we account properly for our uncertainty and fuzziness, there are no Goodhart-style problems in maximising an approximation to U_H. This argument has been disputed, and there are ongoing debates about it.

With a good definition of what it means for the AI to influence the learning process, online learning of U_H becomes possible, even for powerful AIs learning over long periods of time in which the human changes their views (either naturally or as a consequence of the AI's actions).

Thus, we could construct an online version of inverse reinforcement learning without assuming rationality, where the AI learns about partial models and human behaviour simultaneously, constructing the U_H from observations given the right data and the right assumptions.

3.3 Principled patching of other methods

Some of the theoretical ideas presented here can be used to improve other AI alignment ideas. This post explains one of the ways this can happen.

The basic idea is that there exist methods - stated preferences, revealed preferences, an idealised human reflecting for a very long time - that are often correlated with U_H and with each other. However, all of the methods fail - stated preferences are often dishonest (the revelation principle doesn't apply in the social world), revealed preferences assume a rationality that is often absent in humans (and some models of revealed preferences obscure how unrealistic this rationality assumption is), humans that think for a long time have the possibility of value drift or random walks to convergence.

Given these flaws, it is always tempting to patch the method: add caveats to get around the specific problem encountered. However, if we patch and patch until we can no longer think of any further problems, that doesn't mean there are no further problems: simply that they are likely beyond our capacity to predict ahead of time. And, if all that it has is a list of patches, the AI is unlikely to be able to deal with these new problems.

However, if we keep the definition of U_H in mind, we can come up with principled reasons to patch a method. For example, lying on stated preferences means a divergence between stated preferences and internal model; revealed preferences only reveal within the parameters of the partial model that is being used; and value drift is a failure of preference synthesis.

Therefore, each patch can have an explanation for the divergence between method and desired outcome. So, when the AI develops the method further, it can itself patch the method, when it enters a situation where a similar type of divergence. It has a reason for *why* these patches exist, and hence the ability to generate new patches efficiently.

3.4 Simplified U_H sufficient for many methods

It's been argued that many different methods rely upon, if not a complete synthesis U_H, at least some simplified version of it. Corrigibility, low impact, and distillation/amplification all seem to be methods that require some simplified version of U_H.

Similarly, some concepts that we might want to use or avoid - such as "manipulation" or "understanding the answer" - also may require a simplified utility function. If these concepts can be defined, then one can disentangle them from the rest of the alignment problem, allowing us to instructively consider situations where the concept makes sense.

In that case, a simplified or incomplete construction of U_H, using some simplification of the synthesis process, might be sufficient for one of the methods or definitions just listed.

3.5 Applying the intuitions behind U_H to analysing other situations

Finally, one could use the definition of U_H as inspiration when analysing other methods, which could lead to interesting insights. See for example these posts on figuring out the goals of a hierarchical system.

4 Limits of the method

This section will look at some of the limitations and lacuna of the method described above. For some limitations, it will suggest possible ways of dealing with them; but these are, deliberately, chosen to be extras beyond the scope of the method, where synthesising U_H is the whole goal.

4.1 Utility at one point in time

The U_H is meant to be a synthesis of the *current* preferences and meta-preferences of the human H, using one-step hypotheticals to fill out the definition. Human preferences are changeable on a short time scale, without us feeling that we become a different person. Hence it may make sense to replace U_{H_t} with some average U_H, averaged over a short (or longer) period of time. Shorter period lead to more "overfitting" to momentary urges; longer period allow more manipulation or drift.

4.2 Not a philosophical ideal

The U_H is also not a reflective equilibrium or other idealised distillation of what preferences *should be*. Philosophers will tend to have a more idealised U_H, as will those who have reflected a lot and are more willing to be bullet swallowers/bullet bitters. But that is because these people have strong meta-preferences that push in those idealised directions, so any honest synthesis of their preferences must reflect these.

Similarly, this U_H is defined to be the preferences of some human H. If that human is bigoted or selfish, their U_H will be bigoted or selfish. In contrast, moral preferences that can be considered factually wrong will be filtered out by this construction. Similarly, preferences based on erroneous factual beliefs ("trees can think, so...") will be removed or qualified ("if trees could think, then...").

Thus if H is **wrong**, the U_H will not reflect that wrongness; but if H is **evil**, then U_H will reflect that evilness.

Also, the procedure will not distinguish between moral preferences and other types of preferences, unless the human themselves does.

4.3 Individual utility versus common utility

This research agenda will not look into how to combine the U_H of different humans. One could simply weight the utilities according to some semi-plausible scale and add them together.

But we could do many other things as well. I've suggested removing anti-altruistic preferences before combining the U_H's into some global utility function U_H for all of humanity - or for all future and current sentient beings, or for all beings that could suffer, or for all physical entities.

There are strong game-theoretical reasons to remove anti-altruistic preferences. We might also add philosophical considerations (eg moral realism) or deontological rules (eg human rights, restrictions on copying themselves, extra weighting to certain types of preferences), either to the individual U_H or when combining them, or prioritise moral preferences over other types. We might want to preserve the capacity for moral growth, somehow (see Section 4.6).

That can all be done, but is not part of this research agenda, whose sole purpose is to synthesise the individual U_H's, which can then be used for other purposes.

4.4 Synthesising U_H rather than discovering it (moral anti-realism)

The utility U_H will be constructed, rather than deduced or discovered. Some moral theories (such as some versions of moral realism) posit that there is a (generally unique) U_H waiting to be discovered. But none of these theories give effective methods for doing so.

In the absence of such a definition of how to discover an ideal U_H, it would be highly dangerous to assume that finding U_H is a process of discovery. Thus the whole method is constructive from the very beginning (and based on a small number of arbitrary choices).

Some versions of moral realism could make use of U_H as a starting point of their own definition. Indeed, in practice, moral realism and moral anti-realism seem to be initially almost identical when meta-preferences are taken into account. Moral realists often have mental examples of what counts as "moral realism doesn't work", while moral anti-realists still want to simplify and organise moral intuitions. To a first approximation, these approaches can be very similar in practice.

4.5 Self-referential contradictory preferences

There remain problems with self-referential preferences - preferences that claim they should be given more (or less) weight than otherwise (eg "all simple meta-preferences should be penalised"). This was already observed in a previous post.

This includes formal Gödel-style problems, with preferences explicitly contradicting themselves, but those seem solvable - with one or another version of logical uncertainty.

More worrying, from the practical standpoint, is the human tendency to reject values imposed upon them, just because they are imposed upon them. This resembles a preference of the type "reject any U_H computed by any synthesis process". This preference is weakly existent in almost all of us, and a variety of our other preferences should prevent the AI from forcibly re-writing us to become U_H-desiring agents.

So it remains not at all clear what happens when the AI says "this is what you really prefer" and we almost inevitably answer "no!". This concept can be seen, in a sense, as a gratification: we're not objecting to the outcome of the synthesis, per se, but to the way that the outcome was imposed on us.

Of course, since the U_H is constructed rather than real, there is some latitude. It might be possible to involve the human in the construction process, in a way that increases their buy-in (thanks to Tim Genewein for the suggestion). Maybe the AI could construct the first U_H, and refine it with further interactions with the human. And maybe, in that situation, if we are confident that U_H is pretty safe, we'd want the AI to subtly manipulate the human's preferences towards it.

4.6 The question of identity and change

It's not certain that human concepts of identity can be fully captured by identity preferences and meta-preferences. In that case, it is important that human identity be figured out somehow, lest humanity itself vanish even as our preferences are satisfied. Nick Bostrom sketched how this might happen: in the mindless outsourcers scenario, human outsource more and more of their key cognitive features to automated algorithms, until nothing remains of "them" any more.

Somewhat related is the fact that many humans see change and personal or moral growth as a key part of their identity. Can such a desire be accommodated, despite a likely stabilisation of values, without just becoming a random walk across preference space?

Some aspects of growth and change can be accommodated. Humans can certainly become more skilled, more powerful, and more knowledgeable. Since humans don't distinguish well between terminal and instrumental goals, some forms of factual learning resemble moral learning ("if it turns out that anarchism results in the greatest flourishing of humanity, then I wish to be a anarchist; if not, then not"). If we take into account the preferences of all humans in some roughly equal way (see Section 4.3), then we can get "moral progress" without needing to change anyone's individual preferences. Finally, professional roles, contracts, and alliances allow for behavioural changes (and sometimes values changes), in ways that maximise the initial values. Sort of like "if I do PR work for the Anarchist party, I will spout anarchist values" and "I accept to make my values more anarchist, in exchange for the Anarchist party shifting their values more towards mine".

Beyond these examples, it gets trickier to preserve moral change. We might put a slider that makes our own values less instrumental or less selfish over time, but that feels like a cheat: we already know what we will be, we're just taking the long route to get there. Otherwise, we might allow our values to change within certain defined

areas. This would have to be carefully defined to prevent random change, but the main challenge is efficiency: changing values have an inevitable efficiency cost, so there needs to be strong positive pressure to preserve the changes - and not just preserve an unused "possibility for change", but actual, efficiency-losing, changes. This "possibility for change" can be seen as a gratification: a cost we are willing to pay in terms of perfect efficiency, in order to have a process (continued moral learning) that we prefer.

This should be worth investigating more; it feels like these considerations need to be built into the synthesis process for this to work, rather than the synthesis project making them work itself (thus this kind of preferences is one of the "Global meta-preferences about the outcome of the synthesis process").

4.7 Other Issues not addressed

These are other important issues that need to be solved to get a fully friendly AI, even if the research agenda works perfectly. They are, however, beyond the scope of this agenda; a partial list of these is:

1. Actually building the AI itself (left as an exercise to the reader).
2. Population ethics (though some sort of average of individual human population ethics might be doable with these methods).
3. Taking into account other factors than individual preferences.
4. Issues of ontology and ontology changes.
5. Mind crime (conscious suffering beings simulated within an AI system), though some of the work on identity preferences may help in identifying conscious minds.
6. Infinite ethics.
7. Definitions of counterfactuals or which decision theory to use.
8. Agent foundations, logical uncertainty, how to keep a utility stable.
9. Acausal trade.
10. Optimisation daemons/inner optimisers/emergent optimisation.

Note that the Machine Intelligence Research Institute is working heavily on issues 7, 8, and 9.

---

1. A partial preference being a preference where the human considers only a small part of the variables describing the universe; see Section 1.1. ↩

2. Actually, *this* specific problem is not included directly in the research agenda, though see Section 4.3. ↩

3. Likely but not certain: we don't know how effective AIs might become at computing counterfactuals or modelling humans. ↩

4. It makes sense to allow partial preferences to contrast a small number of situations, rather than just two. So "when it comes to watching superhero movies, I'd prefer to watch them with Alan, but Beth will do, and definitely not with Carol". Since partial preferences with n situations can be built out of smaller number of partial preferences with two situations, allowing more situations is a useful practical move, but doesn't change the theory. ↩

5. "One-step" refers to hypotheticals that can be removed from the human's immediate experience ("Imagine that you and your family are in space...") but not very far removed (so no need for lengthy descriptions that could sway the human's opinions by hearing them). ↩

6. Equivalently to reducing the weight, we could increase uncertainty about the partial preference, given the unfamiliarity. There are many options for formalisms that lead to the same outcome. Though note that here, we are imposing a penalty (low weight/high uncertainty) for unfamiliarity, whereas the actual human might have incredibly strong internal certainty in their preferences. It's important to distinguish assumptions that the synthesis process makes, from assumptions that the human might make. ↩

7. Extreme situations are also situations where we have to be very careful to ensure the AI has the right model of all preference possibilities. The flaws of incorrect model can be corrected by enough data, but when data is sparse and unreliable, then model assumptions - including prior - tend to dominate the result. ↩

8. "Natural" does not, of course, mean any of "healthy", "traditional", or "non-polluting". However those using the term "natural" are often assuming all of those. ↩

9. The human's meta-preferences are also relevant to this it. It might be that, whenever asked about this particular contradiction, the human would answer one way. Therefore H's conditional meta-preferences may contain ways of resolving these contradictions, at least if the meta-preferences have high weight and the preferences have low weight.

   Conditional meta-preferences can be tricky, though, as we don't want them to allow the synthesis to get around the one-step hypotheticals restriction. A "if a long theory sounds convincing to me, I want to believe it" meta-preference in practice do away with these restrictions. That particular meta-preference might be cancelled out by the ability of many different theories to sound convincing. ↩

10. We *can* allow meta-preferences to determine a lot more of their own synthesis if we find an appropriate method that a) always reaches a synthesis, and b) doesn't artificially boost some preferences through a feedback effect. ↩

What is the evidence for productivity benefits of weightlifting?

[Mod Note: This question received an answer that seemed worth curating. See the answer by LW user hereisonehand for the curation notice]

I've been weightlifting for a while, and I've heard vaguely good things about it's effect on productivity, like a general increase in energy levels. A recent quick google search session came up empty. If someone looks into the literature and finds something interesting I'll pay a $50 prize.*

Assume the time horizon is <5 years. I'd prefer answers focus predominantly on productivity benefits. Effects on cardiovascular could be part of an analysis, but would not qualify on their own. If the evidence is for something clearly linked to productivity, like sleep, I'd count that. Introspective evidence will also not qualify. Comparisons to other forms of exercise would be especially interesting. Assume a healthy individual, although I'm at least somewhat interested in effects on individuals with depression or anxiety given their prevalence.

*Prize to go to best answer, as judged by me, if there are any that meet some minimal threshold of rigor, also as judged by me.

"The Bitter Lesson", an article about compute vs human knowledge in AI

This is a linkpost for http://www.incompleteideas.net/IncIdeas/BitterLesson.html

The Bitter Lesson Rich Sutton

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

In computer chess, the methods that defeated the world champion, Kasparov, in 1997, were based on massive, deep search. At the time, this was looked upon with dismay by the majority of computer-chess researchers who had pursued methods that leveraged human understanding of the special structure of chess. When a simpler, search-based approach with special hardware and software proved vastly more effective, these human-knowledge-based chess researchers were not good losers. They said that brute force" search may have won this time, but it was not a general strategy, and anyway it was not how people played chess. These researchers wanted methods based on human input to win and were disappointed when they did not.

A similar pattern of research progress was seen in computer Go, only delayed by a further 20 years. Enormous initial efforts went into avoiding search by taking advantage of human knowledge, or of the special features of the game, but all those efforts proved irrelevant, or worse, once search was applied effectively at scale. Also important was the use of learning by self play to learn a value function (as it was in many other games and even in chess, although learning did not play a big role in the 1997 program that first beat a world champion). Learning by self play, and learning in general, is like search in that it enables massive computation to be brought to bear. Search and learning are the two most important classes of techniques for utilizing massive amounts of computation in AI research. In computer Go, as in computer chess, researchers' initial effort was directed towards utilizing human understanding (so that less search was needed) and only much later was much greater success had by embracing search and learning.

In speech recognition, there was an early competition, sponsored by DARPA, in the 1970s. Entrants included a host of special methods that took advantage of human

knowledge---knowledge of words, of phonemes, of the human vocal tract, etc. On the other side were newer methods that were more statistical in nature and did much more computation, based on hidden Markov models (HMMs). Again, the statistical methods won out over the human-knowledge-based methods. This led to a major change in all of natural language processing, gradually over decades, where statistics and computation came to dominate the field. The recent rise of deep learning in speech recognition is the most recent step in this consistent direction. Deep learning methods rely even less on human knowledge, and use even more computation, together with learning on huge training sets, to produce dramatically better speech recognition systems. As in the games, researchers always tried to make systems that worked the way the researchers thought their own minds worked---they tried to put that knowledge in their systems---but it proved ultimately counterproductive, and a colossal waste of researcher's time, when, through Moore's law, massive computation became available and a means was found to put it to good use.

In computer vision, there has been a similar pattern. Early methods conceived of vision as searching for edges, or generalized cylinders, or in terms of SIFT features. But today all this is discarded. Modern deep-learning neural networks use only the notions of convolution and certain kinds of invariances, and perform much better.

This is a big lesson. As a field, we still have not thoroughly learned it, as we are continuing to make the same kind of mistakes. To see this, and to effectively resist it, we have to understand the appeal of these mistakes. We have to learn the bitter lesson that building in how we think we think does not work in the long run. The bitter lesson is based on the historical observations that 1) AI researchers have often tried to build knowledge into their agents, 2) this always helps in the short term, and is personally satisfying to the researcher, but 3) in the long run it plateaus and even inhibits further progress, and 4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning. The eventual success is tinged with bitterness, and often incompletely digested, because it is success over a favored, human-centric approach.

One thing that should be learned from the bitter lesson is the great power of general purpose methods, of methods that continue to scale with increased computation even as the available computation becomes very great. The two methods that seem to scale arbitrarily in this way are search and learning.

The second general point to be learned from the bitter lesson is that the actual contents of minds are tremendously, irredeemably complex; we should stop trying to find simple ways to think about the contents of minds, such as simple ways to think about space, objects, multiple agents, or symmetries. All these are part of the arbitrary, intrinsically-complex, outside world. They are not what should be built in, as their complexity is endless; instead we should build in only the meta-methods that can find and capture this arbitrary complexity. Essential to these methods is that they can find good approximations, but the search for them should be by our methods, not by us. We want AI agents that can discover like we can, not which contain what we have discovered. Building in our discoveries only makes it harder to see how the discovering process can be done.

Asymmetric Weapons Aren't Always on Your Side

Some time ago, Scott Alexander wrote about asymmetric weapons, and now he writes again about them. During these posts, Scott repeatedly characterizes asymmetric weapons as inherently stronger for the "good guys" than they are for the "bad guys". Here is a quote from his first post:

> Logical debate has one advantage over narrative, rhetoric, and violence: it's an *asymmetric weapon*. That is, it's a weapon which is stronger in the hands of the good guys than in the hands of the bad guys.

And here is a quote from his more recent one:

> A symmetric weapon is one that works just as well for the bad guys as for the good guys. For example, violence – your morality doesn't determine how hard you can punch; they can buy guns from the same places we can.
>
> An asymmetric weapon is one that works better for the good guys than the bad guys. The example I gave was Reason. If everyone tries to solve their problems through figuring out what the right thing to do is, the good guys (who are right) will have an easier time proving themselves to be right than the bad guys (who are wrong). Finding and using asymmetric weapons is the only non-coincidence way to make sustained moral progress.

One problem with this concept is that *just because something is asymmetric doesn't mean that it's asymmetric in a good direction.*

Scott talks about weapons that are asymmetric towards those who are right. However, there are many more types of asymmetries than just right vs. wrong - physical violence is asymmetric towards the strong, shouting people down is asymmetric towards the loud, and airing TV commercials is asymmetric towards people with more money. Violence isn't merely symmetric - it's *asymmetric in a bad direction*, since fascists are better than violence than you.

This in turn means that various sides will all be trying to pull things in directions that are asymmetric to their advantage. Indeed, a basic principle in strategy is to try to shift conflicts into areas where you are strong and your opponent is weak.

For instance, people who are good at violence benefit from things getting violent. People who are locally popular benefit from popularity contests. People who have lots of free time benefit from time-consuming processes. People who are better at keeping their composure benefit from discourse norms that punish displays of emotion.

Developing asymmetric processes that point towards truth is a good idea, and I'm all for it. But in practice there are also asymmetric processes that point towards error, or merely asymmetric processes that point towards what's currently popular or faddish. Those processes are, if anything, just as likely to have people trying to promote them than the pro-truth ones - perhaps more likely!

That doesn't make the people promoting those ideas "anti-truth" or whatever - they may not even be aware of what they're doing - but even so, people tend to respond to

incentives, and those incentives may well pull them towards norms and methods that are asymmetric in their favor independent of whether those norms and methods promote truth.

"But It Doesn't Matter"

If you ever find yourself saying, "Even if Hypothesis *H* is true, it doesn't have any decision-relevant implications," *you are rationalizing!* The fact that *H* is interesting enough for you to be considering the question at all (it's not some arbitrary trivium like the 1923th binary digit of π, or the low temperature in São Paulo on September 17, 1978) means that it must have some relevance to the things you care about. It is *vanishingly improbable* that your optimal decisions are going to be the *same* in worlds where *H* is true and worlds where *H* is false. The fact that you're tempted to *say* they're the same is probably because some part of you is afraid of some of the imagined consequences of *H* being true. But *H* is already true or already false! If you happen to live in a world where *H* is true, and you make decisions as if you lived in a world where *H* is false, you are thereby missing out on all the extra utility you would get if you made the *H*-optimal decisions instead! If you can figure out exactly what you're afraid of, maybe that will help you work out what the *H*-optimal decisions are. Then you'll be a better position to successfully notice which world you *actually* live in.

[AN #58] Mesa optimization: what it is, and why we should care

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

Find all Alignment Newsletter resources here. In particular, you can sign up, or look through this spreadsheet of all summaries that have ever been in the newsletter. I'm always happy to hear feedback; you can send it to me by replying to this email.

Highlights

Risks from Learned Optimization in Advanced Machine Learning Systems *(Evan Hubinger et al)*: Suppose you search over a space of programs, looking for one that plays TicTacToe well. Initially, you might find some good heuristics, e.g. go for the center square, if you have two along a row then place the third one, etc. But eventually you might find the minimax algorithm, which plays optimally by searching for the best action to take. Notably, your outer optimization over the space of programs found a program that was *itself* an optimizer that searches over possible moves. In the language of this paper, the minimax algorithm is a **mesa optimizer**: an optimizer that is found autonomously by a **base optimizer**, in this case the search over programs.

Why is this relevant to AI? Well, gradient descent is an optimization algorithm that searches over the space of neural net parameters to find a set that performs well on some objective. It seems plausible that the same thing could occur: gradient descent could find a model that is itself performing optimization. That model would then be a mesa optimizer, and the objective that it optimizes is the **mesa objective**. Note that while the mesa objective should lead to similar behavior as the base objective on the training distribution, it need not do so off distribution. This means the mesa objective is **pseudo aligned**; if it also leads to similar behavior off distribution it is **robustly aligned**.

A central worry with AI alignment is that if powerful AI agents optimize the wrong objective, it could lead to catastrophic outcomes for humanity. With the possibility of mesa optimizers, this worry is doubled: we need to ensure both that the base objective is aligned with humans (called **outer alignment**) and that the mesa objective is aligned with the base objective (called **inner alignment**). A particularly worrying aspect is **deceptive alignment**: the mesa optimizer has a long-term mesa objective, but knows that it is being optimized for a base objective. So, it optimizes the base objective during training to avoid being modified, but at deployment when the threat of modification is gone, it pursues only the mesa objective.

As a motivating example, if someone wanted to create the best biological replicators, they could have reasonably used natural selection / evolution as an optimization algorithm for this goal. However, this then would lead to the creation of humans, who would be mesa optimizers that optimize for other goals, and don't optimize for replication (e.g. by using birth control).

The paper has a lot more detail and analysis of what factors make mesa-optimization more likely, more dangerous, etc. You'll have to read the paper for all of these details. One general pattern is that, when using machine learning for some task X, there are a bunch of properties that affect the likelihood of learning heuristics or proxies rather than actually learning the optimal algorithm for X. For any such property, making heuristics/proxies more likely would result in a lower chance of mesa-optimization (since optimizers are less like heuristics/proxies), but conditional on mesa-optimization arising, makes it more likely that it is pseudo aligned instead of robustly aligned (because now the pressure for heuristics/proxies leads to learning a proxy mesa-objective instead of the true base objective).

Rohin's opinion: I'm glad this paper has finally come out. The concepts of mesa optimization and the inner alignment problem seem quite important, and currently I am most worried about x-risk caused by a misaligned mesa optimizer. Unfortunately, it is not yet clear whether mesa optimizers will actually arise in practice, though I think conditional on us developing AGI it is quite likely. Gradient descent is a relatively weak optimizer; it seems like AGI would have to be much more powerful, and so would require a learned optimizer (in the same way that humans can be thought of as "optimizers learned by evolution").

There still is a lot of confusion and uncertainty around the concept, especially because we don't have a good definition of "optimization". It also doesn't help that it's hard to get an example of this in an existing ML system -- today's systems are likely not powerful enough to have a mesa optimizer (though even if they had a mesa optimizer, we might not be able to tell because of how uninterpretable the models are).

Read more: Alignment Forum version

Technical AI alignment

Agent foundations

Selection vs Control *(Abram Demski)*: The previous paper focuses on mesa optimizers that are explicitly searching across a space of possibilities for an option that performs well on some objective. This post argues that in addition to this "selection" model of optimization, there is a "control" model of optimization, where the model cannot evaluate all of the options separately (as in e.g. a heat-seeking missile, which can't try all of the possible paths to the target separately). However, these are not cleanly separated categories -- for example, a search process could have control-based optimization inside of it, in the form of heuristics that guide the search towards more likely regions of the search space.

Rohin's opinion: This is an important distinction, and I'm of the opinion that most of what we call "intelligence" is actually more like the "control" side of these two options.

Learning human intent

Imitation Learning as f-Divergence Minimization *(Liyiming Ke et al)* (summarized by Cody): This paper frames imitation learning through the lens of matching your model's distribution over trajectories (or conditional actions) to the distribution of an expert policy. This framing of distribution comparison naturally leads to the discussion of f-divergences, a broad set of measures including KL and Jenson-Shannon Divergences.

The paper argues that existing imitation learning methods have implicitly chosen divergence measures that incentivize "mode covering" (making sure to have support anywhere the expert does) vs mode collapsing (making sure to only have support where the expert does), and that the latter is more appropriate for safety reasons, since the average between two modes of an expert policy may not itself be a safe policy. They demonstrate this by using a variational approximation of the reverse-KL distance as the divergence underlying their imitation learner.

Cody's opinion: I appreciate papers like these that connect peoples intuitions between different areas (like imitation learning and distributional difference measures). It does seem like this would even more strongly lead to lack of ability to outperform the demonstrator, but that's honestly more a critique of imitation learning more generally than this paper in particular.

Handling groups of agents

Social Influence as Intrinsic Motivation for Multi-Agent Deep RL *(Natasha Jaques et al)* (summarized by Cody): An emerging field of common-sum multi-agent research asks how to induce groups of agents to perform complex coordination behavior to increase general reward, and many existing approaches involve centralized training or hardcoding altruistic behavior into the agents. This paper suggests a new technique that rewards agents for having a causal influence over the actions of other agents, in the sense that the actions of the pair of agents agents have high mutual information. The authors empirically find that having even a small number of agents who act as "influencers" can help avoid coordination failures in partial information settings and lead to higher collective reward. In one sub-experiment, they only add this influence reward to the agents' communication channels, so agents are incentivized to provide information that will impact other agents' actions (this information is presumed to be truthful and beneficial since otherwise it would subsequently be ignored).

Cody's opinion: I'm interested by this paper's finding that you can generate apparently altruistic behavior by incentivizing agents to influence others, rather than necessarily help others. I also appreciate the point that was made to train in a decentralized way. I'd love to see more work on a less asymmetric version of influence reward; currently influencers and influencees are separate groups due to worries about causal feedback loops, and this implicitly means there's a constructed group of quasi-altruistic agents who are getting less concrete reward because they're being incentivized by this auxiliary reward.

Uncertainty

ICML Uncertainty and Robustness Workshop Accepted Papers (summarized by Dan H): The Uncertainty and Robustness Workshop accepted papers are available. Topics include out-of-distribution detection, generalization to stochastic corruptions, label corruption robustness, and so on.

Miscellaneous (Alignment)

To first order, moral realism and moral anti-realism are the same thing *(Stuart Armstrong)*

AI strategy and policy

Grover: A State-of-the-Art Defense against Neural Fake News *(Rowan Zellers et al)*: Could we use ML to detect fake news generated by other ML models? This paper suggests that models that are used to generate fake news will also be able to be used to *detect* that same fake news. In particular, they train a GAN-like language model on news articles, that they dub GROVER, and show that the generated articles are *better* propaganda than those generated by humans, but they can at least be detected by GROVER itself.

Notably, they do plan to release their models, so that other researchers can also work on the problem of detecting fake news. They are following a similar release strategy as with GPT-2 (AN #46): they are making the 117M and 345M parameter models public, and releasing their 1.5B parameter model to researchers who sign a release form.

Rohin's opinion: It's interesting to see that this group went with a very similar release strategy, and I wish they had written more about why they chose to do what they did. I do like that they are on the face of it "cooperating" with OpenAI, but eventually we need norms for *how* to make publication decisions, rather than always following the precedent set by someone prior. Though I suppose there could be a bit more risk with their models -- while they are the same size as the released GPT-2 models, they are better tuned for generating propaganda than GPT-2 is.

Read more: Defending Against Neural Fake News

The Hacker Learns to Trust *(Connor Leahy)*: An independent researcher attempted to replicate GPT-2 (AN #46) and was planning to release the model. However, he has now decided not to release, because releasing would set a bad precedent. Regardless of whether or not GPT-2 is dangerous, at some point in the future, we will develop AI systems that really are dangerous, and we need to have adequate norms then that allow researchers to take their time and evaluate the potential issues and then make an informed decision about what to do. **Key quote:** "sending a message that it is ok, even celebrated, for a lone individual to unilaterally go against reasonable safety concerns of other researchers is not a good message to send".

Rohin's opinion: I quite strongly agree that the most important impact of the GPT-2 decision was that it has started a discussion about what appropriate safety norms should be, whereas before there were no such norms at all. I don't know whether or not GPT-2 is dangerous, but I am glad that AI researchers have started thinking about whether and how publication norms should change.

Other progress in AI

Reinforcement learning

A Survey of Reinforcement Learning Informed by Natural Language *(Jelena Luketina et al)* (summarized by Cody): Humans use language as a way of efficiently storing knowledge of the world and instructions for handling new scenarios; this paper is written from the perspective that it would be potentially hugely valuable if RL agents could leverage information stored in language in similar ways. They look at both the

case where language is an inherent part of the task (example: the goal is parameterized by a language instruction) and where language is used to give auxiliary information (example: parts of the environment are described using language). Overall, the authors push for more work in this area, and, in particular, more work using external-corpus-pretrained language models and with research designs that use human-generated rather than synthetically-generated language; the latter is typically preferred for the sake of speed, but the former has particular challenges we'll need to tackle to actually use existing sources of human language data.

Cody's opinion: This article is a solid and useful version of what I would expect out of a review article: mostly useful as a way to get thinking in the direction of the intersection of RL and language, and makes me more interested in digging more into some of the mentioned techniques, since by design this review didn't go very deep into any of them.

Deep learning

the transformer … “explained”? *(nostalgebraist)* (H/T Daniel Filan): This is an excellent explanation of the intuitions and ideas behind self-attention and the Transformer architecture (AN #44).

Ray Interference: a Source of Plateaus in Deep Reinforcement Learning *(Tom Schaul et al)* (summarized by Cody): The authors argue that Deep RL is subject to a particular kind of training pathology called "ray interference", caused by situations where (1) there are multiple sub-tasks within a task, and the gradient update of one can decrease performance on the others, and (2) the ability to learn on a given sub-task is a function of its current performance. Performance interference can happen whenever there are shared components between notional subcomponents or subtasks, and the fact that many RL algorithms learn on-policy means that low performance might lead to little data collection in a region of parameter space, and make it harder to increase performance there in future.

Cody's opinion: This seems like a useful mental concept, but it seems quite difficult to effectively remedy, except through preferring off-policy methods to on-policy ones, since there isn't really a way to decompose real RL tasks into separable components the way they do in their toy example

Meta learning

Alpha MAML: Adaptive Model-Agnostic Meta-Learning *(Harkirat Singh Behl et al)*

Instead of "I'm anxious," try "I feel threatened"

This is a linkpost for https://mhollyelmoreblog.wordpress.com/2019/06/19/instead-of-im-anxious-try-i-feel-threatened/

cw: teaching to learn

I have a long history with anxiety, and I'm pretty good at noticing when it's happening. The problem is that I'm always anxious. Noticing anxiety doesn't snap me out of anxiety– in fact, it often produces meta-anxiety, anxiety about feeling anxious. So I've tried a simple reframe lately, and I'm liking the results. Instead of noting "I'm anxious," I say to myself "I feel threatened" or "I feel threatened by x" if I know what set me off.

Anxiety is just chronically being in a state of fight or flight, and fight or flight has a stimulus. I like Sapolsky's thesis, which is roughly that for most animals, the stimulus is always something external, a threat to safety or status. For anxious humans, the threatening stimuli are internalized, and fight or flight is either triggered or sustained by thoughts. Anxiety is the condition of feeling threatened.

And yet, noticing that I feel threatened is much more specific than noticing that I'm anxious, whether I can identify the threat or not. It makes what I'm feeling less about me (*I'm* just anxious; my perception is inaccurate; oh, why don't I just stop???) and more about the pattern of behavior (I'm reacting this way because I perceive that thing to be a threat; is it really a threat?; if it is, is it something I can handle?).

In the short time I've been practicing this, I've identified many things I had not realized I considered threats, although, of course, on the feeling level I had always known. I'm surprised by how mundane most of the threats are. Many of them are just "I feel threatened because that noise startled me." But others are kind of embarrassing or incongruent with my self-concept. For example, I'm threatened by other people being better than me. I would find myself stiff and clearly in fight or flight when singing in a group, for instance, and I used to just nurse that anxiety for the entire practice thinking, "Fuck, I'm anxious, I can't breathe, my singing is therefore terrible, and I must be blushing…" But with this technique, I notice the anxious symptoms and see if I can identify the "threat" that tripped them. To my shock, it was usually as simple as another person singing really well, or me not knowing how to sight read when others could. Such everyday, simple provocations! At this point, I don't have much pride left to be embarrassed with, but it's still humbling to see my mountains of anxiety for the molehills of petty jealousy and insecurity they could have stayed.

I don't blame myself for getting carried away. Anxiety is the master of false narratives. An injection of anxiety causes my thoughts to speed up and start going down rabbitholes of what to do, all premised on unseen assumptions I'm making about the nature and severity of the threat. There's no time or brainpower to examine every hasty conclusion when you're swept up in that wave. Reining in anxiety is necessarily a process. It can be embarrassing to realize just how simple the "threat" that led to hours (or days, or months, or years…) of anxiety was, but it's also such a relief! Admitting I'm jealous or petty or flawed is a small price to pay to reclaim some peace.

Machine Learning Projects on IDA

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

TLDR

We wrote a 20-page document that explains IDA and outlines potential Machine Learning projects about IDA. This post gives an overview of the document.

What is IDA?

Iterated Distillation and Amplification (IDA) is a method for training ML systems to solve challenging tasks. It was introduced by Paul Christiano. IDA is intended for tasks where:

- The goal is to outperform humans at the task or to solve instances that are too hard for humans.
- It is not feasible to provide demonstrations or reward signals sufficient for super-human performance at the task
- Humans have a high-level understanding of how to approach the task and can reliably solve easy instances.

The idea behind IDA is to bootstrap using an approach similar to AlphaZero, but with a learned model of steps of human reasoning instead of the fixed game simulator.

Our document provides a self-contained technical description of IDA. For broader discussion of IDA and its relevance to value alignment, see Ought's presentation, Christiano's blogpost, and the Debate paper. There is also a technical ML paper applying IDA to algorithmic problems (e.g. shortest path in a graph).

ML Projects on IDA

Our document outlines three Machine Learning projects on IDA. Our goal in outlining these projects is to generate discussion and encourage research on IDA. We are not (as of June 2019) working on these projects, but we are interested in collaboration. The project descriptions are “high-level” and leave many choices undetermined. If you took on a project, part of the work would be refining the project and fixing a concrete objective, dataset and model.

Project 1: Amplifying Mathematical Reasoning

This project is about applying IDA to problems in mathematics. This would involve learning to solve math problems by breaking them down into easier sub-problems. The problems could be represented in a formal language (as in this paper) or in natural language. We discuss a recent dataset of high-school problems in natural language, which was introduced in this paper. Here are some examples from the dataset:

> Question: Let u(n) = -n^3 - n^2. Let e(c) = -2*c^3 + c. Let f(j) = -118*e(j) + 54*u(j). What is the derivative of f(a)?
>
> Answer: 546*a^2 - 108*a - 118
>
> Question: Three letters picked without replacement from qqqkkklkqkkk. Give probability of sequence qql.
>
> Answer: 1/110

The paper showed impressive results on the dataset for a Transformer model trained by supervised learning (sequence-to-sequence). This suggests that a similar model could do well at learning to solve these problems by decomposition.

Project 2: IDA for Neural Program Interpretation

There's a research program in Machine Learning on "Neural Program Interpretation" (NPI). Work on NPI focuses on learning to reproduce the behavior of computer programs. One possible approach is to train end-to-end on input-output behavior. However in NPI, a model is trained to mimic the program's *internal* behavior, including all the low-level operations and the high-level procedures which invoke them.

NPI has some similar motivations to IDA. This project applies IDA to the kinds of tasks explored in NPI and compares IDA to existing approaches. Tasks could include standard algorithms (e.g. sorting), algorithms that operate with databases, and algorithms that operate on human-readable inputs (e.g. text, images).

Project 3: Adaptive Computation

The idea of "adaptive computation" is to vary the amount of computation you perform for different inputs. You want to apply more computation to inputs that are hard but solvable.

Adaptive computation seems important for the kinds of problems IDA is intended to solve, including some of the problems in Projects 1 and 2. This project would investigate different approaches to adaptive computation for IDA. The basic idea is to decide whether to rely only on the distilled model (which is fast but approximate) or to additionally use amplification (which is more accurate but slower). This decision could be based on a calibrated model or based on a learned policy for choosing whether to use amplification.

Only optimize to 95 %

I was reading Tom Chivers book "The AI does not hate you" and in a discussion about avoiding bad side effects when asking a magic broomstick to fill a water bucket, it was suggested that somehow instead of asking the broomstick to fill the bucket you could do something like ask it to become 95 percent sure that it was full, and that might make it less likely to flood the house.

Apparently Tom asked Eliezer at the time and he said there was no known problem with that solution.

Are there any posts on this? Is the reason why we don't know this won't work just because it's hard to make this precise?

For the past, in some ways only, we are moral degenerates

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

Have human values improved over the last few centuries? Or is it just that current human values are naturally closer to our (current) human values and so we think that there's been moral progress towards us?

If we project out in the future, the first scenario posits continuing increased moral improvements (as the "improvement trend" continues) and the second posits moral degeneration (as the values drift away from our own). So what is it?

I'll make the case that both trends are happening. We have a lot less slavery, racism, ethnic conflicts, and endorsements of slavery, racism, and ethnic conflicts. In an uneven way, poorer people have more effective rights than they did before, so it's somewhat less easy to abuse them.

Notice something interesting about the previous examples? They can all be summarised as "some people who were treated badly are now treated better". Many people throughout time would agree that these people are actually being treated better. On the issue of slavery, consider the following question:

- "If X would benefit from being a non-slave more than being a slave, and there were no costs to society, would it be better for X not to be a slave?"

Almost everyone would agree to that throughout history, barring a few examples of *extremely* motivated reasoning. So most defences of slavery rest on the idea that some classes of people are better off as slaves (almost always a factual error, and generally motivated reasoning), or that some morally relevant group of people benefited from slavery enough to make it worthwhile.

So most clear examples of moral progress are giving benefits to people, such that anyone who knew all the facts would agree it was beneficial for those people.

That trend we might expect to continue; as we gain greater knowledge how to benefit people, and as we gain greater resources, we can expect more people to be benefited.

Values that we have degenerated on

But I'll argue that there are a second class of values that have less of a "direction" to, and where we could plausibly be argued to have "degenerated". And, hence, where we might expect our descendants to "degenerate" more (ie move further away from us).

Community and extended family values, for example, are areas where much of the past would be horrified by the present. Why are people not (generally) meeting up with their second cousins every two weeks, and why do people waste time gossiping about irrelevant celebrities rather than friends and neighbours?

On issues of honour and reputation, why have we so meekly accepted to become citizens of administrative bureaucracies and defer to laws and courts, rather than taking pride in meeting out our own justice and defending our own honour? "Yes, yes", the hypothetical past person would say, "your current system is fairer and more efficient; but why did it have to turn you all so supine"? Are you not free men?

Play around with vaguely opposite virtues: spontaneity versus responsibility; rationality versus romanticism; pride versus humility; honesty versus tact, and so on. Where is the ideal mean between any of those two extremes? Different people and different cultures put the ideal mean in different places, and there's no reason to suspect that the means are "getting better" rather than just "moving around randomly".

I won't belabour the point; it just seems to me that there are areas where the moral progress narrative makes more sense (giving clear benefits to people who didn't have them) and areas where the "values drift around" narrative makes more sense. And hence we might hope for continuing moral progress in some areas, and degeneration (or at least stagnation) in others.

Map of (old) MIRI's Research Agendas

In late 2016 I independently curated a visual map of the areas of research MIRI was working on at the time, together with descriptions of each area and arrows explaining the relationships between areas.

While MIRI's research has evolved since, I have foud myself going back to this map as a reference while doing research, so I though I would share it with the rest of you in case it might be useful for some.

Map of Research Areas in AI Alignment

The map is based on the Agent Foundations agenda by Soares and Fallenstein of 2014 and the Alignment for Advanced ML Systems agenda by Taylor et al in 2016.

Let me know if this was useful for you! It will be useful evidence on whether spending time curating more maps of the same style is a good use of my time.

Is your uncertainty resolvable?

I was chatting with Andrew Critch about the idea of Reacts on LessWrong.

Specifically, the part where I thought there are particular epistemic states that *don't have words yet*, but should. And that a function of LessWrong might be to make various possible epistemic states more salient as options. You might have reacts for "approve/disapprove" and "agree/disagree"... but you might also want reactions that let you quickly and effortless express "this isn't exactly false or bad but it's subtly making this discussion worse."

Fictionalized, Paraphrased Critch said "hmm, this reminds me of some particular epistemic states I recently noticed that don't have names."

"Go on", said I.

"So, you know the feeling of being uncertain? And how it feels different to be 60% sure of something, vs 90%?"

"Sure."

"Okay. So here's two other states you might be in:

- **75% sure that you'll eventually be 99% sure,**
- **80% sure that you'll eventually be 90% sure.**

He let me process those numbers for a moment.

...

Then he continued: "Okay, now imagine you're thinking about a particular AI system you're designing, which might or might not be alignable.

"If you're feeling 75% sure that you'll eventually be 99% sure that that AI is safe, this means you think that *eventually* you'll have a clear understanding of the AI, such that you feel confident turning it on without destroying humanity. Moreover you expect to be able to convince *other people* that it's safe to turn it on without destroying humanity.

"Whereas if you're 80% sure that eventually you'll be 90% sure that it'll be safe, *even in the future state* where you're better informed and more optimistic, you might still not actually be confident enough to turn it on. And even if for some reason you are, other people might disagree about whether you should turn it on.

"I've noticed people tracking how certain they are of something, without paying attention to whether their uncertainty is *possible to resolve*. And this has important ramifications for what kind of plans they can make. Some plans *require near-certainty.* Especially many plans that require group coordination.

"Makes sense", said I. "Can I write this up as a blogpost?"

I’m not quite sure about the best name here, but this seems like a useful concept to have a handle for. Something like “unresolvable uncertainty?”

Does the _timing_ of practice, relative to sleep, make a difference for skill consolidation?

It is well known that sleep (both mid-day naps and nighttime sleep) has a large effect on the efficacy of motor skill acquisition. Performance on a newly learned task improves, often markedly, following a period of sleep.

A few citations (you can find many more by searching "motor skill acquisition sleep" or similar in google scholar) :

- https://www.nature.com/articles/nn1959
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0000341
- https://link.springer.com/article/10.1111/j.1479-8425.2012.00576.x

I want to know if the _timing_ of practice, relative to sleep, makes a difference for skill acquisition.

For instance, if you practice a skill at 7:00 PM, shortly before a night of sleep, will your performance be better in the morning than if you had practiced at 7:00 AM had a full day of wakefulness, and _then_ gone to sleep? If so, what is the effect size?

Josh Kaufman makes a claim to this effect in his book, *The First 20 Hours*. I have no particular reason to doubt him, 40 minutes of searching on google scholar did not turn up any papers about the importance of sleep and practice timing.

Can you point me at a relevant citation?