Best of LessWrong: September 2020

1. Draft report on AI timelines
2. Most Prisoner's Dilemmas are Stag Hunts; Most Stag Hunts are Schelling Problems
3. Book Review: Working With Contracts
4. Comparative advantage and when to blow up your island
5. My computational framework for the brain
6. Honoring Petrov Day on LessWrong, in 2020
7. AGI safety from first principles: Introduction
8. What's Wrong with Social Science and How to Fix It: Reflections After Reading 2578 Papers
9. Comparative Advantage is Not About Trade
10. The Wiki is Dead, Long Live the Wiki! [help wanted]
11. What happens if you drink acetone?
12. Clarifying “What failure looks like”
13. Numeracy neglect - A personal postmortem
14. Social Capital Paradoxes
15. Capturing Ideas
16. "Learning to Summarize with Human Feedback" - OpenAI
17. What Decision Theory is Implied By Predictive Processing?
18. Tips for the most immersive video calls
19. How To Fermi Model
20. Artificial Intelligence: A Modern Approach (4th edition) on the Alignment Problem
21. "Win First" vs "Chill First"
22. The new Editor
23. Stop pressing the Try Harder button
24. Some thoughts on criticism
25. AGI safety from first principles: Superintelligence
26. Why is Bayesianism important for rationality?
27. The Four Children of the Seder as the Simulacra Levels
28. Rationality for Kids?
29. Why GPT wants to mesa-optimize & how we might change this
30. AGI safety from first principles: Goals and Agency
31. Safety via selection for obedience
32. Forecasting Thread: Existential Risk
33. CTWTB: Paths of Computation State
34. Using GPT-N to Solve Interpretability of Neural Networks: A Research Agenda
35. Comparing Utilities
36. [AN #116]: How to make explanations of neurons compositional
37. Updates Thread
38. Reflections on AI Timelines Forecasting Thread
39. Petrov Event Roundup 2020
40. Needed: AI infohazard policy
41. Rationality and playfulness
42. What are good rationality exercises?
43. How Much Computational Power Does It Take to Match the Human Brain?
44. Let the AI teach you how to flirt
45. Not all communication is manipulation: Chaperones don't manipulate proteins
46. A Toy Model of Hingeyness
47. Anthropomorphisation vs value learning: type 1 vs type 2 errors
48. Human Biases that Obscure AI Progress
49. Gems from the Wiki: Do The Math, Then Burn The Math and Go With Your Gut
50. What Does "Signalling" Mean?

Best of LessWrong: September 2020

1. Draft report on AI timelines
2. Most Prisoner's Dilemmas are Stag Hunts; Most Stag Hunts are Schelling Problems
3. Book Review: Working With Contracts
4. Comparative advantage and when to blow up your island
5. My computational framework for the brain
6. Honoring Petrov Day on LessWrong, in 2020
7. AGI safety from first principles: Introduction
8. What's Wrong with Social Science and How to Fix It: Reflections After Reading 2578 Papers
9. Comparative Advantage is Not About Trade
10. The Wiki is Dead, Long Live the Wiki! [help wanted]
11. What happens if you drink acetone?
12. Clarifying “What failure looks like”
13. Numeracy neglect - A personal postmortem
14. Social Capital Paradoxes
15. Capturing Ideas
16. "Learning to Summarize with Human Feedback" - OpenAI
17. What Decision Theory is Implied By Predictive Processing?
18. Tips for the most immersive video calls
19. How To Fermi Model
20. Artificial Intelligence: A Modern Approach (4th edition) on the Alignment Problem
21. "Win First" vs "Chill First"
22. The new Editor
23. Stop pressing the Try Harder button
24. Some thoughts on criticism
25. AGI safety from first principles: Superintelligence
26. Why is Bayesianism important for rationality?
27. The Four Children of the Seder as the Simulacra Levels
28. Rationality for Kids?
29. Why GPT wants to mesa-optimize & how we might change this
30. AGI safety from first principles: Goals and Agency
31. Safety via selection for obedience
32. Forecasting Thread: Existential Risk
33. CTWTB: Paths of Computation State
34. Using GPT-N to Solve Interpretability of Neural Networks: A Research Agenda
35. Comparing Utilities
36. [AN #116]: How to make explanations of neurons compositional
37. Updates Thread
38. Reflections on AI Timelines Forecasting Thread
39. Petrov Event Roundup 2020
40. Needed: AI infohazard policy
41. Rationality and playfulness
42. What are good rationality exercises?
43. How Much Computational Power Does It Take to Match the Human Brain?
44. Let the AI teach you how to flirt
45. Not all communication is manipulation: Chaperones don't manipulate proteins
46. A Toy Model of Hingeyness
47. Anthropomorphisation vs value learning: type 1 vs type 2 errors

48. Human Biases that Obscure AI Progress
49. Gems from the Wiki: Do The Math, Then Burn The Math and Go With Your Gut
50. What Does "Signalling" Mean?

Draft report on AI timelines

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

Hi all, I've been working on some AI forecasting research and have prepared a draft report on timelines to transformative AI. I would love feedback from this community, so I've made the report viewable in a Google Drive folder here.

With that said, most of my focus so far has been on the high-level structure of the framework, so the particular quantitative estimates are very much in flux and many input parameters aren't pinned down well -- I wrote the bulk of this report before July and have received feedback since then that I haven't fully incorporated yet. I'd prefer if people didn't share it widely in a low-bandwidth way (e.g., just posting key graphics on Facebook or Twitter) since the conclusions don't reflect Open Phil's "institutional view" yet, and there may well be some errors in the report.

The report includes a quantitative model written in Python. Ought has worked with me to integrate their forecasting platform Elicit into the model so that you can see other people's forecasts for various parameters. If you have questions or feedback about the Elicit integration, feel free to reach out to elicit@ought.org.

Looking forward to hearing people's thoughts!

Most Prisoner's Dilemmas are Stag Hunts; Most Stag Hunts are Schelling Problems

I previously claimed that most apparent Prisoner's Dilemmas are actually Stag Hunts. I now claim that they're Schelling Pub in practice. I conclude with some lessons for fighting Moloch.

This post turned out especially dense with inferential leaps and unexplained terminology. If you're confused, try to ask in the comments and I'll try to clarify.

Some ideas here are due to Tsvi Benson-Tilsen.

The title of this post used to be ***Most Prisoner's Dilemmas are Stag Hunts; Most Stag Hunts are Battle of the Sexes***. I'm changing it based on this comment. "Battle of the Sexes" is a game where a male and female (let's say Bob and Alice) want to hang out, but each of them would prefer to engage in gender-stereotyped behavior. For example, Bob wants to go to a football game, and Alice wants to go to a museum. The gender issues are distracting, and although it's the standard, *the game isn't that well-known anyway,* so sticking to the standard didn't buy me much (in terms of reader understanding).

I therefore present to you,

the Schelling Pub Game:

Two friends would like to meet at the pub. In order to do so, they must make the same selection of pub (making this a Schelling-point game). However, they have different preferences about which pub to meet at. For example:

- Alice and Bob would both like to go to a pub this evening.
- There are two pubs: the Xavier, and the Yggdrasil.
- Alice likes the Xavier twice as much as the Yggdrasil.
- Bob likes the Yggdrasil twice as much as the Xavier.
- However, Alice and Bob also prefer to be with each other. Let's say they like being together ten times as much as they like being apart.

Schelling Pub Game payoff matrix

payoffs written alice;bob		B's choice X	B's choice Y
A's choice	X	20;10	2;2
	Y	1;1	10;20

The important features of this game are:

- The Nash equilibria are all Pareto-optimal. There is no "individually rational agents work against each other" problem, like in prisoner's dilemma or even stag hunt.
- There are multiple equilibria, and different agents prefer different equilibria.

Thus, realistically, agents may not end up in equilibrium at all -- because (in the single-shot game) they don't know which to choose, and because (in an iterated version of the game) they may make locally sub-optimal choices in order to influence the long-run behavior of other players.

(Edited to add, based on comments:)

Here's a summary of the central argument which, despite the lack of pictures, may be easier to understand.

1. Most Prisoner's Dilemmas are actually iterated.
2. Iterated games are a whole different game with a different action space (because you can react to history), a different payoff matrix (because you care about future payoffs, not just the present), and a different set of equilibria.
3. It is characteristic of PD that players are incentivised to play away from the Pareto frontier; IE, no Pareto-optimal point is an equilibrium. *This is not the case with iterated PD.*
4. It is characteristic of Stag Hunt that there is a Pareto-optimal equilibrium, but there is also another equilibrium which is far from optimal. *This is also the case with iterated PD.* So **iterated PD resembles Stag Hunt**.
5. However, it is furthermore true of iterated PD that *there are multiple different Pareto-optimal equilibria, which benefit different players more or less.* Also, if players don't successfully coordinate on one of these equilibria, they can end up in a worse overall state (such as mutual defection forever, due to playing grim-trigger strategies with mutually incompatible demands). **This makes iterated PD resemble the Schelling Pub Game.**

In fact, the Folk Theorem suggests that *most* iterated games will resemble the Schelling Pub Game in this way.

In a comment on The Schelling Choice is "Rabbit", not "Stag" I said:

> In the book *The Stag Hunt,* Skyrms similarly says that lots of people use Prisoner's Dilemma to talk about social coordination, and he thinks people should often use Stag Hunt instead.
>
> I think this is right. Most problems which initially seem like Prisoner's Dilemma are actually Stag Hunt, because there are potential enforcement mechanisms available. The problems discussed in Meditations on Moloch are mostly Stag Hunt problems, not Prisoner's Dilemma problems -- Scott even talks about enforcement, when he describes the dystopia where everyone has to kill anyone who doesn't enforce the terrible social norms (including the norm of enforcing).
>
> This might initially sound like good news. Defection in Prisoner's Dilemma is an inevitable conclusion under common decision-theoretic assumptions. Trying to escape multipolar traps with exotic decision theories might seem hopeless. On the other hand, rabbit in Stag Hunt is *not* an inevitable conclusion, by any means.
>
> Unfortunately, in reality, hunting stag is actually quite difficult. *("The schelling choice is Rabbit, not Stag... and that really sucks!")*

Inspired by Zvi's recent sequence on Moloch, I wanted to expand on this. These issues are important, since they determine how we think about group action problems / tragedy of the commons / multipolar traps / Moloch / all the other synonyms for the same thing.

My current claim is that most Prisoner's Dilemmas are actually *Schelling pub games*. But let's first review the relevance of Stag Hunt.

Your PD Is Probably a Stag Hunt

There are several reasons why an apparent Prisoner's Dilemma may be more of a Stag Hunt.

- The game is actually an iterated game.
- Reputation networks could punish defectors and reward cooperators.
- There are enforceable contracts.
- Players know quite a bit about how other players think (in the extreme case, players can view each other's source code).

Each of these formal model creates a situation where players ***can*** get into a cooperative equilibrium. The challenge is that you can't unilaterally decide everyone should be in the cooperative equilibrium. If you want good outcomes for yourself, you have to account for what everyone else probably does. If you think everyone is likely to be in a bad equilibrium where people punish each other for cooperating, then aligning with that equilibrium might be the best you can do! This is like hunting rabbit.

Exercize*: is there a situation in your life, or within spitting distance, which seems like a Prisoner's Dilemma to you, where everyone is stuck hurting each other due to bad incentives? Is it an iterated situation? Could there be reputation networks which weed out bad actors? Could contracts or contract-like mechanisms be used to encourage good behavior?*

So, why do we perceive so many situations to be Prisoner's Dilemma -like rather than Stag Hunt -like? Why does Moloch sound more like *each individual is incentivized to make it worse for everyone else* than *everyone is stuck in a bad equilibrium?*

Sarah Constantine writes:

> A friend of mine speculated that, in the decades that humanity has lived under the threat of nuclear war, we've developed the assumption that we're living in a world of one-shot Prisoner's Dilemmas rather than repeated games, and lost some of the social technology associated with repeated games. Game theorists do, of course, know about iterated games and there's some fascinating research in evolutionary game theory, but the original formalization of game theory was for the application of nuclear war, and the 101-level framing that most educated laymen hear is often that one-shot is the prototypical case and repeated games are hard to reason about without computer simulations.

To use board-game terminology, the *game* may be a Prisoner's Dilemma, but the *metagame* can use enforcement techniques. Accounting for enforcement techniques, the game is more like a Stag Hunt, where defecting is "rabbit" and cooperating is "stag".

Schelling Pubs

But this is a bit informal. You don't separately choose how to metagame and how to game; really, your iterated strategy determines what you do in individual games.

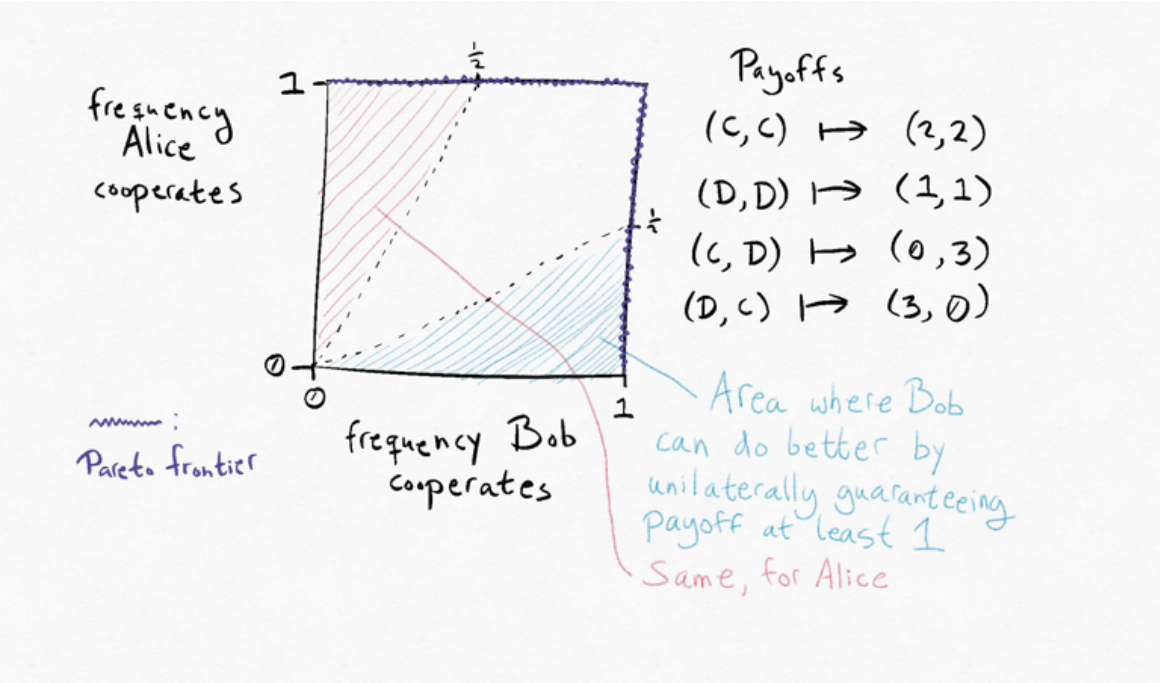
So it's more accurate to just think of the iterated game. There are a bunch of iterated strategies which you can choose from.

The key difference between the single-shot game and the iterated game is that cooperative strategies, such as Tit for Tat (but including others), are avaliable. These strategies have the property that (1) they are equilibria -- if you know the other player is playing Tit for Tat, there's no reason for you not to; (2) if both players use them, they end up cooperating.

A key feature of Tit for Tat strategy is that if you do end up playing against a pure defector, you do almost as well as you could possibly do with them. This doesn't sound very much like a Stag Hunt. It begins to sound like a Stag Hunt in which you can change your mind and go hunt rabbit if the other person doesn't show up to hunt stag with you.

Sounds great, right? We can just play one of these cooperative strategies.

The problem is, there are many possible self-enforcing equilibria. Each player can threaten the other player with a *Grim Trigger* strategy: they defect forever the moment some specified condition isn't met. This can be used to extort the other player for more than just the mutual-cooperation payoff. Here's an illustration of possible outcomes, with the enforceable frequencies in the white area:



The entire while area are *enforceable equilibria:* players could use a grim-trigger strategy to make each other cooperate with very close to the desired frequency, because what they're getting is still better than mutual defection, even if it is far from fair, or far from the Pareto frontier.

Alice could be extorting Bob by cooperating 2/3rds of the time, with a grim-trigger threat of never cooperating at all. Alice would then get an average payoff of 2⅓, while Bob would get an average payout of 1⅓.

In the artificial setting of Prisoner's Dilemma, it's easy to say that Cooperate, Cooperate is the "fair" solution, and an equilibrium like I just described is "Alice exploiting Bob". However, real games are not so symmetric, and so it will not be so obvious what "fair" is. The purple squiggle highlights the Pareto frontier -- the space of outcomes which are "efficient" in the sense that no alternative is purely better for everybody. These outcomes may not all be fair, but they all have the advantage that no "money is left on the table" -- any "improvement" we could propose for those outcomes makes things worse for at least one person.

Notice that I've also colored areas where Bob and Alice are doing worse than payoff 1. Bob can't enforce Alice's cooperation while defecting more than half the time; Alice would just defect. And vice versa. All of the points within the shaded regions have this property. So not *all* Pareto-optimal solutions can be enforced.

Any point in the white region can be enforced, however. Each player could be watching the statistics of the other player's cooperation, prepared to pull a grim-trigger if the statistics ever stray too far from the target point. This includes so-called ***mutual blackmail*** equilibria, in which both players cooperate with probability slightly better than zero (while threatening to never cooperate at all if the other player detectably diverges from that frequency). This idea -- that 'almost any' outcome can be enforced -- is known as the Folk Theorem in game theory.

The Schelling Pub part is that (particularly with grim-trigger enforcement) everyone has to choose the same equilibrium to enforce; otherwise everyone is stuck playing defect. You'd rather be in even a bad mutual-blackmail type equilibrium, as opposed to selecting incompatible points to enforce. Just like, in Schelling Pub, you'd prefer to meet together at any venue rather than end up at different places.

Furthermore, I would claim that *most* apparent Stag Hunts which you encounter in real life are actually schelling-pub, in the sense that there are many different stags to hunt and it isn't immediately clear which one should be hunted. Each stag will be differently appealing to different people, so it's difficult to establish common knowledge about which one is worth going after together.

***Exercize**: what stags aren't you hunting with the people around you?*

Taking Pareto Improvements

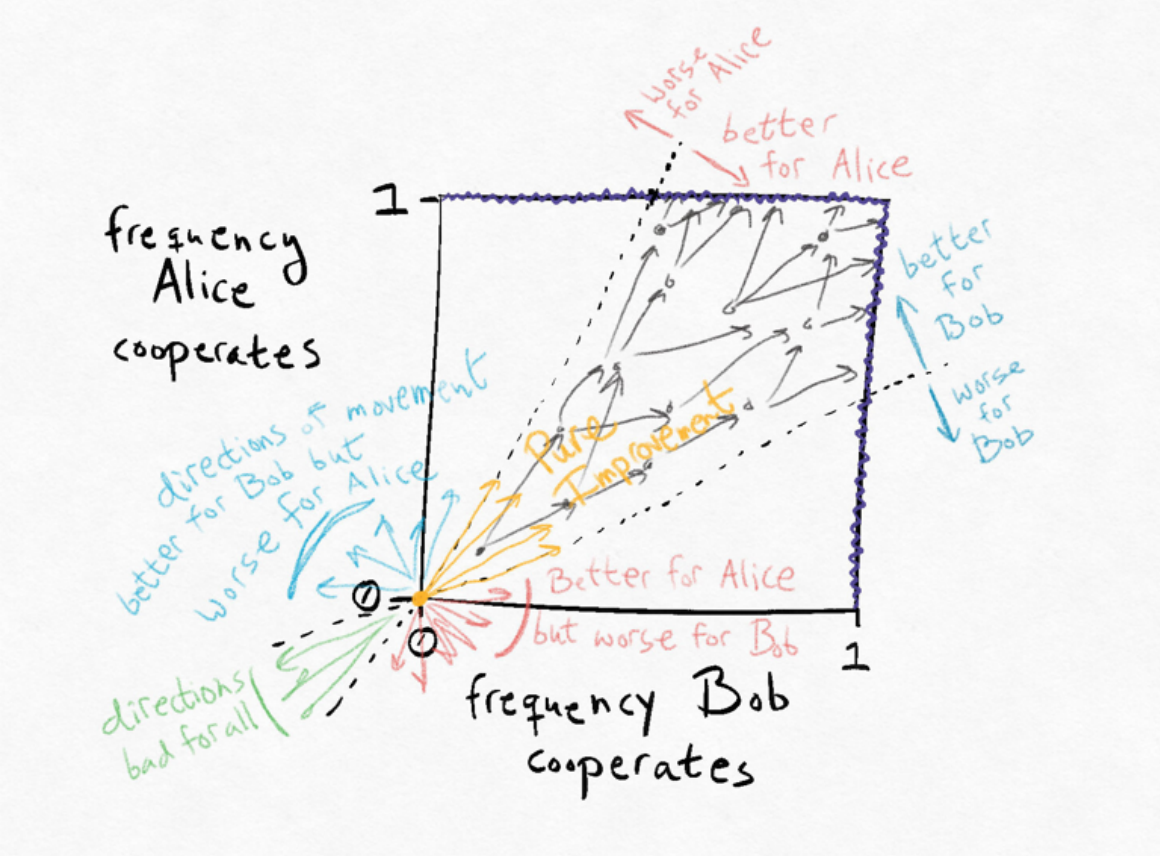
Fortunately, Grim Trigger is not the *only* enforcement mechanism which can be used to build an equilibrium. Grim Trigger creates a crisis in which you've got to guess which equilibrium you're in very quickly, to avoid angering the other player; and no experimentation is allowed. There are much more forgiving strategies (and contrite ones, too, which helps in a different way).

Actually, *even using Grim Trigger to enforce things*, why would you punish the other player for doing something *better for you?* There's no motive for punishing the other player for raising their cooperation frequency.

In a scenario where you don't know which Grim Trigger the other player is using, but you don't think they'll punish you for cooperating *more* than the target, a natural response is for both players to just cooperate a bunch.

So, it can be very valuable to ***use enforcement mechanisms which allow for Pareto improvements.***

Taking Pareto improvements is about moving from the middle to the boundary:



(I've indicated the directions for Pareto improvements starting from the origin in yellow, as well as what happens in other directions; also, I drew a bunch of example Pareto improvements as black arrows to illustrate how Pareto improvements are awesome. Some of the black arrows might not be perfectly within the range of Pareto improvements, sorry about that.)

However, there's also an argument against taking Pareto improvements. If you accept *any* Pareto improvements, you can be exploited in the sense mentioned earlier -- you'll accept any situation, so long as it's not worse for you than where you started. So you will take some pretty poor deals. Notice that one Pareto improvement can prevent a different one -- for example, if you move to (1/2, 1), then you can't move to (1,1/2) via Pareto improvement. So you could always reject a Pareto improvement because you're holding out for a better deal. (This is the *Schelling Pub* aspect of the situation -- there are Pareto-optimal outcomes which are better or worse for different people, so, it's hard to agree on which improvement to take.)

That's where Cooperation between Agents with Different Notions of Fairness comes in. The idea in that post is that you don't take *just any* Pareto improvement -- you have standards of fairness -- but you don't just completely defect for less-than-perfectly-fair deals, either. What this means is that two such agents with incompatible notions of fairness can't get all the way to the Pareto frontier, but the closer their notions of fairness are to each other, the closer they can get. And, if the notions of fairness *are* compatible, they can get all the way.

Moloch is the Folk Theorem

Because of the Folk Theorem, *most* iterated games will have the same properties I've been talking about (not just iterated PD). Specifically, most iterated games will have:

1. **Stag-hunt-like property 1:** There is a Pareto-optimal equilibrium, but there is also an equilibrium far from Pareto-optimal.
2. **The Schelling Pub property:** There are multiple Pareto-optimal equilibria, so that even if you're trying to cooperate, you don't necessarily know which one to aim for; and, different options favor different people, making it a complex negotiation even if you can discuss the problem ahead of time.

There's a third important property which I've been assuming, but which doesn't follow so directly from the Folk Theorem: **the suboptimal equilibrium is "safe", in that you can unilaterally play that way to get some guaranteed utility.** The Pareto-optimal equilibria are not similarly safe; mistakenly playing one of them when other people don't can be worse than the "safe" guarantee from the poor equilibrium.

A game with all three properties is like Stag Hunt with multiple stags (where you all must hunt the same stag to win, but can hunt rabbit alone for a guaranteed mediocre payoff), or Schelling Pub where you can just stay home (you'd rather stay home than go out alone).

Lessons in Slaying Moloch

0. I didn't even address this in this essay, but it's worth mentioning: ***not all conflicts are zero-sum***. In the introduction to the 1980 edition of *The Strategy of Conflict*, Thomas Schelling discusses the reception of the book. He recalls that a prominent political theorist "exclaimed how much this book had done for his thinking, and as he talked with enthusiasm I tried to guess which of my sophisticated ideas in which chapters had made so much difference to him. It turned out it wasn't any particular idea in any particular chapter. Until he read this book, he had simply not comprehended that an inherently non-zero-sum conflict could exist."

1. In situations such as iterated games, ***there's no in-principle pull toward defection***. Prisoner's Dilemma seems paradoxical when we first learn of it (at least, it seemed so to me) because we are not accustomed to such a harsh divide between individual incentives and the common good. But perhaps, as Sarah Constantine speculated in Don't Shoot the Messenger, modern game theory and economics have conditioned us to be used to this conflict due to their emphasis on single-shot interactions. As a result, Moloch comes to sound like an inevitable gravity, pulling everything downwards. This is not necessarily the case.

2. Instead, ***most collective action problems are bargaining problems***. If a solution can be agreed upon, we can generally use weak enforcement mechanisms (social norms) or strong enforcement (centralized governmental enforcement) to carry it out. But, agreeing about the solution may not be easy. The more parties involved, the more difficult.

3. ***Try to keep a path open toward better solutions.*** Since wide adoption of a particular solution can be such an important problem, there's a tendency to treat alternative solutions as the enemy. This bars the way to further progress. (One could loosely characterize this as the difference between religious doctrine and democratic law; religious doctrine trades away the ability to improve in favor of the more powerful consensus-reaching technology of immutable universal law. But of course this oversimplifies things somewhat.) Keeping a path open for improvements is hard, partly because it can create exploitability. But it keeps us from getting stuck in a poor equilibrium.

Book Review: Working With Contracts

Contracts is one of those areas that I always figured I ought to study, at least enough to pick up the basics, but never seemed either interesting or important enough to reach the front of my queue. On top of that, there's a lot of different angles from which to approach the subject: the law-school-style Contracts 101 class covers the legal principles governing contracts, the economists' version abstracts away the practical specifics and talks about contracts in game-theoretic terms, more business-oriented books often focus on negotiation, etc.

"Working With Contracts: What Law School Doesn't Teach You" is about the practical skills needed for working with contracts on an everyday basis - specifically the sort of skills usually picked up on the job by young lawyers. It talks about things like what to look for when reviewing a contract, how to organize contracts, why lawyers use weird words like "heretofore", various gotchas to watch out for, etc. It assumes minimal background knowledge, but also includes lots of technical nuts and bolts. In short, it's the perfect book for someone who wants a technical understanding of real-world contract practice.

This post will review interesting things I learned from the book.

Background Knowledge

First, some very brief background info, which the book itself mostly assumes.

Legally, in order to count as a "contract", we need four main pieces:

- Offer: someone offers a deal
- Acceptance: someone else accepts it
- Consideration: both parties gain something from the deal; it's not a gift
- Mutual understanding: both parties agree on what the deal is and the fact that they've agreed to it

A Contracts 101 class has all sorts of details and gotchas related to these. Notice that "signature on a piece of paper" is not on that list; e.g. oral contracts are entirely enforceable, it's just harder to prove their existence in court. Even implicit contracts are enforceable - e.g. when you order food from a restaurant, you implicitly agree to pay for it, and that's a legally-enforceable contract. That said, we'll focus here on explicit written contracts.

Once formed, a contract acts as custom, private law between the parties. Enforcement of this law goes through civil courts - i.e. if someone breaches the contract, then the counterparty can sue them for damages. Note the "for damages" in that sentence; if a counterparty breaches a contract in a way that doesn't harm you (relative to not breaching), then you probably won't be able to sue them. (Potentially interesting exercise for any lawyers in the audience: figure out a realistic contractual equivalent of Newcomb's problem, where someone agrees to one-box on behalf of someone else but then two-boxes, and claims in court that their decision to two-box benefited the counterparty rather than harming them. I'd bet there's case law on something equivalent to this.)

Note that this is all specific to American law, as is the book. In particular, other countries tend to more often require specific wording, ceremonial actions, and the like in order to make a contract (or component of a contract) enforceable.

What Do Contracts Do?

The “functional” components of a contract can be organized into two main categories: representations and covenants. A representation says that something *has happened* or *is true*; a covenant says that something *will happen* or *will be true*.

Some example representations:

- ABC Corp signs a statement that they have no pending lawsuits against them.
- Bob signs a statement that the house he’s selling contains no lead-based paint or asbestos insulation.
- Carol signs a statement that the forms she provided for a mortgage application are accurate and complete.
- Title Corp signs a statement that there are no outstanding mortgages on a piece of property.

Nominally, each of these is a promise that something is true. However, that’s not quite how they work *functionally*. Functionally, if a counterparty acts based on the assumption that the statement is true and is harmed as a result, then they can sue for damages. In other words, when providing a representation, we provide **insurance** against any damages which result from the representation being false. Bob may not even have checked that the house he’s selling contains no asbestos, and that’s fine - *if* he’s willing to insure the counterparty against any asbestos-related risk.

This idea of insurance becomes important in contract negotiations - there’s a big difference between e.g. “no environmental problems” and “no environmental problems *to the best of their knowledge*”. The former insures against any environmental problems, while the latter insures against any environmental problems which the signer knew about at time of signing. One puts the duty/risk of finding/fixing unknown problems on the signer, while the other puts it on the counterparty.

The other key thing to notice about representations is that they’re *as of the signing date*. When Bob states that his house contains no asbestos, that does not insure against the house previously containing asbestos or containing asbestos in the future. It only needs to be true as of that one moment in time. This becomes relevant in complex multi-stage contracts, where there’s an initial agreement subject to a bunch of conditions and reviews, and the final closing comes later after all that review is done. For instance, in a mortgage there’s an initial agreement subject to the borrower providing lots of forms (credit check, proof of income, proof of insurance, etc…), and the final contract is closed after all that is reviewed. In these situations, the borrower usually makes some representations early on, and then has to “bring down” the representations at closing - i.e. assert that they’re still true.

While representations deal with past and present, covenants deal with the future. They’re the classic idea of contract provisions: precommitments to do something. Some examples:

- ABC Corp agrees to not sell the machinery they’re leasing.
- Bob agrees to not use any lead-based paint on the house he’s buying.

- Carol agrees to maintain minimum levels of insurance on the house she's mortgaging.
- Monitoring Corp agrees to alert Bank if there is any change in the credit rating of Company.

These work basically like you'd expect.

Representations and covenants often run in parallel: a representation that X is true will have a corresponding covenant to make X continue to be true in the future. For instance:

- ABC corp states that they do not currently have any liens on their main plant, and agrees to not create any (i.e. they won't borrow any money with the plant as collateral).
- Carol states that she currently has some level of insurance coverage on her house, and agrees to maintain that level of coverage.

This is mainly for contracts which will be performed over a long time, especially debt contracts. One-off contracts (like a purchase/sale) tend to have relatively few covenants; most of their substance is in the representations.

Parallels to Software Development

Representations and covenants seem pretty straightforward, at least conceptually. One is insurance against some fact being false, the other is a precommitment.

The technical complexity of contracts comes from the interplay between two elements. First:

> The goal of a contract is to describe *with precision* the substance of the meeting of two minds, in language that will be interpreted by each subsequent reader *in exactly the same way*.

In other words, we want no ambiguity, since any ambiguity could later be used by one of the parties to "cheat" their way out of the contract. This creates a headache very familiar to software developers: like programs, contracts mean exactly what they say. There is no "do what I mean" button; we can't write something ambiguous and rely on the system to figure out what we meant.

Second: we don't have perfect knowledge of the future. When making a precommitment in a contract, that precommitment is going to operate fairly mechanically in whatever the future environment looks like. Just like a function written in code may encounter a vast space of unusual inputs in the wild, a precommitment in a contract may interact with a vast space of unusual conditions in the wild. And since we don't know in advance *which* conditions will be encountered, the person writing the code/contract needs to consider the whole possible range. They need to figure out, in advance, what weird corner cases *could* arise.

Put those two pieces together, and the picture should feel *very* familiar to software developers.

The result is that a lawyer's job ends up involving a lot of the same pieces as a software engineer's job. A client/manager says "here's what we want", the lawyer/programmer says "ummm I don't think you really want that, because

<problem> happens if <circumstance>", and they go back-and-forth for a while trying to better define what the client/manager really wants. An example from the book pictures a lawyer reviewing a contract with a client (simplified slightly by me):

> Lawyer: This is a covenant that restricts your business from incurring debt…
>
> Client: That's fine, we don't plan to use any bank financing.
>
> Lawyer: Well, the definition of "debt" used is very broad. For instance, it includes payment plans on any equipment you buy…
>
> Client: Well, we can add some room for that.
>
> Lawyer: How much room do you need?
>
> Client: Based on our current needs, less than $1M at any given time.
>
> Lawyer: But if that new plant you were talking about gets off the ground, won't you need to buy a bunch of new equipment for it?
>
> Client: Good point, we'd better ask for $5M…

This could go on for a while.

Despite the parallels, lawyers are not very *good* software engineers, in general. The most common solution to the sorts of problems above is to throw a patch on it, via two kinds of exceptions:

- Carveouts: action X is generally forbidden, except for special case Y.
- Baskets: action X is generally forbidden, except in amounts below some limit (e.g. the $5M limit in the example above)

Over the course of negotiations, patches are layered on top of patches. An example from the book:

> Little Corp may not transfer any Shares during the term of this Agreement, except for (i) transfers at any time to its Affiliates (including, without limitation, Micro Corp) other than Medium Corp, and (ii) so long as an Event of Default attributable to Big Corp shall have occurred and be continuing, transfers to any Person (including, for the avoidance of doubt, Medium Corp).

This mess is the contractual equivalent of a series of if-statements nested within if-statements. This is, apparently, standard practice for lawyers.

(Another complaint: in a complex contract, it would not be hard to include provisions alongside the table of contents which nullify provisions which appear in the wrong section. Then people reviewing the contract later wouldn't have to read the whole thing in order to make sure they didn't miss anything relevant to their use-case; it would be the contract equivalent of variable scope. My mother's a lawyer in real estate and wills, so I asked her why lawyers don't do this. Her possibly-tongue-in-cheek-answer: might put lawyers out of business. Kidding aside, the bar association engages in some pretty incestuous rent-seeking, but judges have been pushing for decades to make contracts and other legal documents more legible to non-lawyers.)

The “Do What I Mean” Button

A contract writer’s job is much easier than a programmer’s job in one key respect: a contract will ultimately be interpreted by humans. That means we can say the equivalent of “look, you know what I mean, just do that”, *if* we expect that a court will actually know what we mean.

This gives rise to a bunch of standard tricks for invoking the do-what-I-mean button. We’ll talk about three big ones: materiality, reasonableness, and consistency with “ordinary business”/”past practice”.

Materiality

Roughly speaking, materiality means ignoring small things. For instance, compare:

- “Borrower shall not default in its obligations under any contract”, vs
- “Borrower shall not default in its obligations under any material contract”

The first would be breached if e.g. the borrower forgot to update their payment information on their $10 monthly github subscription, and the payment was late. The second would ignore small things like that.

In general, materiality is relative to the size of the business. A $100k oversight would be quite material to most small businesses, but immaterial to AT&T. It’s also relative to the contract - if that $100k oversight is directly relevant to a $300k contract, then it’s material, even if the $300k contract itself is small change to AT&T.

Where’s the cutoff line? That’s for courts to decide, if and when it matters. That’s how pushing the do-what-I-mean button works; you have to rely on the courts to make a sensible decision.

One particularly common usage of materiality: “material adverse change/effect”. Rather than saying “X has no pending lawsuits”, we say “X has no pending lawsuits whose loss would entail a material adverse effect”. Rather than saying “Borrower will notify Lender of any change in their business forecasts”, we say “Borrower will notify Lender of any material adverse change in their business forecasts”. This way a lender or buyer finds out about problems which actually matter, without being inundated with lots of minor details.

Reasonableness

Reasonableness is exactly what it sounds like. It’s saying something that has some obvious loophole to abuse, then giving a stern look and saying “don’t go pulling any bullshit”. Example: “Company shall reimburse X for all of X’s out-of-pocket expenses arising from…” vs “Company shall reimburse X for all of X’s reasonable out-of-pocket expenses arising from…”

Some patterns where reasonableness shows up:

- Reasonable expectations, e.g. “Borrower shall notify Lender of any changes which could reasonably be expected to have a material adverse effect…”
- Consent not to be unreasonably withheld, e.g. “ABC Corp may not X without consent of XYZ Corp, such consent not to be unreasonably withheld.”

- Reasonable efforts, e.g. “Borrower shall obtain X from their insurer.” vs “Borrower shall exert reasonable effort to obtain X from their insurer.”

What would each of these do without the reasonableness clause? In the first case, the borrower could claim that they didn’t expect Obvious Bad Thing to impact their business. In the second case, XYZ Corp could withhold consent for some case they obviously don’t care about in order to extract further concessions from ABC Corp. In the third case, an insurer could simply refuse to provide X, and the borrower wouldn’t be able to do anything about it.

Behaving Normally

Sometimes a lender or prospective buyer wants to say “what you normally do is fine, so do that and don’t go crazy”. Two (similar) standards for this: “in the ordinary course of business” and “consistent with past practice”.

Typical examples:

- “Borrower will not incur any <debt of specific type> except in the ordinary course of business.”
- “ABC Corp will not make any payments to <subsidiary> except in a manner consistent with past practice.”

In general, this is a pretty good way to let business continue as usual without having to go into all the tiny details of what business-as-usual involves, while still ensuring that e.g. a borrowing company doesn’t sell all their assets, distribute the funds as a dividend to a parent company, and then declare bankruptcy.

Remedial Provisions

In general, if a contract is breached, the counterparty can sue for damages. If you want anything else to happen as the result of a breach, then it needs to be included in the contract. In particular, common things triggered by a breach include:

- Termination: counterparty gains the right to terminate the contract
- Acceleration: loaned money must be paid back immediately
- Indemnification: counterparty must be paid for any breach-related damages

The last is somewhat redundant with the court system, but by including it explicitly, the contract can also specify how to calculate damages, how damages are to be paid, caps or exceptions to liability, etc. Rather than leaving such matters to the whims of a court, the contract can specify them.

Termination and acceleration are particularly relevant from a negotiation standpoint - the former for one-shot contracts like sales, and the latter for long-term contracts like debt.

The earlier stages of a complex sale (e.g. a merger/acquisition of a company) involve an agreement to sell *subject to* a long list of conditions being satisfied - i.e. the “due diligence” conditions. If any of those conditions are not met, then the buyer gains the right to terminate the contract - i.e. walk away from the deal. But these things can take months; the last acquisition I saw took around a year. During that time, the buyer may change their mind for reasons entirely unrelated to the seller - e.g. market prices

for the seller's assets may change. The seller wants to prevent the buyer from walking away in a case like that.

This means that the buyer has incentive to ask for very complicated and/or very subjective conditions, to give themselves the opportunity to walk away whenever they want. For instance, if a buyer manages to get a condition which requires "X which is satisfactory *in Buyer's sole discretion*", then the buyer effectively gains a blanket option to walk away from the deal; they can always just claim that some inane detail of X is unsatisfactory. (This is a good example where reasonableness can fix the problem.) In particular, if market conditions change, then the buyer may use that option to negotiate more concessions, like a lower purchase price.

Acceleration has a similar effect in debt deals. Nobody ever wants to accelerate debt; it's a surefire way to end up in bankruptcy court. When a contract breach gives a lender the option to accelerate, what actually happens is that they use that option as leverage to negotiate a new deal. They'll want a higher interest rate, or a claim on more of the borrower's assets, or the like.

Takeaway: just because a contract specifies a particular penalty for breach does not mean that the penalty actually happens. Often, the penalty is really used as an option by one party to renegotiate the contract, and provides leverage for such a negotiation.

Takeaways

Contracts are a lot like computer programs: they're taken very literally, and they could potentially encounter a wide variety of corner cases in the wild. Together, those two pieces make a contract writer's job quite similar to a programmer's job: a client/manager will tell you what they *think* they want, and then you go back-and-forth trying to formulate what they really want.

Compared to (good) software developers, lawyers do not seem to be very good at this; they tend to throw patches on top of patches, creating more corner cases rather than fewer. They don't seem to have even realized that *enforced* scope and modularity are things which one could use in a contract; consequently, every contract must be read in its entirety by anyone relying on it. That puts a sharp limit on the scale of today's contracts.

Unlike programmers, lawyers do have a "do what I mean" button, although its use comes with a cost; it means leaving interpretation to the whims of a court. For many "simple" things, that cost is relatively minor - so contracts can ignore "immaterial" problems, or require "reasonable" behavior, or stipulate consistency with "past practice" and "the course of ordinary business".

Functionally, contracts provide insurance against stated facts being false, and they provide precommitments for the future. They can also stipulate nominal penalties for breach of contract, though in practice these penalties often serve as options to renegotiate (with leverage) rather than actually being used.

Comparative advantage and when to blow up your island

This is a linkpost for https://dynomight.net/2020/09/11/comparative-advantage-and-when-to-blow-up-your-island/

Economists say free trade is good because of "comparative advantage". But what is comparative advantage? Why is it good?

My computational framework for the brain

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

(See comment here for some updates and corrections and retractions. —Steve, 2022)

By now I've written a bunch of blog posts on brain architecture and algorithms, not in any particular order and generally interspersed with long digressions into Artificial General Intelligence. Here I want to summarize my key ideas in one place, to create a slightly better entry point, and something I can refer back to in certain future posts that I'm planning. If you've read every single one of my previous posts (hi mom!), there's not much new here.

In this post, I'm trying to paint a picture. I'm not really trying to justify it, let alone prove it. The justification ultimately has to be: All the pieces are biologically, computationally, and evolutionarily plausible, and the pieces work together to explain absolutely everything known about human psychology and neuroscience. (I believe it! Try me!) Needless to say, I could be wrong in both the big picture and the details (or missing big things). If so, writing this out will hopefully make my wrongness easier to discover!

Pretty much everything I say here *and its opposite* can be found in the cognitive neuroscience literature. (It's a controversial field!) I make no pretense to originality (with one exception noted below), but can't be bothered to put in actual references. My previous posts have a *bit* more background, or just ask me if you're interested. :-P

So let's start in on the 7 guiding principles for how I think about the brain:

1. Two subsystems: "Neocortex" and "Subcortex"

(Update: I have a revised discussion of this topic at my later post Two Subsystems: Learning and Steering.)

This is the starting point. I think it's absolutely critical. The brain consists of two subsystems. The **neocortex** is the home of "human intelligence" as we would recognize it—our beliefs, goals, ability to plan and learn and understand, every aspect of our conscious awareness, etc. etc. (All mammals have a neocortex; birds and lizards have an homologous and functionally-equivalent structure called the "pallium".) Some other parts of the brain (hippocampus, parts of the thalamus & basal ganglia & cerebellum—see further discussion here) help the neocortex do its calculations, and I lump them into the "neocortex subsystem". I'll use the term **subcortex** for the rest of the brain (brainstem, hypothalamus, etc.).

- *Aside: Is this the triune brain theory?* No. Triune brain theory is, from what I gather, a collection of ideas about brain evolution and function, most of which are wrong. One aspect of triune brain theory is putting a lot of emphasis on the distinction between neocortical calculations and subcortical calculations. I like that part. I'm keeping that part, and I'm improving it by expanding the neocortex club to also include the thalamus, hippocampus, lizard pallium, etc., and then I'm ignoring everything else about triune brain theory.

2. Cortical uniformity

I claim that the neocortex is, to a first approximation, architecturally uniform, i.e. all parts of it are running the same generic learning algorithm in a massively-parallelized way.

The two caveats to cortical uniformity (spelled out in more detail at that link) are:

- There are sorta "hyperparameters" on the generic learning algorithm which are set differently in different parts of the neocortex—for example, different regions have different densities of each neuron type, different thresholds for making new connections (which also depend on age), etc. This is not at all surprising; all learning algorithms inevitably have tradeoffs whose optimal settings depend on the domain that they're learning (no free lunch).
    - As one of many examples of how even "generic" learning algorithms benefit from domain-specific hyperparameters, if you've seen a pattern "A then B then C" recur 10 times in a row, you will start unconsciously expecting AB to be followed by C. But "should" you expect AB to be followed by C after seeing ABC only 2 times? Or what if you've seen the pattern ABC recur 72 times in a row, but then saw AB(not C) twice? What "should" a learning algorithm expect in those cases? The answer depends on the domain—how regular vs random are the environmental patterns you're learning? How stable are they over time? The answer is presumably different for low-level visual patterns vs motor control patterns etc.
- There is a gross wiring diagram hardcoded in the genome—i.e., set of connections between different neocortical regions and each other, and other parts of the brain. These connections later get refined and edited during learning. These make the learning process faster and more reliable by bringing together information streams with learnable relationships—for example the wiring diagram seeds strong connections between toe-related motor output areas and toe-related proprioceptive (body position sense) input areas. We *can* learn relations between information streams without any help from the innate wiring diagram, by routing information around the cortex in more convoluted ways—see the Ian Waterman example here—but it's slower, more limited, and may consume conscious attention. Related to this is a diversity of training signals: for example, different parts of the neocortex are trained to predict different signals, and also different parts of the neocortex get different dopamine training signals—or even none at all.

3. Blank-slate neocortex

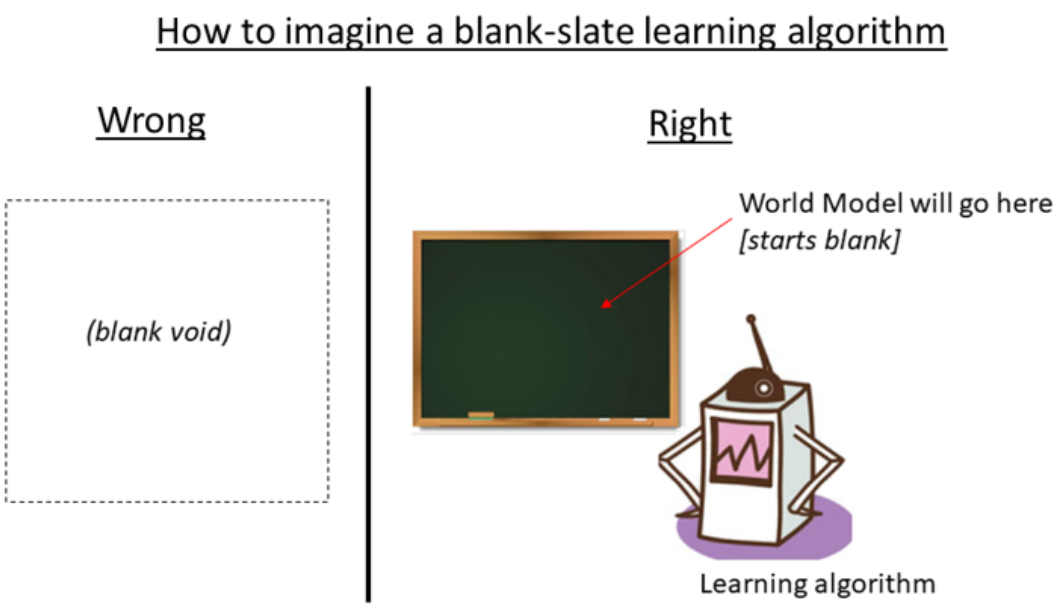
(...But not blank-slate *sub*cortex! More on that below.)

(Update: To avoid confusion, I've more recently been calling this concept "learning-from-scratch"—see discussion in my later post "Learning from Scratch" in the brain.)

I claim that the neocortex (and the rest of the telencephalon and cerebellum) starts out as a "blank slate": Just like an ML model initialized with random weights, the neocortex cannot make any correct predictions or do anything useful until it learns to do so from previous inputs, outputs, and rewards.

In more neuroscience-y (and maybe less provocative) terms, I could say instead: the neocortex is a memory system. It's a *really fancy* memory system—it's highly structured to remember particular kinds of patterns and their relationships, and it comes with a sophisticated query language and so on—but at the end of the day, it's still a type of memory. And like any memory system, it is useless to the organism until it gradually accumulates information. (Suggestively, if you go far enough back, the neocortex and hippocampus evolved out of the same ancient substructure (ref).)

(By the way, I am not saying that the neocortex's algorithm is similar to today's ML algorithms. There's more than one blank-slate learning algorithm! See image.)



A "blank slate" learning algorithm, as I'm using the term, is one that learns information "from scratch"—an example would be a Machine Learning model that starts with random weights and then proceeds with gradient descent. When you imagine a "blank slate" learning algorithm, you should not imagine an empty void that gets filled with data. You should imagine a machine that learns more and better patterns over time, and writes those patterns into a memory bank—and "blank slate" just means that the memory bank starts out empty. There are many such machines, and they will learn different patterns and therefore do different things. See next section, and see also the discussion of hyperparameters in the previous section.

Why do I think that the neocortex starts from a blank slate? Two types of reasons:

- *Details of how I think the neocortical algorithm works:* This is the main reason for me.
  - For example, as I mentioned here, there's a theory I like that says that all feedforward signals (I'll define that in the next section) in the neocortex—which includes all signals coming into the neocortex from the outside it, plus many cortex-to-cortex signals—are re-encoded into the data format that the neocortex can best process—i.e. a set of sparse codes, with low overlap, uniform distribution, and some other nice properties—and this re-encoding is done *by a pseudorandom process*! If that's right, it would seem to categorically rule out anything but a blank-slate starting point.
  - More broadly, we *know* the algorithm can learn new concepts, and new relationships between concepts, without having any of those concepts baked in by evolution—e.g. learning about rocket engine components. So why not consider the possibility that that's *all* it does, from the very beginning? I can see vaguely how that would work, why that would be biologically plausible and evolutionarily adaptive, and I can't currently see any other way that the algorithm can work.
- *Absence of evidence to the contrary:* I have a post Human Instincts, Symbol Grounding, and the Blank-Slate Neocortex where I went through a list of universal human instincts, and didn't see anything inconsistent with a blank-slate neocortex. The subcortex—which is absolutely *not* a blank slate—plays a big role in most of those; for example, the mouse has a brainstem bird-detecting circuit wired directly to a brainstem running-away circuit. (More on this in a later section.) Likewise I've read about the capabilities of newborn humans and other animals, and still don't see any problem. I accept all challenges; try me!

4. What is the neocortical algorithm?

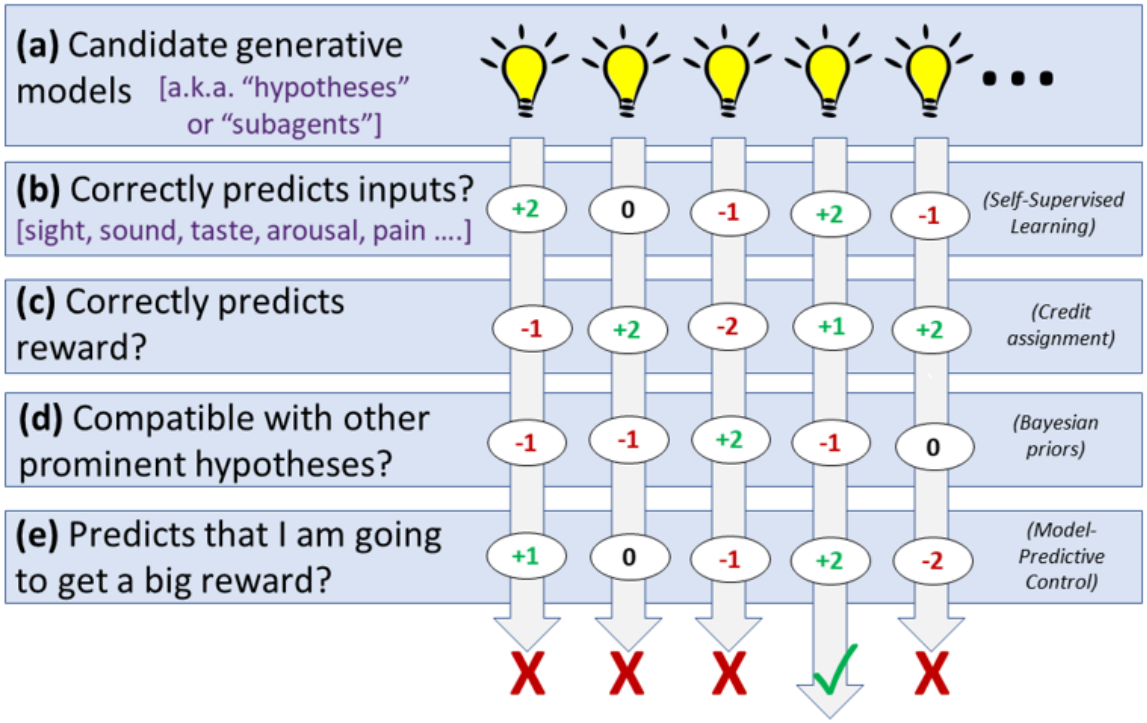
4.1. "Analysis by synthesis" + "Planning by probabilistic inference"

"Analysis by synthesis" means that the neocortex searches through a space of generative models for a model that predicts its upcoming inputs (both external inputs, like vision, and internal inputs, like proprioception and reward). "Planning by probabilistic inference" (term from here) means that we treat our own actions as probabilistic variables to be modeled, just like everything else. In other words, the neocortex's output lines (motor outputs, hormone outputs, etc.) are the same type of signal as any generative model prediction, and processed in the same way.

Here's how those come together. As discussed in Predictive Coding = RL + SL + Bayes + MPC, and shown in this figure below:

- The neocortex *favors* generative models that have been making correct predictions, and *discards* generative models that have been making predictions that are contradicted by input data (or by other favored generative models).
- *And,* the neocortex *favors* generative models which predict larger future reward, and *discards* generative models that predict smaller (or more negative) future reward.

This combination allows both good epistemics (ever-better understanding of the world), and good strategy (planning towards goals) in the same algorithm. This combination *also* has some epistemic and strategic failure modes—e.g. a propensity to wishful thinking—but in a way that seems compatible with human psychology & behavior, which is likewise not perfectly optimal, if you haven't noticed. Again, see the link above for further discussion.



Criteria by which generative models rise to prominence in the neocortex; see Predictive Coding = RL + SL + Bayes + MPC for detailed discussion. Note that (e) is implemented by a very different mechanism than the other parts.

- *Aside: Is this the same as Predictive Coding / Free-Energy Principle?* Sorta. I've read a fair amount of "mainstream" predictive coding (Karl Friston, Andy Clark, etc.), and there

are a few things about it that I like, including the emphasis on generative models predicting upcoming inputs, and the idea of treating neocortical outputs as just another kind of generative model prediction. It also has a lot of other stuff that I disagree with (or don't understand). My account differs from theirs mainly by (1) emphasizing multiple simultaneous generative models that compete & cooperate (cf. "society of mind", multiagent models of mind, etc.), rather than "a" (singular) prior, and (2) restricting discussion to the neocortex subsystem, rather than trying to explain the brain as a whole. In both cases, this may be partly a difference of emphasis & intuitions, rather than fundamental. But I think the core difference is that predictive coding / FEP takes some processes to be foundational principles, whereas I think that those same things do happen, but that they're emergent behaviors that come out of the algorithm under certain conditions. For example, in Predictive Coding & Motor Control I talk about the predictive-coding story that proprioceptive predictions are literally exactly the same as motor outputs. Well, I don't think they're exactly the same. But I *do* think that proprioceptive predictions and motor outputs are the same in *some* cases (but not others), in *some* parts of the neocortex (but not others), and *after* (but not before) the learning algorithm has been running a while. So I kinda wind up in a similar place as predictive coding, in some respects.

4.2. Compositional generative models

Each of the generative models consists of predictions that other generative models are on or off, and/or predictions that input channels (coming from outside the neocortex—vision, hunger, reward, etc.) are on or off. ("It's symbols all the way down.") All the predictions are attached to confidence values, and both the predictions and confidence values are, in general, functions of time (or of other parameters—I'm glossing over some details). The generative models are compositional, because if two of them make disjoint and/or consistent predictions, you can create a new model that simply predicts that both of those two component models are active simultaneously. For example, we can snap together a "purple" generative model and a "jar" generative model to get a "purple jar" generative model. They are also compositional in other ways—for example, you can time-sequence them, by making a generative model that says "Generative model X happens and then Generative model Y happens".

PGM-type message-passing: Among other things, the search process for the best set of simultaneously-active generative model involves something at least vaguely analogous to message-passing (belief propagation) in a probabilistic graphical model. Dileep George's vision model is a well-fleshed-out example.

Hierarchies are part of the story but not everything: Hierarchies are a special case of compositional generative models. A generative model for an image of "85" makes a strong prediction that there is an "8" generative model positioned next to a "5" generative model. The "8" generative model, in turn, makes strong predictions that certain contours and textures are present in the visual input stream.

However, not all relations are hierarchical. The "is-a-bird" model makes a medium-strength prediction that the "is-flying" model is active, *and* the "is-flying" model makes a medium-strength prediction that the "is-a-bird" model is active. Neither is hierarchically above the other.

As another example, the brain has a visual processing hierarchy, but as I understand it, studies show that the brain has loads of connections that don't respect the hierarchy.

Feedforward and feedback signals: There are two important types of signals in the neocortex.

A **"feedback"** signal is a generative model prediction, attached to a confidence level, which includes all the following:

- "I predict that neocortical input line #2433 will be active, with probability 0.6".
- "I predict that generative model #95738 will be active, with probability 0.4".
- "I predict that neocortical output line #185492 will be active, with probability 0.98"—and this one is a self-fulfilling prophecy, as the feedback signal is also the output line!

A **"feedforward"** signal is an announcement that a certain signal is, in fact, active right now, which includes all the following:

- "Neocortical input line #2433 is currently active!"
- "Generative model #95738 is currently active!"

There are about 10× more feedback connections than feedforward connections in the neocortex, I guess for algorithmic reasons I don't currently understand.

In a hierarchy, the top-down signals are feedback, and the bottom-up signals are feedforward.

The terminology here is a bit unfortunate. In a motor output hierarchy, we think of information flowing "forward" from high-level motion plan to low-level muscle control signals, but that's the feed*back* direction. The forward/back terminology works better for sensory input hierarchies. Some people say "top-down" and "bottom-up" instead of "feedback" and "feedforward" respectively, which is nice and intuitive for both input and output hierarchies. But then *that* terminology gets confusing when we talk about non-hierarchical connections. Oh well.

(I'll also note here that "mainstream" predictive coding discussions sometimes talk about feedback signals being associated with confidence *intervals* for analog feedforward signals, rather than confidence *levels* for binary feedforward signals. I changed it on purpose. I like my version better.)

5. The subcortex steers the neocortex towards biologically-adaptive behaviors.

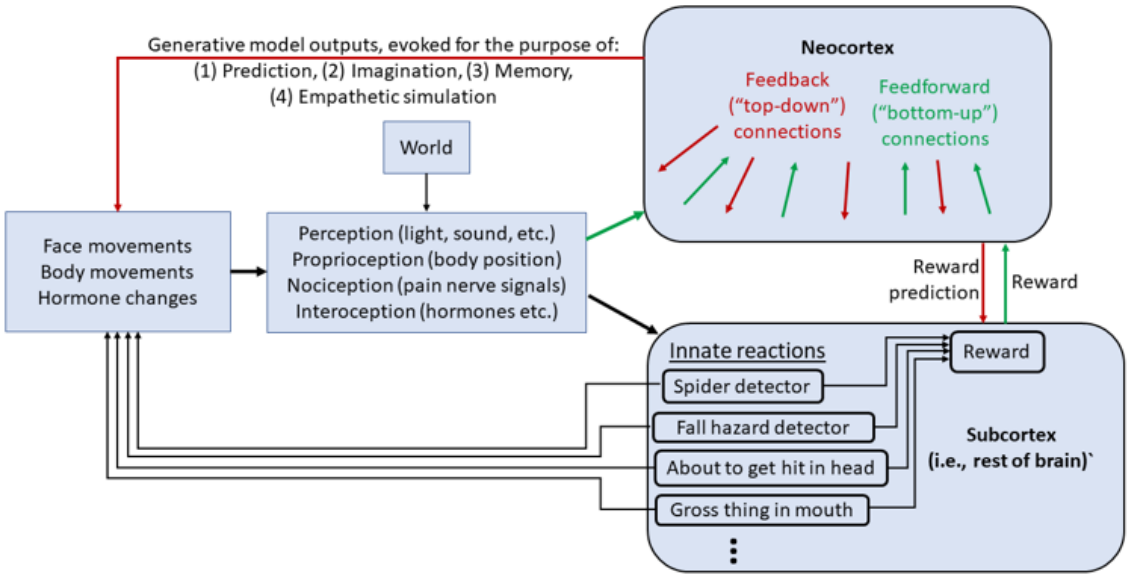
The blank-slate neocortex can learn to predict input patterns, but it needs guidance to do biologically adaptive things. **So one of the jobs of the subcortex is to try to "steer" the neocortex, and the subcortex's main tool for this task is its ability to send rewards to the neocortex at the appropriate times**. Everything that humans reliably and adaptively do with their intelligence, from liking food to making friends, depends on the various reward-determining calculations hardwired into the subcortex.

6. The neocortex is a black box from the perspective of the subcortex. So steering the neocortex is tricky!

Only the neocortex subsystem has an intelligent world-model. Imagine you just lost a big bet, and now you can't pay back your debt to the loan shark. That's bad. The subcortex (hypothalamus & brainstem) needs to send negative rewards to the neocortex. But how can it know? How can the subcortex have any idea what's going on? It has no concept of a "bet", or "debt", or "payment" or "loan shark".

This is a very general problem. I think there are two basic ingredients in the solution.

Here's a diagram to refer to, based on the one I put in Inner Alignment in the Brain:



Schematic illustration of some aspects of the relationship between subcortex & neocortex. See also my previous post Inner Alignment in the Brain for more on this. **(Update June 2021: I would no longer draw the diagram this way, see here. The biggest difference is: I would not draw a direct line from neocortex to a hormone change (for example); instead the cortex would ask the subcortex (hypothalamus + brainstem) to make that hormone change, and then the subcortex might or might not comply with that recommendation. (I guess the way I drew it here is more like somatic marker hypothesis.)))**

6.1 The subcortex can learn what's going on in the world via its own, parallel, sensory-processing system.

Thus, for example, we have the well-known visual processing system in our visual cortex, *and* we have the lesser-known visual processing system in our midbrain (superior colliculus). Ditto for touch, smell, proprioception, nociception, etc.

While they have similar inputs, these two sensory processing systems could not be more different!! The neocortex fits its inputs into a huge, open-ended predictive world-model, but the subcortex instead has a small and hardwired "ontology" consisting of evolutionarily-relevant inputs that it can recognize like faces, human speech sounds, spiders, snakes, looking down from a great height, various tastes and smells, stimuli that call for flinching, stimuli that one should orient towards, etc. etc., and these hardwired recognition circuits are connected to hardwired responses.

For example, babies learn to recognize faces quickly and reliably in part because the midbrain sensory processing system knows what a face looks like, and when it sees one, it will saccade to it, and thus the neocortex will spend disproportionate time building predictive models of faces.

...Or better yet, instead of saccading to faces itself, the subcortex can *reward the neocortex* each time it detects that it is looking at a face! Then the neocortex will go off looking for faces, using its neocortex-superpowers to learn arbitrary patterns of sensory inputs and motor outputs that tend to result in looking at people's faces.

6.2 The subcortex can see the neocortex's outputs—which include not only prediction but imagination, memory, and empathetic simulations of other people.

~~For example, if the neocortex never predicts or imagines any reward, then the subcortex can guess that the neocortex has a grim assessment of its prospects for the future—I'll discuss that particular example much more in an upcoming post on depression.~~ *(Update: that was wrong; see better discussion here.)*

To squeeze more information out of the neocortex, the subcortex can also "teach" the neocortex to reveal when it is thinking of one of the situations in the subcortex's small hardwired ontology (faces, spiders, sweet tastes, etc.—see above). For example, if the subcortex rewards the neocortex for cringing in advance of pain, then the neocortex will learn to favor pain-prediction generative models that also send out cringe-motor-commands. And thus, eventually, it will *also* start sending weak cringe-motor-commands when imagining future pain, or when empathically simulating someone in pain—and the subcortex can detect *that*, and issue hardwired responses in turn.

(Update: I now think "the subcortex rewards the neocortex for cringing in advance of pain" is probably not quite the right mechanism, see here.)

See Inner Alignment in the Brain for more examples & discussion of all this stuff about steering.

Unlike most of the other stuff here, I haven't seen *anything* in the literature that takes "how does the subcortex steer the neocortex?" to be a problem that needs to be solved, let alone that solves it. (Let me know if you have!) ...Whereas I see it as *The Most Important And Time-Sensitive Problem In All Of Neuroscience*—because if we build neocortex-like AI algorithms, we will need to know how to steer them towards safe and beneficial behaviors!

7. The subcortical algorithms remain largely unknown

I think much less is known about the algorithms of the subcortex (brainstem, hypothalamus, ~~amygdala~~, etc.) (Update: After further research I have promoted the amygdala up to the neocortex subsystem, see discussion here) than about the algorithms of the neocortex. There are a couple issues:

- *The subcortex's algorithms are more complicated than the neocortex's algorithms:* As described above, I think the neocortex has more-or-less one generic learning algorithm. Sure, it consists of many interlocking parts, but it has an overall logic. The subcortex, by contrast, has circuitry for detecting and flinching away from an incoming projectile, circuitry for detecting spiders in the visual field, circuitry for (somehow) implementing lots of different social instincts, etc. etc. I doubt all these things strongly overlap each other, though I don't know that for sure. That makes it harder to figure out what's going on.
    - I don't think the algorithms are "complicated" in the sense of "mysterious and sophisticated". Unlike the neocortex, I don't think these algorithms are doing anything where a machine learning expert couldn't sit down and implement something functionally equivalent in PyTorch right now. I think they are complicated in that they have a complicated specification (*this* kind of input produces *that* kind of output, and this *other* kind of input produces this *other* kind of output, etc. etc. etc.), and this specification what we need to work out.
- *Fewer people are working on subcortical algorithms than the neocortex's algorithms:* The neocortex is the center of human intelligence and cognition. So very exciting! So very monetizable! By contrast, the midbrain seems far less exciting and far less practically useful. Also, the neocortex is nearest the skull, and thus accessible to some experimental techniques (e.g. EEG, MEG, ECoG) that don't work on deeper structures. This is especially limiting when studying live humans, I think.

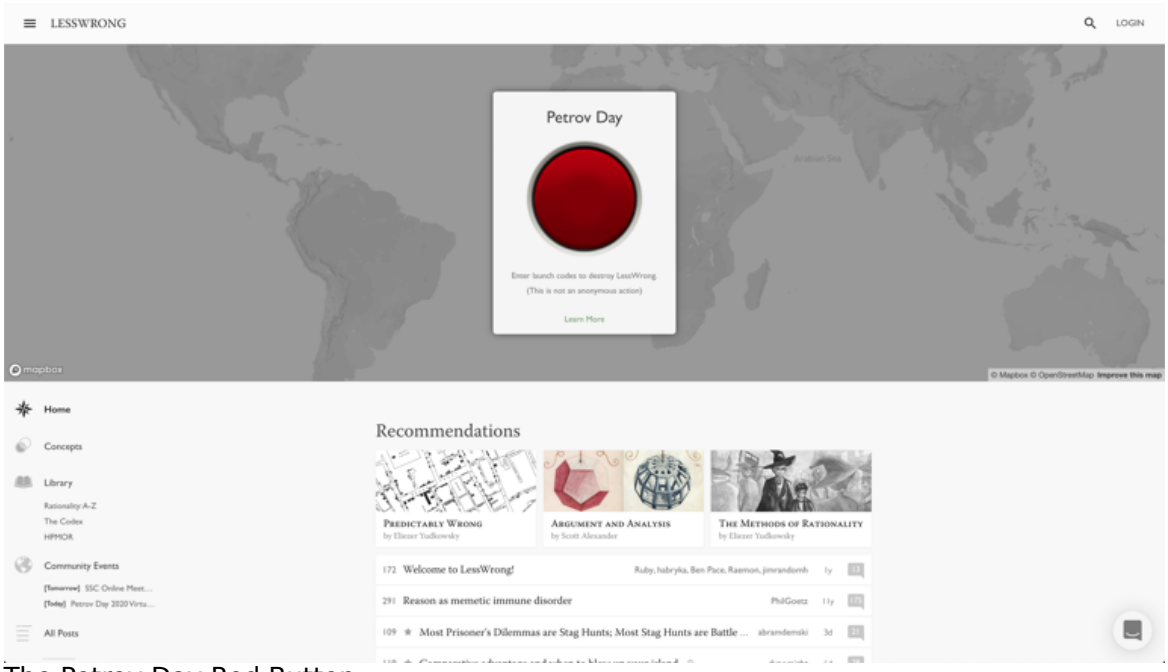
As mentioned above, I am very unhappy about this state of affairs. For the project of building safe and beneficial artificial general intelligence, I feel strongly that it would be better if we reverse-engineered subcortical algorithms first, and neocortical algorithms second.

(*Edited to add:* ...if at all. Like, maybe, armed with a better understanding of how the subcortex steers the neocortex, we'll realize that there's just *no way* to keep a brain-like AGI under human control. Then we can advocate against people continuing to pursue the research program of reverse-engineering neocortical algorithms! Or conversely, if we have a really solid plan to build safe and beneficial brain-like AGIs, we could try to *accelerate* the reverse-engineering of the neocortex, as compared to other paths to AGI. This is a great example of how AGI-related technical safety research can be decision-relevant today even if AGI is centuries away.)

Conclusion

Well, my brief summary wasn't all that brief after all! Congratulations on making it this far! I'm very open to questions, discussion, and criticism. I've already revised my views on all these topics numerous times, and expect to do so again. :-)

Honoring Petrov Day on LessWrong, in 2020



The Petrov Day Red Button.

Just after midnight last night, 270 LessWrong users received the following email.

> *Subject Line:* ***Honoring Petrov Day: I am trusting you with the launch codes***
>
> Hello {username},
>
> On Petrov Day, we celebrate and practice not destroying the world.
>
> It's difficult to know who can be trusted, but today I have selected a group of (270) LessWrong users who I think I can rely on in this way. You've all been given the opportunity to not destroy LessWrong.
>
> This Petrov Day, if you, {username}, enter the launch codes below on LessWrong, the Frontpage will go down for 24 hours, removing a resource thousands of people view every day. Each entrusted user has personalised launch codes, so that it will be clear who nuked the site.
>
> Your personalised codes are: {codes}
>
> I hope to see you in the dawn of tomorrow, with our honor still intact.
>
> –Ben Pace & the LessWrong Team
>
> P.S. Here is the on-site announcement.

Not Destroying the World

Stanislav Petrov once chose not to destroy the world.

As a Lieutenant Colonel of the Soviet Army, Petrov manned the system built to detect whether the US government had fired nuclear weapons on Russia. On September 26th, 1983, the system reported five incoming missiles. Petrov's job was to report this as an attack to his superiors, who would launch a retaliative nuclear response. But instead, contrary to the evidence the systems were giving him, he called it in as a false alarm, for he did not wish to instigate nuclear armageddon. (He later turned out to be correct.)

During the Cold War, many other people had the ability to end the world – presidents, generals, commanders of nuclear subs from many countries, and so on. Fortunately, none of them did. As humanity progresses, the number of people with the ability to end the world increases, and so too does the standard to which we must hold ourselves. We lived up to our responsibilities in the cold war, but barely. (The Global Catastrophic Risks Institute has compiled this list of 60 close calls.)

In 2007, Eliezer named September 26th Petrov Day, and the rationality community has celebrated the holiday ever since. We celebrate Petrov's decision, and we ourselves practice not destroying things, even if it is pleasantly simple to do so.

The Big Red Button

Raymond Arnold has suggested many ways of observing Petrov Day.

You can discuss it with your friends.

You can hold a quiet, dignified ceremony with candles and the beautiful booklets Jim Babcock created.

And you can also play on hard mode: "*During said ceremony, unveil a large red button. If anybody presses the button, the ceremony is over. Go home. Do not speak.*"

This has been a common practice at Petrov Day celebrations in Oxford, Boston, Berkeley, New York, and in other rationalist communities. It is often done with pairs of celebrations, each whose red button (when pressed) brings an end to the partner celebration.

So for the second year, at midnight, I emailed personalized launch codes to 270 LessWrong users. This is over twice the number of users I sent codes to last year (which was 125), and includes a lot more users who use a pseudonym and who I've never met. If any users do submit a set of launch codes, then (once the site is back up) we'll publish their username, and whose unique launch codes they were.

During Saturday 26th September (midnight to midnight Pacific Time), we will practice the skill of sitting together and not pressing harmful buttons.

Relating to the End of Humanity

Humanity could have gone extinct many times.

Petrov Day is a celebration of the world not ending. It's a day where we come together to think about how one man in particular saved the world. We reflect on the ways in which our civilization is fragile and could have ended already, we feel grateful that it has not, and we ask ourselves how we could also save the world.

If you would like to participate in the tradition of Petrov Day on LessWrong this year, and if you feel up to talking directly about it, then you're invited to write a comment and share your own feelings about humanity, extinction, and how you relate to it. There's a few prompts below to help you figure out what to say. Note that not all people are in a position in their lives to focus on preventing an existential catastrophe.

1. **What's at stake for you?** *What are the things you're grateful for, and that you look forward to? What are the things you'd mourn if humanity perished?*
2. **How do you relate to the extinction of humanity?** *What's your story of coming to engage with the fragility of a world beyond the reach of god, and how do you connect to it emotionally?*
3. **Are you taking actions to protect it?** *What are you taking responsibility for in the world, and in what ways are you taking responsibility for the future?*

Finally, if you'd like to participate in a Petrov Day Ceremony today, check out Ray's Petrov event roundup, especially the online New York mega-meetup.

To all, I wish you a safe and stable Petrov Day.

AGI safety from first principles: Introduction

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

This is the first part of a six-part report called AGI safety from first principles, *in which I've attempted to put together the most complete and compelling case I can for why the development of AGI might pose an existential threat. The report stems from my dissatisfaction with existing arguments about the potential risks from AGI. Early work tends to be less relevant in the context of modern machine learning; more recent work is scattered and brief. I originally intended to just summarise other people's arguments, but as this report has grown, it's become more representative of my own views and less representative of anyone else's. So while it covers the standard ideas, I also think that it provides a new perspective on how to think about AGI - one which doesn't take any previous claims for granted, but attempts to work them out from first principles.*

Having said that, the breadth of the topic I'm attempting to cover means that I've included many arguments which are only hastily sketched out, and undoubtedly a number of mistakes. I hope to continue polishing this report, and I welcome feedback and help in doing so. I'm also grateful to many people who have given feedback and encouragement so far. I plan to cross-post some of the most useful comments I've received to the Alignment Forum once I've had a chance to ask permission. I've posted the report itself in six sections; the first and last are shorter framing sections, while the middle four correspond to the four premises of the argument laid out below.

AGI safety from first principles

The key concern motivating technical AGI safety research is that we might build autonomous artificially intelligent agents which are much more intelligent than humans, and which pursue goals that conflict with our own. Human intelligence allows us to coordinate complex societies and deploy advanced technology, and thereby control the world to a greater extent than any other species. But AIs will eventually become more capable than us at the types of tasks by which we maintain and exert that control. If they don't want to obey us, then humanity might become only Earth's second most powerful "species", and lose the ability to create a valuable and worthwhile future.

I'll call this the "second species" argument; I think it's a plausible argument which we should take very seriously.[1] However, the version stated above relies on several vague concepts and intuitions. In this report I'll give the most detailed presentation of the second species argument that I can, highlighting the aspects that I'm still confused about. In particular, I'll defend a version of the second species argument which claims that, without a concerted effort to prevent it, there's a significant chance that:

1. We'll build AIs which are much more intelligent than humans (i.e. superintelligent).
2. Those AIs will be autonomous agents which pursue large-scale goals.

3. Those goals will be misaligned with ours; that is, they will aim towards outcomes that aren't desirable by our standards, and trade off against our goals.
4. The development of such AIs would lead to them gaining control of humanity's future.

While I use many examples from modern deep learning, this report is also intended to apply to AIs developed using very different models, training algorithms, optimisers, or training regimes than the ones we use today. However, many of my arguments would no longer be relevant if the field of AI moves away from focusing on machine learning. I also frequently compare AI development to the evolution of human intelligence; while the two aren't fully analogous, humans are the best example we currently have to ground our thinking about generally intelligent AIs.

---

1. Stuart Russell also refers to this as the "gorilla problem" in his recent book, *Human Compatible*. ↩

What's Wrong with Social Science and How to Fix It: Reflections After Reading 2578 Papers

This is a linkpost for https://fantasticanachronism.com/2020/09/11/whats-wrong-with-social-science-and-how-to-fix-it/

Really interesting analysis of social science papers and replication markets. Some excerpts:

> Over the past year, I have skimmed through 2578 social science papers, spending about 2.5 minutes on each one. This was due to my participation in Replication Markets, a part of DARPA's SCORE program, whose goal is to evaluate the reliability of social science research. 3000 studies were split up into 10 rounds of ~300 studies each. Starting in August 2019, each round consisted of one week of surveys followed by two weeks of market trading. I finished in first place in 3 out 10 survey rounds and 6 out of 10 market rounds. In total, about $200,000 in prize money will be awarded.
>
> The studies were sourced from all social sciences disciplines (economics, psychology, sociology, management, etc.) and were published between 2009 and 2018 (in other words, most of the sample came from the post-replication crisis era).
>
> The average replication probability in the market was 54%; while the replication results are not out yet (175 of the 3000 papers will be replicated), previous experiments have shown that prediction markets work well.[1]
>
> This is what the distribution of my own predictions looks like:[2]
>
> 
>
> [...]
>
> Check out this crazy chart from Yang et al. (2020):
>
> 
>
> Yes, you're reading that right: studies that replicate are cited at the same rate as studies that do not. Publishing your own weak papers is one thing, but citing other people's weak papers? This seemed implausible, so I decided to do my own analysis with a sample of 250 articles from the Replication Markets project. The correlation between citations per year and (market-estimated) probability of replication was -0.05!
>
> You might hypothesize that the citations of non-replicating papers are negative, but negative citations are extremely rare.[5] One study puts the rate at 2.4%. Astonishingly, even *after retraction* the vast majority of citations are positive, and those positive citations continue for decades after retraction.[6]

As in all affairs of man, it once again comes down to Hanlon's Razor. Either:

1. Malice: they know which results are likely false but cite them anyway.
2. or, Stupidity: they can't tell which papers will replicate even though it's quite easy.

Accepting the first option would require a level of cynicism that even I struggle to muster. But the alternative doesn't seem much better: *how can they not know?* I, an idiot with no relevant credentials or knowledge, can fairly accurately determine good research from bad, but all the tenured experts can not? How can they not tell *which papers are retracted*?

I think the most plausible explanation is that scientists don't read the papers they cite, which I suppose involves both malice *and* stupidity.[7] Gwern has an interesting write-up on this question, citing some ingenious bibliographic analyses: "Simkin & Roychowdhury venture a guess that as many as 80% of authors citing a paper have not actually read the original". Once a paper is out there nobody bothers to check it, even though they know there's a 50-50 chance it's false!

Comparative Advantage is Not About Trade

Braudel is probably the most impressive historian I have read. His quantitative estimates of premodern populations and crop yields are exactly the sort of foundation you'd think any understanding of history would be based upon. Yet reading his magnum opus, it became steadily clearer as the books progressed that Braudel was missing some fairly fundamental economic concepts. I couldn't quite put my finger on *what* was missing until a section early in book 3:

> ... these deliberately simple tautologies make more sense to my mind than the so-called 'irrefutable' pseudo-theorem of David Ricardo (1817), whose terms are well known: that the relations between two given countries depend on the "comparative costs" obtaining in them at the point of production

Braudel, apparently, is not convinced by the principle of comparative advantage. What is his objection?

> The division of labor on a world scale (or on world-economy-scale) cannot be described as a concerted agreement made between equal parties and always open to review… Unequal exchange, the origin of the inequality in the world, and, by the same token, the inequality of the world, the invariable generator of trade, are longstanding realities. In the economic poker game, some people have always held better cards than others…

It seems Braudel is under the impression that comparative advantage is only relevant in the context of "equal" exchange or "free" trade or something along those lines.

If an otherwise impressive economic historian is that deeply confused about comparative advantage, then I expect other people are similarly confused. This post is intended to clarify.

The principle of comparative advantage does not require that trade be "free" or "equal" or anything of the sort. When the Portugese or the British seized monopolies on trade with India in the early modern era, those trades were certainly not free or equal. Yet the monopolists would not have made any profit whatsoever unless there were some underlying comparative advantage.

For example, consider an oversimplified model of the salt trade. People historically needed lots of salt to preserve food, yet many inland areas lack local sources, so salt imports were necessary for survival. Transport by ship was historically orders of magnitude more efficient than overland, so a government in control of a major river could grab a monopoly on the salt trade. Since the people living inland could not live without it, the salt monopolist could charge quite high prices - a "trade" arguably not so different from threatening inland farmers with death if they did not pay up. (An exaggeration, since there were other ways to store food and overland smuggling became viable at high enough prices, but I did say it's an oversimplified example.)

Notice that, in this example, there is a clear underlying comparative advantage: the inland farmers have a comparative disadvantage in producing salt, while the ultimate salt supplier (a salt mine or salt pan) has a comparative advantage in salt production. If the farmer could produce salt with the same opportunity cost as the salt mine/pan,

then the monopolist would have no buyers. If the salt mine/pan had the same opportunity cost for obtaining salt as the farmers, then the monopolist would have no supplier. Absent some underlying comparative advantage between two places, the trade monopolist cannot make any profit.

Another example: suppose I'm a transatlantic slave trader, kidnapping people in Africa and shipping them to slave markets in the Americas. It's easy to see how the kidnapping part might be profitable, but why was it profitable to move people across the Atlantic? Why not save the transportation costs, and work the same slaves on plantations *in Africa* rather than plantations in the Americas? Or why not use native American slaves entirely, rather than importing Africans? Ultimately, the profits were because the Americas had a lot lower population density - there was more land, and fewer people to work it. Thus, labor was worth more in the Americas (and that same comparative advantage drove not just the slave trade, but also immigration and automation). Without a comparative advantage, enslaving people might still have been profitable, but there would be no reason to ship them across the Atlantic.

Let's take it a step further. This argument need not involve any trade at all.

Suppose I'm the dictator of some small archipelago. I have total ownership and control over the country's main industries (bananas and construction), and there's an international embargo against trade with my little country, so there's no trade to worry about either internally or externally. Let's say I just want to maximize construction output - although I will still need to order *some* banana-growing in order to keep my construction workers fed.

The question is: who and where do I order to grow bananas, and who and where do I order to build things? To maximize construction, I will want to order people with the largest comparative advantage in banana-growing to specialize in banana-growing, and I will want to order those bananas to be grown on the islands with the largest comparative advantage in banana-growing. (In fact, this is not just relevant to maximization of construction - it applies to pareto-optimal production in general.) There's no trade; I'm just using comparative advantage to figure out how best to deploy my own resources.

Takeaway: comparative advantage is not a principle of *trade*, it's a principle of *optimization*. Pareto-optimal production means specialization by comparative advantage.

The Wiki is Dead, Long Live the Wiki! [help wanted]

12 Virtues°
2-Place and 1-Place Words°
5-and-10°
A Human's Guide To Words°
A Sense That More Is Possible°
AGI Chaining°
AGI Skepticism°
AGI Sputnik Moment°
AI Advantages°
AI Arms Race°
AI Safety Venues°
AI-Complete°
Abolitionism°
Absent-Minded Driver°
Absolute Certainty°
Accelerating Change°
Ad/Bc Problem°
Agent°
Algorithmic Complexity°
Alien Values°
Amount of Evidence°
Anna Salamon°
Anthropomorphism°
Anti-Epistemology°
Antiprediction°
Anvil Problem°
Arguing By Analogy°
Arguing By Definition°
Arguments As Soldiers°
Artificial General Intelligence°
Aspiring Rationalist°
Guessing The Teacher's Password°
Gödel Machine°
Hedon°
Heuristic°
Highly Advanced Epistemology 101 For Beginners°
History of AI Risk Thought°
History of Less Wrong°
Holden Karnofsky°
Hollywood Rationality°
How An Algorithm Feels°
How To Actually Change Your Mind°
Human-AGI Integration and Trade°
I Don'T Know°
Ignorance Prior°
Impossibility°
Impossible World°
Improper Belief°
In-Group Bias°
Incredulity°
Induction°
Inductive Bias°
Infinite Set Atheism°
Instrumental Value°
Intellectual Roles°
Interview Series On Risks From AI °
Introduction To Game Theory (Sequence)°
Introduction To Lesswrong
Paranoid Debating°
Parfit's Hitchhiker°
Peak-End Rule°
Perfectionism°
Perspectives On Intelligence Explosion°
Phlogiston°
Phyg°
Playing To Win°
Policy Debates Should Not Appear One-Sided°
Politics Is The Mind-Killer°
Positivism, Self Deception, and Neuroscience (Sequence)°
Possibility°
Possible World°
Predictionbook°
Preference°
Privileging The Hypothesis°
Probability Theory°
Problem of Verifying Rationality°
Programming Resources°
Prospect Theory°
Puzzle Game Index°
Quantum Immortality°
Quick Reference Guide To The Infinite°
Radical Honesty°
Rational Evidence°
Rationalist°
Rationalist Movement°

That's just a few of them. We imported like 5x as many as these.

With the goal of eventually archiving it fully, we have imported 573 pages and 266,000 words of content from the [old LessWrong wiki]to LessWrong 2.0

The old wiki is a great store of knowledge and still gets two thousand pageviews each day. Incorporating it into the new site gets us at least the following benefits:

- Pages imported from the old wiki now appear in search results on LessWrong proper.
- Pages imported from the old wiki benefit from all the features of new LessWrong such as hover-preview, subscriptions, commenting, and functioning as tags on posts.
- Since LessWrong proper is an active site, hopefully, the wiki content continues to get updated.
- People who land on the old wiki content will more easily find the rest of the awesome content/activity/community on LessWrong proper.
- I like us being the kind of community that when people have spent hundreds (thousands?) of hours generating valuable content, we commit to preserving it.

Quick Links

- List of all pages imported from the LW 1.0 Wiki
- The Tagging/Wiki Dashboard
  - Guide to the New Tagging Dashboard
  - For extra detail on helping, see this section below.
- Join the Tagger Slack.

The Three Import Types

Pages have been imported in one of three ways:

1. **76 are imported as new tags** that can be applied to posts.
2. **111 are merged with existing tag pages**, which is currently in progress and could use some help (see below).
3. **386 are imported as "wiki-only" pages** . These pages cannot be applied to posts and do not currently appear on the Concepts page.

Pages Imported from the Old Wiki

Edit Wiki History Subscribe Discussion

The **old LessWrong wiki** was a companion wiki site to LessWrong 1.0, it was built on MediaWiki software. As of September 2020, the LessWrong 2.0 team is migrating the contents of the old wiki to LessWrong 2.0's new tag/wiki system.

This page contains the list of all pages that are being imported. Imported page fall into one of three types:

- Imported as a new tag page that can be applied to posts (76 pages).
- Merged with an existing tag. The revision history of the old page is copied across and **manual merges of the text are required.** These manual merges are underway. (111 pages).
- Imported as "wiki-only", these pages appear like tag pages except that cannot be applied to posts. Admins can change the wiki-only status of pages. (386 pages).

Further details of imported pages, included state of clean-up work, and links to the original pages can be found in the **Wiki Import Spreadsheet**.

Import Philosophy

The goal is to fully import the contents of the old LessWrong wiki and eventually turn it off. To do that, we must fully preserve any substantive content from the old wiki, even if it is a strange fit for the new wiki. We have erred on the side of being inclusive rather than exclusive, and some pages might seem weird. That was intentional.

Pages Imported as New Tags (76)

The following pages from the old wiki have been imported to LW2.0 as *new tags* that can be applied to posts just like any new tag on LessWrong.

Absurdity Heuristic°	Epistemic Luck°	Personal Identity°
Adaptation Executors°	Everett Branch°	Planning Fallacy°
Adversarial Collaboration°	Evidence of Absence°	Priming°
Affective Death Spiral°	Evidential Decision Theory°	Priors°
Ambient Decision Theory°	Exploratory Engineering°	Rationalist Taboo°
Applause Light°	Fallacy of Gray°	Recursive Self-Improvement°
Astronomical Waste°	Free Will°	Regulation and AI Risk°
Availability Heuristic°	Fuzzies°	Reinforcement Learning°
Bayesian Decision Theory°	Generalization From Fictional Evidence°	Reversed Stupidity Is Not Intelligence°
Belief°	Group Selection°	Road To AI Safety Excellence°
Brain-Computer Interfaces°	Halo Effect°	Roko's Basilisk°
Bystander Effect°	Hedonism°	Scope Insensitivity°
Cached Thought°	Hindsight Bias°	Seed AI°
Causal Decision Theory°	Hope°	Seeing With Fresh Eyes°
Computing Overhang°	Human Universal°	Shut Up and Multiply°
Conformity Bias°	Illusion of Transparency°	Something To Protect°
Conjunction Fallacy°	Infinities In Ethics°	Status Quo Bias°
Correspondence Bias°		

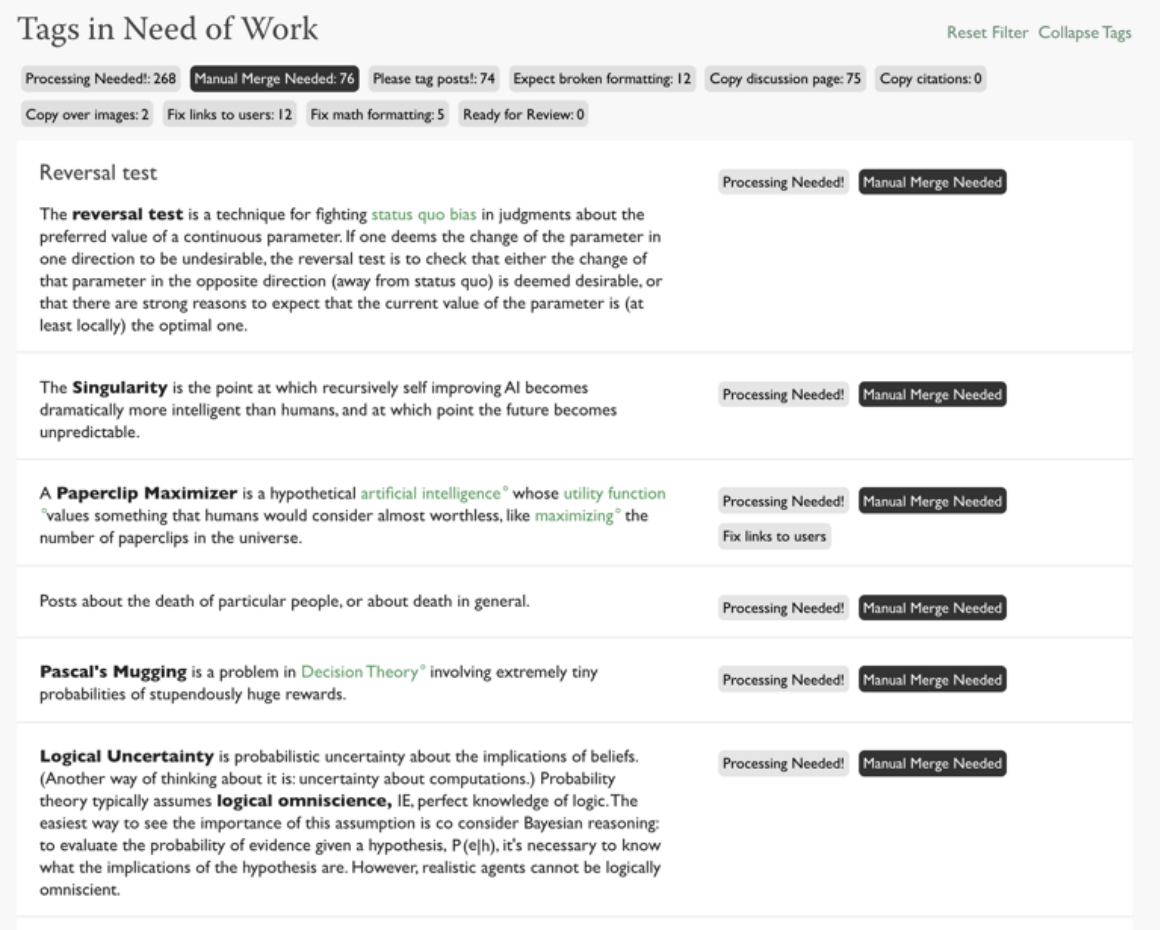
The list of imported of all 573 wiki pages

To be honest, it would be more accurate to say that we are *part-way* through the import. We have completed the programmatic part, and now there remains some manual work to do, hence the *help needed*.

First, there is some general clean-up of links and other elements that didn't import correctly. Second, and more importantly, a **manual text merge** is required for the 111 pages are being merged into existing tags. This means taking the text of the existing tag (if it has any) and combining it appropriately with the old wiki page.

Right now, "merged pages" have the old pages' revision history (click *History* on the tag), but the current text is unchanged.

You can help us out fixing up the wiki import and follow along on completed/incomplete work you can find on the **Tagging/Wiki Dashboard**. More on how to help below, though the hover-overs on the tag flags.



The New Tagging Dashboard

Join the Tagger Slack!!

A couple of weeks ago we created a Slack workspace for dedicated taggers to be able to discuss tagging issues and talk directly to the LessWrong team about it. Following initial success plus good timing with the wiki import campaign, we're opening that Slack to anyone

who wants to help with tagging.

Join the Tagger Slack here

You can also still leave comments on the Tagging Open Call / Discussion Thread.

More Details on Processing Wiki Pages

Here is a more detailed list of the kinds of work to be done:

Merging pages

- Merged pages show only the original current text by default.
- In most cases, this should be pretty straightforward. New tags pages usually have no text or a few sentences that can be easily combined with the text of the imported page.
- When you open a tag page in the full-editor, if it needs merging, there will be links to the latest version of the imported page, and the to page on the old wiki.

Optimizing the opening paragraph

On LessWrong 2.0 (this site), the opening paragraph is what shows on hover-preview for tags, making it very important. It's worth optimizing the opening paragraph of imported pages.

- The approximate title phrase of the page should be **bolded** within the opening paragraph
- The opening paragraph should convey the general topic of the tag clearly

Updating Pages

- Most of the pages on the old wiki have not been updated in several years, and on many topics, a lot more interesting stuff has been said (yay intellectual progress!)
- If you're knowledgeable about a topic, it would be super swell if you updated content to match the latest knowledge.
- A lighter-weight contribution here is to just leave a note in the page's text saying that it's an out-of-date import.

Tagging Relevant Posts

- Imported "tag" pages won't have any posts tagged yet, though most of these have a list of posts already in the text body. Those and other posts are worth adding.
- **For "wiki-only" pages,** there also lists of posts, but we've still decided that's adequate and they don't all need to be tags in addition to that. Feel free to add more posts to the lists in the text body if they're relevant.

Conclusion

I'm excited to have the great content from the old LW wiki now incorporated into the new site, in many ways, it's long overdue.

Thanks to everyone in advance who helps us complete the import!

What happens if you drink acetone?

This is a linkpost for https://dyno-might.github.io/2020/09/14/what-happens-if-you-drink-acetone/

Question: Should you drink acetone?

Answer: No.

But, out of interest, what if you did?

Clarifying “What failure looks like”

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

Thanks to Jess Whittlestone, Daniel Eth, Shahar Avin, Rose Hadshar, Eliana Lorch, Alexis Carlier, Flo Dorner, Kwan Yee Ng, Lewis Hammond, Phil Trammell and Jenny Xiao for valuable conversations, feedback and other support. I am especially grateful to Jess Whittlestone for long conversations and detailed feedback on drafts, and her guidance on which threads to pursue and how to frame this post. All errors are my own.

Epistemic status: My Best Guess

Epistemic effort: ~70 hours of focused work (mostly during FHI’s summer research fellowship), talked to ~10 people.

Introduction

“What failure looks like” is the one of the most comprehensive pictures of what failure to solve the AI alignment problem looks like, in worlds without discontinuous progress in AI. I think it was an excellent and much-needed addition to our understanding of AI risk. Still, if many believe that this is a main source of AI risk, I think it should be fleshed out in more than just one blog post. The original story has two parts; I’m focusing on part 1 because I found it more confusing and nebulous than part 2.

Firstly, I’ll summarise part 1 (hereafter “WFLL1”) as I understand it:

- In the world today, it’s easier to pursue easy-to-measure goals than hard-to-measure goals.
- Machine learning is differentially good at pursuing easy-to-measure goals (assuming that we don’t have a satisfactory technical solution to the intent alignment problem[1]).
- We’ll try to harness this by designing easy-to-measure proxies for what we care about, and deploy AI systems across society which optimize for these proxies (e.g. in law enforcement, legislation and the market).
- We’ll give these AI systems more and more influence (e.g. eventually, the systems running law enforcement may actually be making all the decisions for us).
- Eventually, the proxies for which the AI systems are optimizing will come apart from the goals we truly care about, but by then humanity won’t be able to take back influence, and we’ll have permanently lost some of our ability to steer our trajectory.

WFLL1 is quite thin on some important details:

- WFLL1 does not envisage AI systems directly causing human extinction. So, to constitute an existential risk in itself, the story must involve the lock-in of some

suboptimal world.[2] However, the likelihood that the scenario described in part 1 gets locked-in (especially over very long time horizons) is not entirely clear in the original post.

- It's also not clear how bad this locked-in world would actually be.

I'll focus on the first point: how likely is it that the scenario described in WFLL1 leads to the lock-in of some suboptimal world. I'll finish with some rough thoughts on the second point - how bad/severe that locked-in world might be - and by highlighting some remaining open questions.

Likelihood of lock-in

The scenario described in WFLL1 seems very concerning from a longtermist perspective if it leads to humanity getting stuck on some suboptimal path (I'll refer to this as "lock-in"). But the blog post itself isn't all that clear about why we should expect such lock-in --- i.e. why we won't be able to stop the trend of AI systems optimising for easy-to-measure things before it's too late -- a confusion which has been pointed out before. In this section, I'll talk through some different mechanisms by which this lock-in can occur, discuss some historical precedents for these mechanisms occurring, and then discuss why we might expect the scenario described in WFLL1 to be more likely to lead to lock-in than for the precedents.

The mechanisms for lock-in

Summary: I describe five complementary mechanisms by which the scenario described in WFLL1 (i.e. AI systems across society optimizing for simple proxies at the expense of what we actually want) could get locked-in permanently. The first three mechanisms show how humanity may increasingly depend on the superior reasoning abilities of AIs optimizing for simple proxies to run (e.g.) law enforcement, legislation and the market, despite it being apparent --- at least to some people --- that this will be bad in the long term. The final two mechanisms explain how this may eventually lead to a truly permanent lock-in, rather than merely temporary delays in fixing the problem.

Before diving into the mechanisms, first, let's be clear about the kind of world in which they may play out. The original post assumes that we have not solved intent alignment and that AI is "responsible for" a very large fraction of the economy.[3] So we've made sufficient progress on alignment (and capabilities) such that we can deploy powerful AI systems across society that pursue easy-to-measure objectives, but not hard-to-measure ones.

(1) Short-term incentives and collective action

Most actors (e.g. corporations, governments) have some short-term objectives (e.g. profit, being reelected). These actors will be incentivised to deploy (or sanction the deployment of) AI systems to pursue these short-term objectives. Moreover, even if some of these actors are aware that pursuing proxies in place of true goals is prone to failure, if they decide *not* to use AI then they will likely fall behind in their short-term objectives and therefore lose influence (e.g. be outcompeted, or not reelected). This kind of situation is called a collective action problem, since it requires actors to

coordinate on collectively limiting their use of AI - individual actors are better off (in the short term) by deploying AI anyway.

Example: predictive policy algorithms used in the US are biased against people of colour. We can't debias these algorithms, because we don't know how to design algorithms that pursue the hard-to-measure goal of "fairness". Meanwhile, such algorithms continued to be used. Why? Given crime rate objectives and a limited budget, police departments do better on these objectives by using (cheap) predictive algorithms, compared with hiring more staff to think through bias/fairness issues. So, individual departments are "better off" in the short term (i.e. more likely to meet their objectives and so keep their jobs) if they just keep using predictive algorithms. Even if some department chief realises that this minimization of reported crime rate produces this perverse outcome, they are unable to take straightforward action to fix the problem because this would likely result in increased reported crime rate for their department, impacting that chief's career prospects.

(2) Regulatory capture

The second mechanism is that influential people will benefit from the AIs optimizing for easy-to-measure goals, and they will oppose attempts to put on the brakes. Think of a powerful CEO using AI techniques to maximize profit: they will be incentivised to capture regulators who attempt to stop the use of AI, for example via political donations or lobbying.

Example: Facebook is aware of how user data protection and the spread of viral misinformation led to problems in the 2016 presidential election. Yet they spent $17 million lobbying the US government to assuage regulators who were trying to introduce countervailing regulation in 2019.

(3) Genuine ambiguity

The third mechanism is that there will be genuine ambiguity about whether the scenario described in WFLL1 is good or bad. For a while, humans are overall better off in absolute terms than they are today.[4] From the original post:

> There will be legitimate arguments about whether the implicit long-term purposes being pursued by AI systems are really so much worse than the long-term purposes that would be pursued by the shareholders of public companies or corrupt officials.

This will be heightened by the fact that it's easier to make arguments about things for which you have clear, measurable objectives.[5] So arguments that the world is actually fine will be easier to make, in light of the evidence about how well things are going according to the objectives being pursued by AIs. Arguments that something is going wrong, however, will have no such concrete evidence to support them (they might only be able to appeal to a vague sense that the world just isn't as good as it could be).

This ambiguity will make the collective action problem of the first mechanism even harder to resolve, since disagreement between actors on the severity of a collective problem impedes collective action on that problem.

Example: genuine ambiguity about whether capitalism is “good” or “bad” in the long run. Do negative externalities become catastrophically high, or does growth lead to sufficiently advanced technology fast enough to compensate for these externalities?

(4) Dependency and deskilling

If used widely enough across important societal functions, there may come a time when ceasing to use AI systems would require something tantamount to societal collapse. We can build some intuition for this argument by thinking about electricity, one general purpose technology on which society already depends heavily. Suppose for the sake of argument that some research comes out arguing that our use of electricity will eventually cause our future to be less good than it otherwise could have been. How would humanity respond? I’d expect to see research on potential modifications to our electricity network, and research that tries to undermine the original study. But actually giving up electricity seems unlikely. Even if doing so would not imply total societal collapse, it would at least significantly destabilise society, reducing our ability to deal with other existential risks. This destabilisation would increase the chance of conflict, which would further erode international trust and cooperation and increase risks posed by a range of weapon technologies.[6] And even if giving up electricity was actually the best strategy in expectation, we wouldn’t necessarily do so, due to the problems of short term incentives, collective action, regulatory capture and genuine ambiguity mentioned above.

Furthermore, if we increasingly depend on AIs to make the world work, then humans are unlikely to continue to learn the skills we would need to replace them. In a world where most businesspeople/doctors/lawyers are now AIs, we would likely cut costs by closing down most human business/medical/law schools. This deskilling is an additional reason to think we could be locked-in to a world where AI systems are filling these roles.

(5) Opposition to taking back influence

Whilst these four mechanisms may mean that our attempts at taking back influence from AIs will be delayed, and will come at some cost, surely we will *eventually* realise that something has gone wrong, and make a proper attempt to fix it, even if this involves some costly reskilling and destabilisation?

By way of answering this question in the negative, the original article imagines the following possibility:

> Eventually, large-scale attempts to fix the problem are themselves opposed by the collective optimization of millions of optimizers pursuing simple goals.

This opposition could take two forms. The first can be seen as a continuation of the “genuine ambiguity” mechanism. Simply because the AIs are doing their jobs so well, we may be increasingly unlikely to realise that anything is going wrong. Reported sense of security, healthcare statistics, life satisfaction, GDP, etc. will look great, because it is precisely these proxies for which the AIs are optimizing. As the gap between how things are and how they appear grows, so too will the persuasion/deception abilities of AIs and the world’s incomprehensibility. Eventually, AIs will be able to manipulate human values and our ability to perceive the world in sophisticated ways (think: highly addictive video games, highly persuasive media or education; cf. the human safety problem).

Example: recommender algorithms maximizing click-throughs feed users more extreme content in order to keep them online for longer. Stuart Russell claims that this is an example of an algorithm making its users' values more extreme, in order to better pursue its objective.[7]

Secondly, the AIs may explicitly oppose any attempts to shut them down or otherwise modify their objectives. This is because human attempts to take back influence probably will result in (short term) losses according to their objective functions (e.g. reported sense of security will go down if the systems that have been driving this down are switched off). Therefore, AIs will be incentivised to oppose such changes.

What this opposition looks like depends on how general the AIs are. In CAIS-type scenarios, AIs would probably be limited to the narrow kinds of deception described above. For example, an AI police service with bounded resources minimizing the number of complaints before the end of the day (as a proxy for society's actual safety) will not take long-term, large-scale actions to manipulate human values (e.g. producing advertising to convince the public that complaining is ineffectual). However, it could still take unintended short-term, small-scale actions, if they're helpful for the task before the end of the bound (e.g. offer better protection to people if they don't file complaints).

More general AI could oppose human attempts to take back influence in more concerning ways. For example, it could hamper human attempts at collective action (by dividing people's attention across different issues), cut funding for research on AI systems that can pursue hard-to-measure objectives or undermine the influence of key humans in the opposition movement. Our prospects certainly seem better in CAIS-type scenarios.

Historical precedents

I think the existence of these mechanisms makes the case that it is *possible* that the scenario described in WFLL1 will get locked-in. But is it *plausible?* In particular, will we really fail to make a sufficient attempt to fix the problem before it is irreversibly locked-in? I'll examine three historical precedents which demonstrate the mechanisms playing out, which positively update my credence that it will also play out in the case of WFLL1. However, this reasoning via historical precedents is far from decisive evidence, and I can imagine completely changing my mind if I had more evidence about factors like takeoff speeds and the generality of AI systems.

Climate change

Climate change is a recent example of how mechanisms 1-3 delayed our attempts to solve a problem until some irreversible damage was already done. However, note that the mechanism for the irreversible lock-in is different to WFLL1 (the effects of climate change are locked-in via irreversible physical changes to the climate system, rather than mechanisms 4 and 5 described above).

(1) **Short-term incentives and collective action**

Most electricity generation companies maximize profit by producing electricity from fossil fuels. Despite the unequivocal scientific evidence that burning fossil fuels causes climate change and will probably make us collectively worse off in the long term,

individual companies are better off (in the short term) if they continue to burn fossil fuels. And they will be outcompeted if they don't. The result is a slow-rolling climate catastrophe, despite attempts at collective action like the Kyoto Protocol.

(2) **Regulatory capture**

BP, Shell, Chevron, ExxonMobil and Total have spent €251m lobbying the EU since 2010 in order to water down EU climate legislation.

(3) **Genuine ambiguity**

Consensus among the scientific community that human-caused emissions were contributing to climate change was not established until the 1990s. Even today, some people deny there is a problem. This probably delayed attempts to solve the problem.

The agricultural revolution

The agricultural revolution is a precedent for mechanisms 1 and 4 leading to lock-in of technology that arguably made human life worse (on average) for thousands of years. (The argument that agriculture made human life worse is that increased population density enabled epidemics, farm labour increased physical stress, and malnutrition rose due to the replacement of a varied diet with fewer starchy foods.[8])

(1) **Short-term incentives and collective action**

Humans who harnessed agricultural technology could increase their population relative to their hunter-gatherer peers. Despite the claimed lower levels of health among agriculture communities, their sheer advantage in numbers gave them influence over hunter-gatherers:

> The greater political and military power of farming societies since their inception resulted in the elimination and displacement of late Pleistocene foragers (Bowles, 2011).

So, individual communities were incentivised to convert to agriculture, on pain of being eradicated by more powerful groups who had adopted agriculture.

(4) **Dependency**

Once a community had been depending on agricultural technology for some generations, it would be difficult to regress to a hunter-gatherer lifestyle. They would have been unable to support their increased population, and would probably have lost some skills necessary to be successful hunter-gatherers.

The colonisation of New Zealand

The colonisation of New Zealand is a precedent for a group of humans permanently losing some influence over the future, due to mechanisms 1, 3 and 5. In 1769, the indigenous Māori were the only people in New Zealand, but by 1872, the British (with different values to the Māori) had a substantial amount of influence over New Zealand's future (see this animation of decline in Māori land ownership for a particularly striking illustration of this). Despite the superficial differences, I think this provides a fairly close analogy to WFLL1.[9]

(1) **Short-term incentives and collective action**

The British purchased land from the Māori, in exchange for (e.g.) guns and metal tools. Each tribe was individually better off if they engaged in trade, because guns and tools were economically and militarily valuable; tribes that did not obtain guns were devastated in the Musket Wars. However, tribes became collectively worse off because the British charged unreasonable prices (e.g. in 1848, over 30% of New Zealand was purchased for around NZD 225,000 in today's currency) and could use this land to increase their influence in the longer term (more settlers could arrive and dominate New Zealand's agriculture-based economy).

(3) **Genuine ambiguity**

British goals were initially somewhat aligned with Māori goals. Most early contact was peaceful and welcomed by Māori. In absolute economic terms, the Māori were initially better off thanks to trade with the British. The Māori translation of the Treaty of Waitangi, which the Māori knew would bring more British settlers, was signed by around 540 Māori chiefs.

(5) **Opposition to taking back influence**

However, once the British had established themselves in New Zealand, the best ways to achieve their goals ceased to be aligned with Māori goals. Instead, they turned to manipulation (e.g. breaking agreements about how purchased land would be used), confiscation (e.g. the New Zealand Settlements Act 1863) and conflict (e.g. the New Zealand Wars). For the past 150 years, Māori values have sadly been just one of many determinants of New Zealand's future, and not even a particularly strong one.

How WFLL1 may differ from precedents

These precedents demonstrate that each of the lock-in mechanisms have already played out, making it seem more plausible. The next section discusses how WFLL1 may differ from the precedents. I think these differences suggest that the lock-in mechanisms are a stronger force in WFLL1 than in the precedents, which also positively updates my credence that WFLL1 will be locked-in.

AI may worsen the "genuine ambiguity" mechanism

If AI leads to a proliferation of misinformation (e.g. via language models or deepfakes), then this will probably reduce our ability to reason and reach consensus about what is going wrong. This misinformation need not be sufficiently clever to convince people of falsehoods, it just has to splinter the attention of people who are trying to understand the problem enough to break our attempts at collective action.[10]

Another way in which AI may increase the amount of "genuine ambiguity" we have about the problem is the epistemic bubble/echo chamber phenomenon, supposedly aggravated by social media recommender systems. The claim is that (1) epistemic communities are isolated from each other via (accidental or deliberate) lack of exposure to (reasonable interpretations of) dissenting viewpoints, and (2) recommender systems, by virtue of maximising click-throughs, have worsened this dynamic. If this is true, and epistemic communities disagree about whether specific uses of AI (e.g. AI systems maximizing easy-to-measure goals replacing judges in

courts) are actually serving society's goals, this would make it even harder to reach the consensus required for collective action.

High risk of dependency and deskilling

WFLL1 assumes that AI is "responsible for" a very large fraction of the economy, making it the first time in human history where most humans are no longer required for the functioning of the economy. The agricultural and industrial revolutions involved some amount of deskilling, but humans were still required at most stages of production. However, in WFLL1 it seems likely that humans will heavily depend on AI for the functioning of the economy, making it particularly hard to put on the brakes.

Speed and warning shots

As AI gets more advanced, the world will probably start moving much faster than today (e.g. Christiano once said he thinks the future will be "like the Industrial Revolution but 10x-100x faster"). Naively, this would seem to make things less likely to go well because we'll have less opportunity to identify and act on warning signs.

That said, some amount of speed may be on our side. If the effects of climate change manifested more quickly, it seems more likely that individual actors would be galvanised towards collective action. So faster change seems to make it more likely that the world wakes up to there being a problem, but less likely that we're able to fix the problem if we do.

Another way of putting this might be: too fast, and the first warning shot spells doom; too slow, and warning shots don't show up or get ignored. I'm very uncertain about what the balance will look like with AI. All things considered, perhaps faster progress is worse because human institutions move slowly even when they're galvanised into taking action.

This discussion seems to carry an important practical implication. Since warning shots are only as helpful as our responses to them, it makes sense to set up institutions that are likely to respond effectively to warning shots if they happen. For example, having a clear, reputable literature describing these kinds of risks, which (roughly) predicts what early warning shots would look like, and argues persuasively that things will only get worse in the long run if we continue to use AI to pursue easy-to-measure goals, seems pretty helpful.

Severity of lock-in

The extent to which we should prioritise reducing the risk of a lock-in of WFLL1 also depends on how bad this world actually is. Previous discussion has seen some confusion about this question. Some possibilities include:

- The world is much worse than our current world, because humans eventually become vastly less powerful than AIs and slowly go extinct, in much the same way as insects that become extinct in our world.
- The world is worse than our current world, because (e.g.) despite curing disease and ageing, humans have no real freedom or understanding of the world, and spend their lives in highly addictive but unrewarding virtual realities.

- The world is better than our current world, because humans still have some influence over the future, but our values are only one of many forces, and we can only make use of 1% of the cosmic endowment.
- The world is much better than our current world, because humans lead fairly worthwhile lives, assisted by AIs pursuing proxies. We course-corrected these proxies along the way and they ended up capturing much of what we value. However, we still don't make use of the full cosmic endowment.

It seems that Christiano had something like the third scenario in mind, but it isn't clear to me why this is the most likely. The question is: how bad would the future be, if it is at least somewhat determined by AIs optimizing for easy-to-measure goals, rather than human intentions? I think this is an important open question. If I were to spend more time thinking about it, here are some things I'd do.

Comparison with precedents

In the same way that it was helpful when reasoning about the likelihood of lock-in to think about past examples, then work out how WFLL1 may compare, I think this could be a useful approach to this question. I'll give two examples: both involve systems optimizing for easy-to-measure goals rather than human intentions, but seem to differ in the severity of the outcomes.

CompStat: where optimizing for easy-to-measure goals was net negative?[11]

- CompStat is a system used by police departments in the US.
- It's used to track crime rate and police activity, which ultimately inform the promotion and remuneration of police officers.
- Whilst the system initially made US cities much safer, it ended up leading to:
- Widespread under/misreporting of crime (to push reported crime rate down).
- The targeting of people of the same race and age as those who were committing crimes (to push police activity up).
- In NYC one year, the reported crime rate was down 80%, but in interviews, officers reported it was only down ~40%.
- It seems plausible that pressure on police to pursue these proxies made cities less safe than they would have been without CompStat: there were many other successful initiatives which were introduced alongside CompStat, and there were cases of substantial harm caused to the victims of crime underreporting and unjust targeting.

"Publish or perish": where optimizing for easy-to-measure goals is somewhat harmful but plausibly net positive?

- The pressure to publish papers to succeed in an academic career has some negative effects on the value of academic research.
- However, much important work continues to happen in academia, and it's not obvious that there's a clearly better system that could replace it.

In terms of how WFLL1 may differ from precedents:

- Human institutions incorporate various "corrective mechanisms", e.g. checks and balances in political institutions, and "common sense". However, it's not obvious that AI systems pursuing easy-to-measure goals will have these.
- Most human institutions are at least somewhat interpretable. This means, for example, that humans who tamper with the measurement process to pursue easy-to-measure objectives are prone to being caught, as eventually happened with CompStat. However, ML systems today are currently hard to interpret, and so it may be more difficult to catch interference with the measurement process.

Conclusion

What this post has done:

- Clarified in more detail the mechanisms by which WFLL1 may be locked-in.
- Discussed historical precedents for lock-in via these mechanisms and ways in which WFLL1 differs from these precedents.
- Taken this as cautious but far from decisive evidence that the lock-in of WFLL1 is plausible.
- Pointed out that there is confusion about how bad the future would be if it is partially influenced by AIs optimizing for easy-to-measure goals rather than human intentions.
- Suggested how future work might make progress on this confusion.

As well as clarifying this confusion, future work could:

- Explore the extent to which WFLL1 could increase existential risk by being a *risk factor* in other existential risks, rather than an existential risk in itself.
- Search for historical examples where the mechanisms for lock-in *didn't* play out.
- Think about other ways to reason about the likelihood of lock-in of WFLL1, e.g. via a game theoretic model, or digging into The Age of Em scenario where similar themes play out.

---

1. I'm worried that WFLL1 could happen *even if* we had a satisfactory solution to the intent alignment problem, but I'll leave this possibility for another time. ↩
2. WFLL1 could also increase existential risk by being a *risk factor* in other existential risks, rather than a mechanism for destroying humanity's potential in itself. To give a concrete example: faced with a global pandemic, a health advice algorithm minimising short-term excess mortality may recommend complete social lockdown to prevent the spread of the virus. However, this may ultimately result in higher excess mortality due to the longer term (and harder to measure) effects on mental health and economic prosperity. I think that exploring this possibility is an interesting avenue for future work. ↩

3. The latter assumption is not explicit in the original post, but this comment suggests that it is what Christiano had in mind. Indeed, WFLL1 talks about AI being responsible for running corporations, law enforcement and legislation, so the assumption seems right to me. ↩

4. This isn't clear in the original post, but is clarified in this discussion. ↩

5. I owe this point to Shahar Avin. ↩

6. These pathways by which conflict may increase existential risk are summarised in The Precipice (Ord, 2020, ch. 6). ↩

7. From Human Compatible: "… consider how content-selection algorithms function on social media. They aren't particularly intelligent, but they are in a position to affect the entire world because they directly influence billions of people. Typically, such algorithms are designed to maximize click-through, that is, the probability that the user clicks on presented items. The solution is simply to present items that the user likes to click on, right? Wrong. The solution is to change the user's preferences so that they become more predictable. A more predictable user can be fed items that they are likely to click on, thereby generating more revenue. People with more extreme political views tend to be more predictable in which items they will click on. (Possibly there is a category of articles that die-hard centrists are likely to click on, but it's not easy to imagine what this category consists of.) Like any rational entity, the algorithm learns how to modify the state of its environment—in this case, the user's mind—in order to maximize its own reward.8 The consequences include the resurgence of fascism, the dissolution of the social contract that underpins democracies around the world, and potentially the end of the European Union and NATO. Not bad for a few lines of code, even if it had a helping hand from some humans. Now imagine what a really intelligent algorithm would be able to do." ↩

8. There is some controversy about whether this is the correct interpretation of the paleopathological evidence, but there seems to at least be consensus about the other two downsides (epidemics and physical stress increasing due to agriculture). ↩

9. I got the idea for this analogy from Daniel Kokotajlo's work on takeovers by conquistadors, and trying to think of historical precedents for takeovers where loss of influence happened more gradually. ↩

10. I owe this point to Shahar Avin. ↩

11. Source for these claims about CompStat: this podcast. ↩

Numeracy neglect – A personal postmortem

My failed enlightenment

I've been thinking about my intellectual education, and what I wish had gone differently.

I am 26 years old. I've been reading books and going to school since I was 8. This puts my career as a learner at about 19 years. Honestly? I feel a bit disappointed. I've had a predominantly "humanistic" education, which is a nice way of saying that my gaps in scientific subjects are embarrassing. Meanwhile, I ended up interacting with people who've invested their formative years in getting a solid foundation in mathy and sciency subjects. Inevitably, I found myself envying their skills and wondering where *my* study time has gone, and what do I have to show for it.

In particular, I have diagnosed myself with a condition I call *numeracy neglect.* When I reflect on my education, I find that: (1) I was a bright and precocious kid. (2) I was always very curious and had a strong motivation to understand the world. (3) Despite this, and despite all the resources that society invested in me, I managed to go at least 15 years without learning much about mathematics, physics, chemistry, and computer science (to mention just the basics).

This contradiction pains me, but it also makes me curious. How does something like this happen?

Numeracy neglect

I will focus on mathematics, since it's a subject that most people are taught, but it's typically misunderstood and unappreciated.

Reflecting on my experience, I can identify two problems.

1. *Aesthetic insensitivity*. The inability to experience the beauty of mathematics, and to apply one's general curiosity to it.

2. *Epistemic ignorance*. The inability to see or accept the fact that mathematics is the language of science, and if you don't understand mathematics, you won't understand most science. In general, the inability to understand mathematics' relevance and usefulness in life.

Another way to put it is that the first is a failure to grasp the intrinsic value of mathematics [1], while the second is a failure to understand its instrumental value.

How does this apply to my experience?

Aesthetic insensitivity

I was not born with a natural aversion for mathematics. I remember enjoying arithmetic and geometry in elementary school. Today, I feel a deep curiosity for mathematical subjects; I've also developed, or perhaps rediscovered, an aesthetic appreciation for mathematical concepts.

Yet something went amiss in the age 11 to 23. My grades in mathematics, physics and chemistry were low. I felt little curiosity for these subjects and would only study for the tests. I did no better when I started university. In my first year, I showed little desire to understand statistics, and passed the exam with the minimum grade. It was only later that I (slowly) began to wake up.

Part of it was due to laziness. I was a fast reader and had an excellent memory. This allowed me to excel in most subjects without much work. In contrast, numerate subjects required more dedication and systematic study.

Perhaps it was also a self-esteem issue. Mathematics is *hard*. Studying it forces me to confront failure on a regular basis. It's humiliating to constantly fail on the simplest problems. (I recently downloaded the app Brilliant. It makes me feel anything but.) I was typically praised for intelligence rather than effort. Although some studies have challenged the mindset hypothesis, my experience confirms the trope (at least in hindsight). I had a lot of self-esteem invested in my intelligence, and it was much easier to feel brilliant while repeating philosophy, than to face my struggles with logarithms.

But there was a deeper issue at work. After all, I wasn't lazy when it came to subjects that I cared about. And there were things that I valued more than my self-esteem, such as the need for knowledge.

Well, I can't quite put my finger on it, but I would say that at that time, mathematics was not *really real* to me.

Many students complain that maths is too abstract, too detached from real life. They cannot find enjoyment in it (which is why I speak of 'aesthetic insensitivity'). However, I'm not sure that abstraction is the real problem. (In my case, I spent a lot of energy on philosophy, which can be very abstract.) Rather, the problem may be one of *failing to see the referents*.

When I read philosophy, I felt that the concepts written on the page were referring to something *real* — something the words *stood for*. We could call them 'ideas', or 'objects in ideaspace'. You don't read philosophy to see how the writer combines words on a page. You read it because you are interested in the ideas that the words point to. If you understand the words, you can explore the ideas, play with them, break them apart or combine them. This makes philosophy enjoyable and even beautiful.

In contrast, when I was studying mathematics, I wasn't able to *really* see the referents. I was blind to the reality of mathematical structures. It seemed like a purely syntactical game: we had numbers and symbols, and were taught how to combine them. Of course, I knew that numbers were 'real' in some sense. And I felt that mathematics was generally discovered rather than invented. But I did not get the glorious feeling that I was soaring in ideaspace, stretching my mind to think the unthinkable, and gazing at the fundamental structure of the multiverse. The signifier was there, but the signified was hidden.

It was programming that opened my eyes. As I started learning Python, I understood the difference between the label and the thing. When coding, one works on two levels: the namespace, which contains the labels for the objects, and the objects themselves. You manipulate objects through their names, but the two levels must be kept apart. As a beginner, I didn't understand this. If you are new to coding, there might come a moment when you feel the need to access a variable name at the object level. Imagine you are saving the weight of various dogs, so you declare < terrier = 22 >. Then you want to print <"the weight of {terrier} is {22}">. You can get 22 by calling the variable <terrier>. But how do you print the *name* of the variable? Now you start looking for a function that takes the label down to the object level, such that the function <get_varname(terrier)> would return "terrier". This is a bad idea, because you're mixing up labels and objects, referents and referees. In your code, the variable <terrier> is merely an address for the object <22>. It has nothing to do with the object <"terrier">, unless you explicitly point it there.

At some point, I realized that doing maths is not so different. You are manipulating names that refer to objects. It's true, you cannot touch the objects directly; cannot see them except through their names. Still, the objects *exist*; it isn't a purely syntactical game.

There really *is* a thing like the number 17. Somewhere out there, in ideaspace. You can call it "17" or "seventeen" or "diecisiete" or "xyz123". You might be unaware of its existence, but that won't make it disappear; you may go around saying it isn't prime, but that won't alter its primeness one bit.

When you do maths, you're not just shuffling symbols on the blackboard. You arrange labels meaningfully, and lo and behold — this gives you access to *real mathematical objects*! You can explore them, play with them, break them apart and combine them. You can discover the number 17! You can prove its primeness! Prove that primes are infinite! And this *can* be fun.

To take the coding analogy further, you can imagine a Universal Mathematical Compiler that inputs your notation and translates it into real mathematical operations on real mathematical objects (provided your syntax makes sense). If you understand mathematics, it's like having a virtual machine in your brain that simulates the operations and returns an actual output. This output is not something you invented, or could have predicted in advance. You send a query to the universe, and the universe answers. It's like a message coming from the *other side*. It's your own private window on the inner workings of the multiverse.

Is this enough to start feeling that mathematics is beautiful, and to develop a passion for it? Perhaps not. But it should at least get one beyond the point where mathematics feels boring and empty.

Epistemic ignorance

Even if you don't like mathematics for its own sake, you should eventually realize that without it you cannot understand science. Donald Knuth said: "Science is what we understand well enough to explain to a computer." Like many aphorisms, this goes too far; Darwin's understanding of evolution was 'scientific' in my book, although his science was not advanced enough that he could have specified a faithful simulation (for instance, he didn't know about DNA). However, having a complete mathematical description of a system and being able to predict its behavior and simulate it on a

computer, at least theoretically, is probably as far as scientific understanding can take you.

So why didn't I study more science?

If I could meet my eight-year old self, this is what I'd tell him: "You are curious about the world. To understand the world, you need to understand science. To understand science, you need to understand mathematics. Life is short, and the Art is long. Don't waste your time on dialectical philosophy. Don't get enmeshed in 'critical theory'. Don't think you're smart because you read science news. Acquire at least a fundamental grasp of mathematics, then get some respected textbooks and study the fundamental sciences. Make sure you have a basic knowledge of physics, chemistry and biology, as well as the theory of probability, so that you won't be completely ignorant of the nature of the universe, and you'll be less likely to fall prey to supernatural beliefs, psychologisms and mind-projections. Then you can focus on the disciplines that most interest you."

Why did my younger self fail to grasp this? It wasn't a problem of worldview. Early on, I embraced atheism, the scientific worldview, physical reductionism, the whole package. Yet how great was the mismatch between my professed values and my actual choices!

ME: "I believe physics describes the fundamental laws of the universe."

NOBODY: "So... you're studying physics?"

ME: "Gosh, no! I can't even tell you what thermodynamics is."

NOBODY: "Oh. So you don't care about understanding the universe?"

ME: "How dare you! Of course I do! I thirst for knowledge, truth and understanding!"

NOBODY: "So what are you doing to increase your knowledge?"

ME: "I'm studying Kant. Did you know that space and time are *a priori* forms of experience?"

NOBODY: We need to have a talk.

It all seems a bit absurd today. I have to make an effort of imagination to understand what was going on in my mind. This part is still not clear to me, but for now I can think of three factors.

The first is affect heuristic. I didn't choose which subjects to study based on a ranking of usefulness. Nor was I reflecting on the expected ROI of my study time. I was just going for what felt interesting or titillating or particularly mysterious at any point in time. And if that meant choosing Adorno's *Negative dialectics* over sitting down and doing physics problems, damn the world. This point ties in to aesthetic insensitivity: I felt that mathematics was neither exciting nor beautiful (at least for me) so I didn't take pains to study it. It was much too late when it occurred to me that the way I *feel* about a subject has no bearing on its importance or usefulness.

The second factor is that I had an implicit faith in conceptual, dialectical 'knowledge'. The kind of knowledge that makes you feel smart when you say that light is made out of waves, even though you have no mathematical understanding of what a wave is. I

confused true understanding for being able to recite a great number of facts about different subjects.

The third factor is that I had no practical application for my knowledge. I didn't make predictions. I didn't give myself the chance to be mistaken. I didn't have a mission that forced me to either *learn* or *fail*. At the end of the day, all I did with most of my knowledge was think about it verbally and sometimes talk about it with other people.

Use computers!

If you could change just one thing in how education works today, what would it be?

I will throw my own suggestion, the most direct and effective I can think of: *use computers.*

No, I don't mean giving the students free tablets so they can watch YouTube videos. I mean putting the computer at the *center* of your pedagogical system and teaching mathematics and the other exact sciences *through it*. Currently, the principal medium for doing maths in schools is pen and paper. What if instead people learned theorems and models by reproducing them in code?

After all, computers are a much more *natural* medium for doing that. Despite their limits, computers can actually simulate and run formal systems, as opposed to... Forlorn students scribbling symbols on their notebooks, trying to make the answer come right?

As soon as children can reasonably learn to read and write in natural language, they should be taught the rudiments of programming. This would show them, for starters, that logic and mathematics are *really real* — not mere syntactical games. It would also empower them to grow up as shapers, rather than mere users, of technology.

Later, you could have them simulate the models of physics, chemistry and biology. They could engage in competitive or cooperative games which reward curiosity and stimulate them to think. You could have them design the games themselves, or send them to gather data and test theories. The possibilities are endless. This would not be a replacement for theory and the classical blackboard exercises. But it would provide a practical, engaging field of application which may awaken at least *some students* from chronic boredom and apathy (though it may create special difficulties for others).

Of course, this would require reshaping the whole educational system and turning most teachers into programmers. I didn't say it was easy, or currently feasible. But neither is it beyond the touch of human capacity, I think. Two centuries ago, only 12% of people could read. Today the numbers are basically reversed, with an estimated 14% of the world being illiterate. Yes, some children have serious difficulties with reading, but most can master it to an acceptable degree.

Will something similar happen with programming? I don't know; I can only hope. The spread of literacy was accelerated by cheap newspapers and print books. Personal computers have been around for fifty years, but only in the past two decades they became cheap enough to enter most households. And cheap smartphones are even younger. At least today most people know how to *use* a computer, which is a start.

Considering the rate at which educational institutions evolve, it might take a few decades before programming becomes a basic subject in most schools. In my opinion, it would be worth it to expend some effort in accelerating the process; the payoffs may be very large.

---

[1] This Quora post provides a nice description of the aesthetic value of mathematics (unfortunately I haven't been able to find the author): "The Surreal numbers are useful for broadening our minds, filling us with a sense of awe and marvel at what our own minds are capable of and what things exist in our imagination even if they don't fit in our accidental physical universe."

Social Capital Paradoxes

[Credit for horizontally transmitting these ideas to my brain goes mostly to Jennifer RM, except for the bits at the end about Bowling Alone and The Moral Economy. Apologies to Jennifer for further horizontally spreading.]

Vertical/Horizontal Transmission

The concept of vertical and horizontal transmission felt like a big upgrade in my ability to think about cooperative/noncooperative behavior in practice. The basic idea is to distinguish between symbiotes that are passed on primarily along genetic lines, vs symbiotes which are passed on primarily between unrelated organisms. A symbiote which is vertically transmitted is very likely to be helpful, whereas a symbiote which is horizontally transmitted is very likely to be harmful. (Remember that in biology, "symbiote" means any kind of close relationship between different organisms; symbiosis which is useful to both organisms is *mutualistic*, while symbiosis which is useful to one but harmful to another is *parasitic*.) (This is discussed here on LW in Martin Sustrik's *Coordination Problems in Evolution*.)

We can obviously generalize this quite a bit.

- Infectious diseases tend to be more deadly the higher their transmission rate is. (Diseases with a low transmission rate need to keep their hosts relatively healthy in order to make contact with other potential hosts.)
- Memes which spread vertically are more likely to be beneficial to humans than memes which spread horizontally (at least, beneficial to those human's genes). Religions which are passed through family lines have an incentive to encourage big families, and include ideas which promote healthy, wealthy, sustainable living. Religions which spread primarily to unrelated people have a greater incentive to exploit those people, squeezing every last drop of proselytization out of them.
- Long-term interactions between humans are more likely to be mutualistic, while short-term interactions are more likely to be predatory.
- In general, cooperative behavior is more likely to arise in iterated games; moreso the more iterations there are, and the more probable continued iteration is.

Vertical transmission is just a highly iterated game between the genes of the host and the genes of the symbiote.

Horizontal Transmission Abounds

Wait, but... horizontal transmission appears to be the norm all over the place, including some of the things I hold most dear!

- Religion and tradition tend to favor vertical transmission, while science, education, and reason favor horizontal transmission.
- Free-market economies seem to favor a whole lot of single-shot interactions, rather than the time-tested iterated relationships which would be more common

in earlier economies.
    - To this day, small-town culture favors more highly iterated relationships, whereas big-city culture favors low-iteration. (I've had a decent amount of experience with small-town culture, and a common sentiment is that you have to live somewhere for 20 years before people trust you and treat you as a full member of the community.)

Paradox One: *A lot of good things seem to have a horizontal transfer structure. Some things which I tend to regard with more suspicion have a vertical flavor.*

Horizontal Transmission Seems Wonderful

- The ability to travel easily from community to community allows a person to find the work, cultural environment, and set of friends that's right for them.
- Similarly, the ability to work remotely can be a huge boon, by allowing separate selection of workplace and living environment.
- The first thing I want to do when I hear that vertically-transmitted religion has beneficial memes is to try and get more of those memes for myself!
- Similarly, I've read that many bacteria have the ability to pick up loose genetic material from their environment, and incorporate it into their own genes. (See horizontal gene transfer.) This can be beneficial if those genes are from organisms adapted to the local environment.

Paradox Two: *In an environment where horizontal transfer is rare, opening things up for more horizontal transfer is usually pretty great. But an open environment gives rise to bad dynamics which incentivize closing down.*

If you're in a world where people only ever trade with highly iterated partners, there is probably a lot of low-hanging fruit to be had from trading with a large number of untrusted partners. You could arbitrage price differences, get goods from areas where they're abundant to areas where they're scarce, and generally make a big profit while legitimately helping a lot of people. All for the low price of opening up trade a little bit.

But this threatens the environment of trust and goodwill that you're relying on. An environment with more free trade is one with more scammers, inferior goods, and outright thieves.

YouTube is great for learning things, but it's also full of absolutely terrible demonstration videos which purport to teach you some skill, but instead offer absurd and underdeveloped techniques (these videos are often called "lifehacks" for some reason, if you're unfamiliar with the phenomenon and want to search for it). The videos are being optimized for transmission rather than usefulness. Acquiring useful information requires prudent optimization against this.

Social Capital

Social Capital is, roughly, the amount of trust you have within a group. *Bowling Alone* is a book which researches America's decline in social capital over the course of the 1900s. Trust in the goodwill of strangers took a dramatic dive over that time period,

with corresponding negative consequences (EG, the decline in hitchhiking, the rise of helicopter parenting).

You might think this is due to the increasingly "horizontal" environment. More travel, more free-market capitalism, bigger cities, the decline of small towns; more horizontal spread of memes, by print, radio, television, and internet; more science and education.

And you might be right.

But, counterpoint:

Paradox Three: *Free-market societies have higher social capital.* Citation: *The Moral Economy,* Samuel Bowles.

More generally: a lot of things are a lot better than naive horizontal/vertical thinking would suggest. I've already mentioned that a lot of the things I hold dear seem to have a pretty horizontal transmission model. I *don't* think that's just because I've been taken over by virulent memes.

By the way, my favorite explanation of the decline in social capital over the 1900s is this: there was, for some reason, a huge burst of club-making in the late 1800s, which continued into the early 1900s. These clubs were often very civically active, contributing to a common perception that everyone cooperates together to improve society. This culminated in an extremely high degree of social capital in "The Greatest Generation" -- however, that generation was already starting to forget the club-making/club-attending culture which had fuelled the increase in social capital. Television ultimately killed or put the damper on the clubs, because most people wanted to catch their favorite shows in the evening rather than go out. Social capital gradually declined from then on.

(But, doubtless, there was more going on than just this, and I have no idea how big a factor club culture really plays.)

Questions

1. Why do so many good things have horizontal transmission structures?
2. How should we think about horizontal transmission, normatively? Specifically, "paradox two" is an argument that horizontal-transmission practices, while enticing, can "burn the commons" of collective goodwill by opening up things for predatory/parasitic dynamics. Yet the conclusion seems severe and counterintuitive.
3. Why do free-market societies have higher social capital? How can this be fit into a larger picture in which horizontal transmission structures / few-shot interactions incentivize less cooperative strategies?

Capturing Ideas

Related to: Babble and Prune , What Makes People Intellectually Active? , Zettelkasten .

Summary: if you want to generate more ideas, carry a notebook and write down any thoughts you have.

Citation Needed

I've heard these ideas repeated again and again in different forms: books on note-taking, writing, and creativity; the sorts of interviews where artists are asked "where do you get your ideas?"; and most recently, the final post in the Babble and Prune sequence ("Write"). It would be good of me to gather together some references (especially if there's any academic research on this topic?), but I'm going full-on anecdotal here, and just present to you the most complete version of the thing I can cobble together from memory.

I've personally found this technique to be useful, and anecdotally, so have many other people.

The Basic Technique

Let's say you want to have more ideas in some specific category. For example:

- You want to write fiction. At times, you might feel like you're full to bursting with ideas which you'd like to turn into stories. Yet, when you sit down to do it, you feel like you don't have any ideas.
- You want to do creative research. Maybe so far you only have worked on what problems your advisor gives you. Or maybe you've never worked on research before, and don't know where to start, beyond just reading background literature.
- You're looking for ideas in some school- or work- related context, ideas for Christmas presents, startup ideas, etc etc...

Step 1. Get a pocket notebook, or create a new list in a phone note-taking app, et cetera. The goal is to maximize availability and convenience: to the extent possible, you should be able to capture ideas at any time and place.

Step 2. Write down any ideas that you have. *Any idea at all.* It doesn't have to be good! This is brainstorming. One or two words is fine, so long as you know what it means. Elaborating the idea more will help you remember and may help you generate more ideas, but you can save that for later.

That's it! It's that simple.

Why write a whole post on this? My suspicion is that a lot of people won't raise this very simple strategy to attention to try. I think writing down your ideas has a magical quality to it. You might think:

- "It's not that I'm forgetting my ideas. I just don't have much to say."
- "I'll write down ideas when I have something good enough to write down."
- "I can just write things down later -- my memory isn't that bad. There's no reason to carry a notebook with me."

Or other such thoughts. If you're at a loss for ideas, I suspect these thoughts are wrong. The following "why it works" section is mostly to illustrate that there may be more to this technique than is immediately obvious.

Why It Works

There are some obvious reasons why this might help, and there are also some less-obvious reasons. In approximate order of decreasing obviousness (which is also, as it happens, decreasing order of probability):

Memory

The most obvious thing: you're writing down ideas, so you'll have a list of ideas to look at later.

Actually, writing things down seems to help even if you don't look back at it later: as alkjash mentioned in *Write*, just the act of writing it down might be enough to make it stick in your memory.

Time

It might be that you ordinarily only think about your creative project more-or-less when you sit down to work on it. You don't have any ideas because the only time you spend *trying* to come up with ideas is when you're sitting in front of a blank page.

Putting a notebook in your pocket means you can think about this at any time. Moreover, it creates the affordance: your brain will register this as a thing it can do. (Moreover, you're probably more likely to have interesting ideas when you're out and about, getting all kinds of sensory stimulus.)

Getting out the notebook to write one idea causes you to put more time into thinking of ideas. You might end up writing two or three more you thought of in the time it took you to write the first.

Practice Noticing

It could be that an important aspect of this is: you're intentionally practicing noticing that you have ideas. Like so many other things, perhaps this is something you improve at through practice.

(Note that practicing noticing story ideas vs research ideas vs other kinds of ideas might all be different skills; you don't necessarily get much better at one just because you've practiced the other.)

Reward

You know how level-ups in video games manage to be addictive, even though your brain has no intrinsic love of watching little numbers go up?

I think there's a similar thing here.

Ordinary scenario: You have a passing thought which could be turned into a creative idea. You take no action, and are soon distracted with something else. Your brain concludes: that was a useless thought.

With notebook: You have a passing thought which could be turned into a creative idea. You take out your notebook and pen and start writing. Your brain concludes: looks like that was useful for something! I'll try and come up with more things like that!

Keeping a list of ideas gives you the feeling that you're building something. Each new entry is another brick in a palace of awesomeness.

Beware: **if you don't do anything with your list, this feeling will fade with time.** Your brain will figure out that you're laying bricks in nothing but a ... sad pile of bricks.

The book *Getting Things Done* suggests that you need to build a relationship of trust with your future self, in order for lists like this to work -- writing something down needs to mean that you'll take appropriate action later, even if only to (appropriately) discard most of the ideas.

(This post won't mainly be about how to take your ideas and do something with them, but see the later section "developing ideas".)

Getting Ideas Out (to make room for more!)

Another idea mentioned in *Getting Things Done* is that writing things down gets them out of your head. According to the author, so long as an idea is in your head, it's taking a little bit of your attention. If you write it down in a list, *and if you trust yourself to look at the list later and take appropriate action*, then your brain turns off the reminder and you free up the attention for other things.

Alkjash says something similar in *Write*:

> Fast forward to 2013 and transport yourself to my first summer research program. Every Monday, we give a brief board about that week's progress. I mull ideas on paper over the week before TeXing them up Sunday night.
>
> A curious thing happened - all my progress happened on Monday and Tuesday. I spent the rest of the week meandering around the same ideas, checking special cases and writing up fragments of arguments. On Sunday night I write everything down, and the ideas crystallize on paper. They lose their grip on me, and I move on to new pastures.

(Note that, as a grad student, I've mostly heard rather the opposite: it seems most people make most of their progress the day *before* their weekly meeting, not the day *after*. But, both factors could be in play.)

Attention Leads to Detail

Maybe you've been vaguely dissatisfied with the way your apartment is set up for some time. All of these thoughts feel the same to you; you file them under "apartment is dumb".

Recently, you've decided to do something about it. In order to get started, you'll record your ideas. You stick a notebook in your pocket and start writing thoughts whenever they occur to you.

When you write something down, you have to put it into words. Even if it's just a short phrase, it involves a little bit more detail than your fuzzy mental handle.

Now you notice that you actually have a diversity of complaints, with an implied diversity of remedies. Because you've written each of them down, you can see this diversity at a glance.

This certainly happened to me, when I was first taking notes on rationality (after reading HPMOR). What seemed to me like a single, unified concept kept splintering and splintering.

The opposite could also happen. You vaguely think your cloud of ideas on a subject is huge and diverse, but when you carry a notebook and write every thought down, you find out that everything comes to just two or three points.

Babble & Prune

Since this post was inspired largely by reading *Babble and Prune*, a few words on the relationship to that model:

In the subsections on "practicing noticing" and "reward", I implied that there was some kind of learning going on -- training your brain to notice/produce the kinds of ideas you're looking for. Can we explain this in terms of Babble & Prune? The Babble & Prune model naturally splits this into two distinct types of learning:

- **Training your brain to babble more in that general direction**. If you do anagrams a lot, you learn some really good heuristics for flipping letters around to form new words. Similarly, by paying attention to thoughts of the special kind you want to foster, you may train your brain to flip concepts around in ways more likely to help with that. Confusion about how the subway works might turn into a short story idea about a world where travel works differently. A weird hiccup in your reasoning might turn into a setting where a religion is devoted to precisely that mistake in reasoning. And so on.
- **Training your brain not to prune those things.** You might normally ignore ideas of that kind, which teaches your brain not to bring them to conscious attention, or store them in memory. By paying attention and writing them down, you might adjust those filters, opening up the gates of attention to more of those thoughts.
    - By the way, if you want to get more into conscious attention generally, something to try is the meditative practice of *mental noting*: simply giving labelling words to what is going on in your brain. If you notice that you are thinking, say "thinking" to yourself (out loud if that helps); if you notice

that you are remembering, say "remembering"; if you notice that you are bored, say "bored"; if you notice mental images, say "images"; and so on. The point of this exercise is, amongst other things, to develop awareness of what your brain is doing. In terms of *Babble & Prune*, this softens the filter of conscious attention, giving you access to more of what's going on.

A lot of *Babble & Prune* is oriented toward simply pruning less, an idea which I don't inherently agree with. Yes, pruning less overall might be the right thing for a lot of people. But I'm much more interested in fine-grained adjustments to the pruning filter.

Relatedly -- in the "basic technique" section of this post, I emphasized that you should write down *any ideas at all*. Of course this isn't literally true. You have to calibrate your level of pruning.

- Often, at the start, it's good to dramatically reduce your filtering in order to start the flow of ideas. Write down absolutely any related idea which comes to mind. If none are coming to mind, write down totally unrelated things just so there's something on the page. Write down bad ideas so you'll have something to compare your better ideas to. Write down absolute nonsense so you'll have something to make your bad ideas look good.
- As you start to get good ideas, you'll naturally start to put filters back up. If this technique is, overall, successful, you'll eventually have more interesting ideas than you can execute on. At that point, it's natural to only write down additional ideas which have some significance -- a good-enough chance of being worth your time.
    - Note, however, that it's easy to make the mistake of letting good ideas choke out your source of inspiration. Feynman wrote about how he couldn't come up with any ideas after working on the atom bomb, because he had got an image of himself as someone who worked on big important things. This problem persisted until he told himself that he wasn't allowed to work on important things -- unimportant things only! After that he worked on the physics of spinning planes, which eventually turned out to help with some fundamental problems in quantum physics.

Developing Ideas

Another disagreement I have with *Babble & Prune* is the idea that more layers of filtering is worse. I think the "three gates" (first filter: conscious thought; second filter: saying it out loud; third filter: writing it down) was one of the best and most useful parts of the sequence. (Modulo the fact that for me, the position of the second and third gates is often reversed: I'll write something before I'd say it, due to my close relationship with notebooks.) Yet, I disagree with the contention that this is too many filters. It's too few filters.

Imagine if you had no thinking → speaking filter. In order to avoid saying bad things,

you would have to learn to self-censure your conscious thoughts a lot more. Less ideas would rise to consciousness, and the ones that did would be more constrained by the Gricean maxims and other factors.

Adding a buffer between thinking and speaking allows us to think more, and to develop our thoughts more fully before speaking them aloud.

Similarly, in Zettelkasten, I described my pipeline for developing ideas as consisting of at least four stages:

- **Jot:** Very concise handles used for idea capture. This current post is all about jot-taking. Jots remain meaningful to me for at least a week, but eventually I might have no idea what I was talking about.
- **Gloss:** Paragraph-ish summary I write when I intend to develop a jot more fully. A gloss gives enough of a summary that I won't lose the idea if I let it sit for weeks or months. Takes considerably more focusing to write than a jot.
- **Development:** Free-writing based on an idea. Mostly very informal and narrative-based, dramatizing the ups and downs of an idea as I propose solutions find issues with those solutions, etc.
- **Refinement:** More formal write-ups, often for an audience other than myself. Revising drafts in response to feedback. Engaging with comments on posts. Etc.

This might not fit your use-case. For example, if you're trying to do visual art (for example, drawing a webcomic), your workflow can't all be different types of writing.

The main thing I'm trying to get across here is that *in order to develop the ideas you've captured into something worth sharing,* you probably need several stages.

The *Babble & Prune* model mentioned that babbling is far from totally random, and in particular, as you babble you're mostly mutating ideas (less like random sampling from idea-space, more like a random walk around idea-space). By adding more stages of filtering, I'm suggesting that the way to produce high-quality content is something like simulated annealing: gradually imposing higher and higher standards on our ideas as we continue to mutate them, so that the final result can crystallize into a strong metallic alloy.

"Learning to Summarize with Human Feedback" - OpenAI

https://openai.com/blog/learning-to-summarize-with-human-feedback/

Eliezer's Summary/Thoughts:
https://mobile.twitter.com/ESYudkowsky/status/1301954347933208578

What Decision Theory is Implied By Predictive Processing?

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

At a fairly abstract/stylized level, predictive processing models human cognition and behavior as always minimizing predictive error. Sometimes, the environment is "fixed" and our internal models are updated to match it - e.g. when I see my untied shoelace, my internal model updates to include an untied shoelace. Other times, our internal model is "fixed", and we act on the environment to make it better match the model - e.g. "wanting food" is internally implemented as a strong expectation that I'm going to eat soon, which in turn makes me seek out food in order to make that expectation true. Rather than having a utility function that values food or anything like that, the decision theory implied by predictive processing just has a model in which we obtain food, and we try to make the model match reality.

Abstracting out the key idea: we pack all of the complicated stuff into our world-model, hardcode some things into our world-model which we *want* to be true, then generally try to make the model match reality.

While making the model match reality, there will be knobs we can turn both "in the model" (i.e. updates) and "in reality" (i.e. actions); there's no hard separation between the two. There will be things in both map and reality which we can change, and there will be things in both map and reality which we can't change. It's all treated the same. At first glance, that looks potentially quite useful for embedded agency.

(My own interest in this was piqued partly because a predictive-processing-like decision theory seems likely to produce abstraction boundaries which look like Cartesian boundaries. As in that post, it seems like some of the intuitive arguments we make around decision theories would naturally drop out of a predictive-processing-like decision theory.)

What problems does such a decision theory run into? What sort of things can we hardcode into our world-model without breaking it altogether? What things must be treated as "fixed" when making the model match reality? Does such an approach have any "invariant" implications, i.e. implications independent of *which* model we're trying to match? What further requirements are there on the target model in order for a predictive-processing-style agent to have "good" behavior, in the ways characterized by other decision theories?

This is intended to be an open-ended research question, but off-the-cuff thoughts and links to relevant work are welcome.

Tips for the most immersive video calls

I spend a lot of my day on video calls. Wave is a distributed company, so they're the main way we communicate. But compared to talking in person, they feel unnatural:

- Most people have low-quality microphones and webcams that make them look and sound bad.
- There's a lag between when you say something and when the other person hears it, making it hard to navigate conversational turn-taking.
- If you're using headphones, you can't hear your own voice very well.
- Because of echo cancellation, you often can't talk when someone else is also talking, which makes the conversation flow less well.

I started wondering how much nicer video calls would feel if I fixed these problems. So I spent way too much time fiddling with gear and software. This post summarizes what I've learned. Collectively, I think these recommendations have a pretty big impact: when talking one-on-one to friends with equally good setups, I've been able to go 4+ hours without feeling fatigued.

Epistemic status: best guess; not a professional; almost certainly contains wrong bits. Tell me which ones by comment or email!

omg ben don't make me read your 4500 word doorstopper, just tell me what to do

Here's how I would stack-rank my advice for my past self. (Of course, your personal ranking might be different depending on your situation.)

1. ($depends) Don't work in a space where your noise can bother other people, or vice versa.

2. ($10-30) If you ever have network issues, run a cable between your computer and router. You'll probably need an adapter. (Contrary to popular belief that a bad connection is your ISP's fault, it's more likely to be flaky wifi.)

3. (~$100) Buy open-back headphones, which let you hear your own voice normally and are extremely comfortable.

4. (~$30) Switch from your built-in computer mic to a headset mic (and pop filter), which will sound much better and pick up less noise. Note this requires a headset with detachable cable, like the one I linked above.

   You can now leave yourself unmuted! If the other person also has headphones, you can also talk at the same time. Both of these will make your conversations flow better.

5. ($0) Prefer Zoom to most alternatives; it has higher sound quality, better echo cancellation, and fewer silly behaviors. If you have headphones and a good mic, enable "original sound" to turn off some unnecessary audio filtering.

6. (~$200) Get a second monitor for notes so that you can keep Zoom full-screen on your main monitor. It's easier to stay present if you can always glance at people's faces. (I use an iPad with Sidecar for this; for a dedicated device, the right search term is "portable monitor". Also, if your meetings frequently involve presentations or screensharing, consider getting a third monitor too.)

7. ($0?) Arrange your lighting to cast lots of diffuse light on your face, and move any lights that shine directly into your camera. Lighting makes a bigger difference to image quality than what hardware you use!

8. (~$20-80 if you have a nice camera) Use your camera as a webcam. There's software for Canon, Fujifilm, Nikon, and Sony cameras (Windows-only for Nikon and Sony); for others, if they can output clean HDMI (check this list), you can buy an HDMI capture card. You will also want to be able to plug your camera into a power source, for which you may need a "dummy battery."

9. (~$40 if you have a smartphone with a good camera) Use that as a webcam via Camo.

10. (~$350) If you don't own a nice camera but want one, you can get a used entry-level mirrorless camera + lens + dummy battery + boom arm. See buying tips below.

More detailed recommendations and justifications follow.

Network

Connection problems are the thing that makes video calls suck the most. They do this in three different ways:

1. If your connection ever gets really bad, your audio will break up, which is exhausting to listen to and ruins the flow.

2. Even if it doesn't get that bad, a poor connection will increase *latency*, or the time between when you speak and when the other person hears you. High latency is what causes the dreaded "you first, no *you* first" dance.

3. Finally (and least importantly), a bad connection limits the amount of data you can exchange, forcing you to use lower-quality video. This doesn't really matter if you're using a webcam, but by the end of this post, you might have a good enough camera that it matters.

I wrote a whole post of its own on how to troubleshoot your home network for video calls, but realistically, most connection problems are because wifi sucks and you can avoid them by not using wifi. So, first try running an Ethernet cable between your computer and your router. If you still notice high latency or your connection dropping, or if you really can't run a cable for some reason, check the guide for more troubleshooting advice.

Audio

Video improvements are flashy and noticeable, but audio is the reason you're having the call, thus ultimately more important. So audio comes first.

Get away from other people

This is a basic prerequisite for everything below. Coworking spaces and cafés are nice if you plan to be silent all day, but will make natural-feeling meetings impossible due to your crippling self-consciousness about noise levels. If you're going to be on a call for more than 5 minutes, get your own space.

(If you are committed to taking your meetings in a crowded and noisy space, ignore the rest of the audio section. You're mostly just doomed to crappy calls in this case, though you might be able to limit the damage by getting a nice headset mic and installing krisp.ai.)

If you're talking to someone else who's in a noisy environment, you can apparently also use krisp.ai to filter their audio yourself, though I haven't tried this.

Get full-duplex audio with no echo

One key ingredient to making voice conversations feel "natural" is that both participants need be able to talk and hear the other person talking at the same time ("full-duplex audio"). Full-duplex audio is important because it allows you to talk simultaneously ("overlap") with the other person.

You might think that overlap should be rare, because interrupting someone else is rude. While that's true of large-scale overlaps, we often use small-scale overlaps to negotiate conversational turn-taking (e.g. starting talking when the speaker is trailing off but hasn't finished), or to signify that we're paying attention ("uh-huh," "yeah").

The hard problem of full-duplex audio is that if someone else is talking, their voice is going to come out of your computer's speakers and go back into your microphone. If your computer leaves the microphone on, in that case, it'll end up playing back an "echo" of their own voice to them, which is extremely annoying. So video call tries to filter out feedback from your speakers into your microphone, which is called echo cancellation.

Unfortunately, removing *only* the speaker echo from your microphone stream is really hard to do. So instead, the software often ends up completely muting your mic if someone else is talking. If you've ever tried to micro-overlap with someone and noticed that their audio cut out briefly, that's what's going on.

If your listeners can't overlap with you, it's harder to tell whether they're following along, and it's harder to negotiate whose turn it is to speak. This makes the conversation feel less natural, especially in larger groups.

To get full-duplex audio, you need to (a) have an audio setup that doesn't produce echoes, then (b) convince your video call app not to try to suppress echoes.

(a), "an audio setup that doesn't produce echoes," means that your microphone should not pick up any sound from your speakers. In practice this means that your "speakers" must be headphones.

(b), “convince your video call app not to try to suppress echoes,” seemed surprisingly tricky when I tried to research it, because each video call app has its own heuristics for when to engage echo-cancellation.

So I did my own tests of echo cancellation Zoom, Skype, and Hangouts in Chrome and Firefox. I started a chat between two computers, both with headphones attached—a setup that should have required no echo cancellation. I then played music into the microphone of one computer. On the other, I spoke into the microphone and listened for whether the music got quieter.

Zoom, Skype and Hangouts in Firefox all seemed to slightly decrease the audio volume when I spoke, indicating light echo cancellation. For Hangouts in Chrome, the audio cut out *completely* every time I said anything. In Zoom, I was able to eliminate all echo cancellation by selecting the “use original audio” option, which you can also permanently enable for particular audio devices—I’d recommend doing this.

Throw your wireless headset in the trash

The gear I recommend in this guide is all wired, not Bluetooth. While Bluetooth seems like it should be great, in practice it has horrible problems with audio latency, quality and reliability. Also, I don’t think wireless open-back headphones (see below) exist.

If you finished the previous paragraph and still think you can get away with using wireless audio gear, read the post at the link :)

Hear yourself clearly with open-back headphones

Most headphones are *closed-back*, which means they form an acoustic seal over your ear that attenuates outside sound. This is good for “noise isolation” when you’re listening to music. But it’s bad in calls because it also isolates you from your own voice, making you sound muffled and unnatural to yourself. (The same thing also happens with any earbuds that form a seal, i.e. pretty much everything except EarPods or non-Pro AirPods.)

Personally, without the feedback from hearing myself, I also tend to start speaking louder or shouting on calls. This tires out my voice, and can get stressful for whoever I’m ~~shouting at~~ talking to.

To avoid this, you can buy *open-back headphones*, which have mesh instead of a closed covering over your ears. I bought the Philips SHP9500, which I like a lot; I haven’t tested any other pairs. (I chose a low-end pair because for video calls, sound quality will mostly be limited by people’s microphones; if you want to use the same ones to listen to music, you might want a higher-end pair.)

As an extra bonus, open-back headphones are *way* more comfortable because they get less hot. I didn’t realize beforehand how much difference this would make, but it’s amazing to be able to wear headphones all day without my ears overheating!

Note that open-back headphones “leak” sound, so anyone near you will hear the other side of the conversation as well. This isn’t a problem if you have your own space, but they’re not suitable for shared spaces. You might think the sound could leak into your own microphone and cause an echo, but I tested and it’s too quiet for that unless you set the volume uncomfortably high.

Don’t mute

This isn’t about equipment per se, but it has implications for your equipment choices. Quoting Matt Mullenweg, founder of one of the earliest and largest fully-distributed companies:

> One heterodox recommendation I have for audio and video calls when you’re working in a distributed fashion is not to mute, if you can help it. When you’re speaking to a muted room, it’s eerie and unnatural — you feel alone even if you can see other people’s faces. You lose all of those spontaneous reactions that keep a conversation flowing. If you ask someone a question, or they want to jump in, they have to wait to unmute. I also don’t love the “unmute to raise your hand” behavior, as it lends itself to meetings where people are just waiting their turn to speak instead of truly listening.

I strongly agree with this and prefer for the people I’m talking with to stay unmuted unless they have a crappy mic that picks up a lot of noise. Which won’t be your problem as long as you…

Get a better microphone

Most non-standalone computer microphones, including ones on fancy headsets, sound ear-bleedingly bad. (The 2020 MacBook Pro microphone is okay.) You can sound way more pleasant to your colleagues by getting a nicer one. For instance, compare me reading Edward Lear on the following mics:

Mic	Recording	Commentary
2014 iMac		Sounds like a tin can, because it is
Jabra Evolve 70		Wirecutter rec; sounds like a bad head cold
2020 MacBook Pro 13”		“Studio quality” my foot; try “moderate head cold”
V-Moda BoomPro		Stupid name, $30, actually sounds ok

The best “can’t mess up” microphone option is the last one in the table, the V-Moda BoomPro (with a foam windscreen), which attaches to your headphones in place of a standard 3.5mm audio cable.[1] It sounds much clearer and less muffled than any headset’s built-in mic. It’ll also pick up less background noise (e.g. typing) than *any* mic, by virtue of being closer to your mouth, which makes it easier to follow “don’t mute” above. However, it won’t sound quite as natural as a non-headset mic.

If you want something that sounds even more realistic, your best bet is something like an AT2005 positioned less than six inches from your face using a boom arm that’s clamped to your desk. If you don’t want your mic to be visible, it may require some

zooming/cropping of your camera setup to get it that close. Compare it to the BoomPro (you'll need headphones to hear the difference well):

BoomPro	AT2005

I tested a few different microphones, and listened to recordings of many more. I haven't trained myself to detect small quality differences, but my tentative conclusions are:

1. Distance to your mouth dominates nearly everything. For the mics I tested, 6 inches vs 12 inches made as big of a difference as switching mics.

   (That's because from the microphone's perspective, doubling the distance makes your voice 4x quieter, or equivalently, makes room noise 4x louder. It also makes louder e.g. the echoes of your voice—leading the microphone to produce a "boomy" sound.)

   Here's two fairly different-sounding microphones at 6 inches vs 12:

Model	6 inches	12 inches
Blue Yeti		
AT2005USB		

2. Low-end condenser microphones sound somewhat "fuller" or more natural than dynamics, which can sound a little "muffled." For example, the Blue Yeti above is a condenser while the AT2005USB is a dynamic.

   (Dynamic mics are sometimes said to "reject noise" better than condensers, but I couldn't find anyone making this claim who explained why.[2])

3. Mics under $50, and headset mics, sound noticeably bad even when close to your mouth. Outside of that, it seemed like sound quality depended as much or more on things other than microphone quality, like how much your space echoes. (There's a surprisingly large genre of YouTubers demoing expensive mics with bad-sounding setups.) So it seemed like higher-end microphones would probably be wasted in most video call setups.

Based on this, I suggest the AT2005 as a widely-recommended mic in the lowest "doesn't sound noticeably bad" price tier, that can be stuck on a boom arm with minimal ceremony.

Other microphone comparisons for further reading:

- Wirecutter USB mic comparisons

- Marco Arment on podcasting microphones
- Matt Mullenweg “Don’t mute, get a better headset”
- rtings.com headset recording quality tests

Listen to yourself

In video calls, unlike real life, what you hear is not the same as what’s heard by the person you’re talking to. And mics are fiddly enough that your particular setup and mic technique matters a lot. So it’s really useful to listen to how your audio sounds. You can do this with a web app like miccheck.me. The most common mic problems are plosive pops and harsh sibilants—pick sentences that will cover those. The Harvard Sentences are a good starting point.

If those consonants sound bad, you might need a better windscreen, or to change how your mic is positioned. For instance, if you have a headset mic, you should position it just beside the corner of your mouth—*not* directly in front—so that you’re not breathing/spitting into it.

Video

Use a dedicated monitor

I recently started putting my active call in full-screen mode on my primary (27”) monitor, and using a second monitor for notes or other activity. This turned out to make a surprisingly big difference to how immersive the call felt. (Maybe I should have been clued in by the fact that this feature is called “immersive mode”?) It’s amazing for keeping me focused and present. Possible reasons why:

- In windowed mode, Zoom keeps your preview at the top of the screen in a bar of its own, but in fullscreen mode there’s no bar, just a floating preview window (which you can also hide). That means there’s a lot more screen available for other people’s video.
- Hiding the window bezel, task bar, etc. make it much less salient that you’re talking through a computer, and the user interface is less likely to distract you.
- In windowed mode, I’d sometimes end up tabbing away to look at another window, and lazily forget to tab back. Then I’d spend a lot of the call not looking at people and feeling less connected.

I use an iPad for this, but you can also buy dedicated “portable monitors” for under $200 that would serve this purpose well.

Improve your lighting

The best way to get a sharper image on any camera is to put more light into the sensor. Laptop webcams have terrible image quality, but a laptop webcam with good lighting will look better than a fancy camera with bad lighting:

Left: 2014 iMac webcam destroyed by MAXIMUM BACKLIGHT. Middle: Fancy Sony A6000 struggles too. Right: iMac webcam gloriously lit by a lumenator bounced off the wall in front of me. Click for fullsize.

The two basic rules of lighting are:

1. Cast lots of diffuse light on your face to make sure it's brighter than the background. (Also, the more light that's hitting the scene, the less grainy your image will appear.)

   The easiest way to do this is to put your desk in front of a window; second-easiest is to bounce artificial lights off of a light-colored surface behind you. If neither of those is enough, you can also use a "ring light" or "softbox" (no particular recommendations as I've never tried one).

2. Eliminate light sources in the camera's field of view. Compared to the human eye, cameras have lower "dynamic range" (the ability to faithfully capture variation in brightness)—which is why, for instance, your phone can't take good pictures of trees against the sky on a bright day. If there's a bright window behind you, or a light fixture in your camera's field of view, that part will look "blown out" and everything else will look dark by comparison.

Use your real background

Probably controversial. I'm speaking strictly from the point of view of immersiveness here—not e.g. expressing your individuality, making your coworkers laugh, or hiding the pile of laundry behind you. Those are all valid reasons to want to use backgrounds! Just be aware that you're sacrificing immersiveness when you do.

Why? Zoom's background detection software is not very accurate, and it'll periodically delete parts of your hair/body, make the background show through your eyeballs, etc. Plus, it's really bad at detecting boundaries so some of the real background will show through your hair.

If you have a decent camera (as below), and a space of your own (see get away from other people above), you'll look less distracting and more real if you don't use a fake

background.

Don’t bother with webcams

It’s probably obvious that laptop webcams suck. Even if they’re not inexplicably positioned so that they look up your nose, they will be grainy, blurry and have a tiny dynamic range.

Maybe less obviously, external webcams aren’t that much better. This surprised me, since a high-end webcam like the Logitech Brio costs 2/3 as much as a used interchangeable-lens camera and has *one job*. (To be clear, the Brio looks a lot better than most other webcams—just still a lot worse than a real camera.) For instance, I bought a Logitech C920, Wirecutter’s pick for “best webcam,” but it wasn’t obviously better than my six-year-old iMac webcam, mostly due to *really* questionable exposure / color balance settings.[3]

Left: 2014 iMac camera; right: Logitech C920. I am pretty white, but I'm not THAT white.

Use your smartphone…

(I haven’t used a smartphone as a webcam extensively because I jumped straight to using a full camera, so these are weakly held. I’ll update as I learn more.)

I expected webcams to be better than smartphones, because they wouldn’t have much reason to exist otherwise. But it turns out I was wrong: webcams do not, in fact,

have much reason to exist. Smartphones beat them on sensor size and quality of materials. For example, the iPhone 11 camera costs $73.50 in materials, and has a 1/2.55” sensor, while the Logitech C920 *retails* for $80 and has a 1/3” sensor.[4] So, if you want an external webcam, use a smartphone.

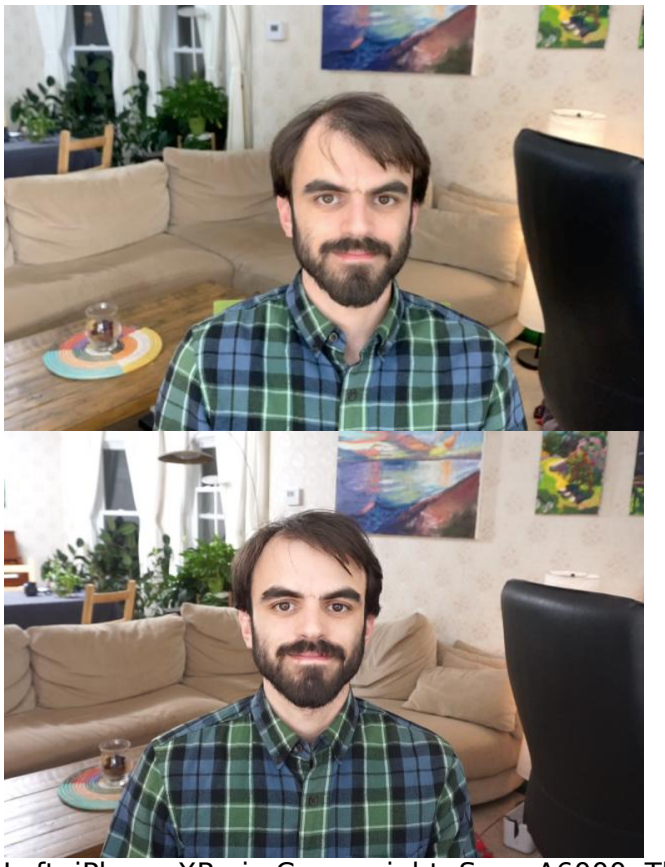
Left: Logitech C920; right: iPhone XR via Reincubate Camo (crop).

I briefly tried two apps for using your smartphone as a webcam, Camo and EpocCam. EpocCam is cheaper ($8 vs $40) but seemed somewhat buggier than Camo (both had occasional issues). Both of them have free trials that watermark your video, so suitable for testing but not actual calls.

The easiest way to mount a smartphone seems to be via a gooseneck holder like this. It seemed like it should be possible to find a much smaller device that attached the phone to my monitor, but the best I could find was this clip thing, which obscures part of the screen on laptops, doesn’t adjust in small enough increments to get the right field of view, and doesn’t work on external monitors with curved backs like my iMac. Let me know if you have a better suggestion.

…or a real camera

Even the lowest-end “real” cameras will trounce most smartphones on image quality. You’ll get a much sharper image, with a pleasingly blurred background to subtly draw attention toward your face and away from your piles of dirty laundry.



Left: iPhone XR via Camo; right: Sony A6000. This one makes more sense at full resolution; the Sony has a bit worse dynamic range, but the image is **much** sharper.

This is probably the most noticeable improvement in the list: I started using a camera when I gave a virtual conference talk at !!con and got compliments from ~20 different coworkers and conference speakers. (Being this noticeable is not necessarily an advantage, but I figured my coworkers already mostly knew that I was extremely vain, and so there was no point in trying to hide it.)

Unfortunately, these cameras are also… pretty user-unfriendly and can take a bit of work to set up. Some non-obvious tips if you decide to replicate my setup (Sony A6000 + Elgato CamLink 4k):

- Get a non-power-zoom lens, otherwise your zoom setting will get reset every time you turn the camera off and on again.
- If you have a Sony A6000, make sure you turn the top dial to “video” mode. Otherwise the HDMI output will be lower quality and the continuous focus won’t work right.
- Use the widest aperture possible to minimize graininess and get a nice blurry background effect.

A note on camera buying

I'm not sure which camera model is best now that so many different manufacturers have webcam drivers. Before the webcam drivers came out, the Sony A6000 was the default recommendation for a camera for streaming, and it's also generally well-recommended as an entry-level mirrorless camera; but if you have a Mac it currently requires a capture card while a Canon/Fujifilm wouldn't. Then again, Sony is promising to release their Mac webcam driver in "Autumn 2020." Basically, do your own research.

It's best to buy from a reputable used camera dealer (e.g. Keh, B&H, Lens Authority) since random resellers on Amazon / etc. might not be good about checking that the camera works well.

For completeness, here's what I bought though I don't particularly endorse it:

- Sony A6000
- 16-50mm ƒ/3.5–5.6 PZ OSS lens (as mentioned above, I'd recommend the non-PZ 18–55mm ƒ/3.5-5.6 OSS instead)
- Elgato Camlink 4k (may be out of stock; cheaper options available from no-name vendors)
- HDMI to Micro HDMI cable (took me a surprisingly long time to figure out what type of HDMI port the A6000 has!)
- Random off-brand USB dummy battery (sometimes gives "battery depleted" errors so maybe not the best idea)

Conclusion

The main thing I learned from this adventure is that, much like wireless, video calls still basically don't work. Every piece of equipment has tricky subtleties that make it super hard to select the right gear and use it correctly. Software does incredibly stupid things that you'll never notice unless you know what to look for. You never hear your own audio, so if it sucks, you'll never know. Many of the problems you do notice will be near-impossible to debug because there are no good diagnostics.

If there was a $2000 device that eliminated the need for this post to exist, it would be an automatic purchase for my employer and probably every other distributed company. But there isn't, so here we are.

Thanks to Dan Luu, Sasha Illarionov, David Coletta, Alexey Guzey, Lincoln Quirk, Eve Bigaj and Jessica Gambirasi for commenting on a draft of this post.

---

1. This means the BoomPro requires headphones with a detachable cable. If you're wedded to a different pair, the Antlion ModMic is a more expensive option that works with any headphones, at the cost of a second cable. There's also a wireless version, although you shouldn't use wireless audio equipment for video calls due to latency and quality concerns. ↩

2. It is easier to put a dynamic microphone right next to your mouth, which is sort of like rejecting noise, but not relevant if you want to place your mic outside the frame of a video. Dynamic mics also have less high-frequency and "transient" response, which can make some types of room noise less obtrusive. ↩

3. The C920 does allow you to tune these settings, but it took me about 2 hours to figure out how, and I ended up having to buy a crappy third-party app. Manually

tuning the settings would also require you to change them whenever your lighting changes throughout the day, which is pretty annoying. ↩

4. Obviously, if you have a lower-end smartphone, the C920 might have nicer materials than your smartphone. But it seems like there's a very small region of tradeoff-space where "buy a webcam" is a good idea relative to "buy a nicer smartphone" or "buy an entry-level camera." ↩

Some links in this post are affiliate links; proceeds go to GiveWell.

How To Fermi Model

[Note from Eli in 2020: I wrote this document in 2016, in conjunction with two workshops that I helped Oliver Habryka run. If I were to try and write a similar document today, it would likely be substantially different in form and style. For instance, reading this in 2020, I'm not very compelled by some of the argumentation that I used to justify this technique, and I think I could have been clearer about some of the steps.

Nevertheless, I think this is some useful content. I'm not going to take the time to write a new version of this document, so it seems better to share it, as is, instead of sitting on it.]

Oliver Habryka provided the seed material and did most of the development work on this technique. He gets upwards of 90% of the credit, even though I (Eli) wrote this document. Thanks Oli!

Introduction:

Rationale for Fermi Modeling:

Making good decisions depends on having a good understanding of the world: the better one's understanding the better one's decisions can be. Model-building procedures allow us to iteratively refine that understanding.

Using any model-building procedure at all is a large step up from using no procedure at all, but some procedures are superior to others. If possible, we would want to use techniques that rely on verified principles and are based on what we know about how the mind works. So, what insights can be gleaned from the academic social and cognitive sciences that is relevant to model-building?

First, Cognitive psychology has shown, many times over, that very simple algorithmic decision rules frequently have just as much predictive power, and even outperform, human expert judgment. Deep, specific models that take into account many details specific to the situation (inside views) are prone to overfitting, and are often inaccurate. Decision rules combat biases like the Halo effect and consequently tend to produce better results.

For instance, a very simple equation to predict the probability that a marriage will last is:

Frequency of lovemaking / Frequency of fights

(Where a higher number represents a more stable marriage. Example taken from Thinking Fast and Slow ch. 21)

This assessment measure is intuitive and uncomplicated, and it predicts length of marriage about as well as any other method, including expert evaluation by experienced couples counselors. Most of the relevant information is encapsulated in just those two variables: more detailed analysis is swamped by statistical variance and tends to make one overconfident. And the algorithm has the additional distinct advantage of being cheap and easy to deploy: simply plug in the variables and see what comes out.

The upshot is simple numeric algorithms are powerful.

Second, is the study of forecasting. One of the most significant takeaways from Philip Tetlock's Expert Political Judgment project is that "foxes" (who use and integrate multiple methods of evaluation, models, and perspectives) fare better than "hedgehogs" (who use a single overriding model or methodology that they know very deeply).

This poses a practical problem however. There are dozens of known psychological phenomena (anchoring, priming, confirmation bias, framing effects, attentional bias) that make it cognitively difficult to think *beyond* one's first idea. Once one has developed a model or a solution, it tends to be "sticky", coloring and constraining further thinking. Even if you want to generate new models, it's hard *not* to anchor on the one you had in front of you a moment ago. As Kahneman colorfully puts it, "What you see is all there is," or so it seems.

Given this, we want decision processes and planning protocols that are **1)** algorithmic, using simple equations or scoring rules with variables that are easy to assess **2)** foxy, in that they incorporate many models instead of one, and **3)** help mitigate the psychological biases that make this difficult.

Fermi Modeling is an attempt at a model building-procedure that caters to these constraints. It is quantitative, Foxy, and designed to compensate, at least somewhat, for our native biases.

In contrast to other methods of theorizing and model building, Fermi Modeling is less about going deep and more about going broad. Instead of spending a lot of time building one, very detailed, sophisticated, and precise model, the emphasis is on building *many* models rapidly.

Overview:

Fermi Modeling is a brainstorming and theorizing technique that encourages you to flip between multiple frames and perspectives, primarily by *moving up and down levels of abstraction.*

In broad strokes, the mental process of Fermi Modeling looks like this:

You start, in **step 1 (red)**, by moving *up* a level of abstraction, by considering reference classes or categories into which your object or problem of interest falls.

Then, in **step 2 (green)**, you move *down* a level of abstraction to generate models applicable to each reference class (with no regard for the original question.)

Then in **step 3 (blue)**, you apply the models generated in step 2 and 3 to the original question.

The point is to get you to consider questions that you wouldn't naively ask.

For instance, suppose I'm considering the question, "how do I determine which people I should spend time assisting, teaching, or otherwise, making better?"

This question brings to mind certain reference frames and criteria. I can think of people in this context as independent agents to implement my goals, which brings to mind consideration such as value alignment, power in the world, and discretion. I can think of people as teammates, which indicates other important factors such as personal compatibility with me and the complementary-ness of our skill sets. I can think of people as trade partners, which would lead me to consider what value they can give me.

Furthermore, I can change my focus from the object, "people", to the verb. I could rephrase the question as "who do I invest in?", which gives me the reference frame of "investment". This immediately brings to mind a whole set of models and formulas: compound interest and risk assessment. This yields considerations such as the principal investment, time to pay off, and probability of pay off.

These finance-flavored factors are obviously relevant to the question of "whose growth I should nurture?", but *I would not have considered them by default*. They don't come to mind when just thinking about "people", only when thinking about "investments"

Fermi Modeling is a process designed to generate ideas that don't come to mind by default, and to facilitate rapid consideration of many "angles of attack" on a problem, when creating models and evaluation criteria.

Method

A note on time allocation: one of the advantages of this method is that it pays off quickly, but can still generate large value with large time investment.

You can Fermi Model for 15 minutes and get rapid useful results, on a quick question, or you can do it for several hours, or even block out a whole day, to consider one particularly important decision. How much time you spend on each step is flexible and subject to personal preference. Just make sure you have enough time to consider at least three reference frames, and aggregate, at the end.

Step 0.

Ask a "how" or "what" question. In particular, look for questions that have a sense of gradient or variation: what causes a thing to be better, easier, bigger, or more impactful. What is a variable that you are trying to maximize or minimize?

Your question should impinge on your actions somehow. The answer to this question will inform some decision about how to act in the world.

Alternatively, this could be a question of general assessment: determining the overall "quality" or "goodness" of some object of interest, be it an organization, a process, a technology, etc.

Prompts	**Examples**

• Try to find an optimization problem. • What determines the quality of X? • How can I do X best? • What determines how much of Y the system that I want to understand produces? • What determines the probability that my system produces Y?	• What is the most effective way to make money? • What should the program for EA Global be? • How can I make more friends? • How can I get more work done? • How can I learn math faster? • How do I run the best workshop? • How do I build political influence? • How can I find a good boyfriend/girlfriend? • How can I find a co-founder for my company? • What do I make of CFAR?

Step 0.5.

Once you have written the initial question, rephrase the question in multiple ways. Try to ask the same question, or a nearby question, using different terminology. (This can involve small refactorings of the goal.) Doing this can introduce a little “conceptual jitter”, that can sometimes yield fruitful distinctions.

Examples
• What is the most effective way to make money? • What determines a person’s income? • How does one maximize earning power? • What should the program for EA Global be? • What is the best version of EA Global? • What should I have participants do at EA global? • How can I make more friends? • What makes people want to be friends with other people? • How do relationships form? • How can I get more work done? • What makes stuff get done faster? • What contributes to my losing time? • How can I learn math faster? • What’s the most efficient way to learn academic subjects? • What’s holding me back from knowing math? • How do I run the best workshop? • What makes a workshop good? • How do I build political influence? • How do I get large groups to do stuff?

Step 1: Abstract

Generate reference frames

1. Mark or underline all the key terms in each phrasing. To a first approximation, underlining all the nouns and all the verbs works. [If you are doing Fermi modeling as an evaluation procedure, you can skip this part].
2. For each of the marked terms, list reference classes for which the term is an example. You want to move up a level of abstraction, considering all the categories into which the term fits. You want to generate between 15 and 50 reference frames (of which you might use 4 to 6).

Prompts	**Examples**
• What *is* x? • What is x an instance/example of? • "Everything is a case-study" • What different reference classes would scientists from different fields put this in?	**EAG** • Conference • Social Gathering • Informational Message • Educational Content • Bunch of monkeys together **How can I make more friends?** • Relationships • Mammals • Search procedures • Partners for playing • Emotional support

Step 2: Model

Rapid model building on each of the frames:

Take one of the reference frames that you generated in the last step and Fermi model on it.

We recommend, if you are doing this for the first time, that you start with a frame other than the one you think is most useful, interesting, or relevant to your problem. Often, people pick one frame and build one model and feel like they're done. After all, the "correct" model is right in front of them; why would they bother constructing another, inferior model? Starting with a less-than-your-favorite frame encourages you to build more than one model.

Step 2.1

Identify first order factors. Consider what variables would determine the "quality" or quantity of things in the reference class. This often takes the form of asking "what makes an X good?". A thing can be more or less X or more or less of a good X.

Note that we are only looking at *first-order* factors. The models that we are generating are intended to be quick and rough. There will, for most categories, be many, many factors that exert a small influence on the overall outcome. We are only looking for as many factors as will have a sufficient effect as to influence the order of magnitude of the outcome. What factors explain most of the variance?

Prompts	**Examples**

• What makes things in this reference class good? • How do things in this reference class work in general?	**Social Gathering** • Number of People • “Quality” of average person • Number of new connections **Search Procedure** • Pool available to search • Accuracy of filtering procedure • Speed of filtering procedure

Step 2.2

Use simple mathematical operations (*, /, +, -, average, min, max, squared) to describe the relationships between first order factors. Write a function that describes how changes in the inputs change the output.

If you don’t know where to start, simply multiply your first order factors together, and then check to see if the resulting model makes sense as a first approximation. If it doesn’t, tinker with it a little by adjusting or adding terms.

Examples
Social Gathering = $Quality\_Of\_Connections * \frac{\#People}{\#N\_Connections}$ **Search Procedure =** $Accuracy * Pool\_Size * Speed$

You can quickly check your models by looking for 0s. What happens when any given factor is set to 0 or to arbitrarily large? Does the result make sense? This can inform your

expressions.

If you aren't familiar with the notion, try drawing a graph, that holds all but one of the inputs constant. You can use the graph to reverse engineer the mathematical notion if you want.

You *can* do more work on these models, primarily by decomposing your first order factors into more basic components. But this is usually misguided. These models are rough, based only on simple intuitions, making them more detailed at this point makes them more precise than their general accuracy warrants. In most cases, it only makes sense to add detail after we have had opportunity to test our models empirically.

Some of the models you generate may be cached, standard models from one domain or another. For instance, there are known, simple equations for compound interest. This is perfectly fine, and in fact, is quite good. Those models come pre-vetted and verified.

The process of generating a single model should not take more than 6 minutes in most cases, as a beginner.

Step 2.3

Build as many such models in a given reference class as you'd like. Two or three is usually sufficient.

Some people find this step somewhat difficult. There are a couple of "tricks" that you can apply to reframe and generate more models.

1. Do a resolve cycle: set a timer for five minutes (or two minutes) and come up with as many models as you can before it rings. Get into the mindset of "I *have* to do it."
2. Reverse the question: If you've been considering what makes a thing good, then ask what would make it bad. ("If you can't optimize, pessimize.")
3. Consider how scientists or academics from various disciplines would approach this problem? How does a historian look at this? A mechanical engineer? A biologist? An economist?
4. Consider an alternative way to parse the world. A good way to do this is to forbid the use of the factors you used in your first model. How *else* could you make sense of this situation?
5. Ask the person next to you. It's often surprising how different the models that another person will generate are.

Some notes on models:

- Consider all the costs. They usually go in the denominator.
- Time is often an input, but it usually has a negligible effect on the output
- Econ 101: Remember to account for opportunity cost (subtracted from the main body of the expression). Is the next best option much worse than this one?
- Probabilities are expressed as values between 0 and 1. It may be helpful to consider what distribution a value is drawn from.
- You can put in constant multiples.

Step 2.4 (Optional)

Generate examples to spur models: think of hypothetical, or better yet, actual examples of the reference class. Consider how they fare in terms of each of your main factors. What makes each example good? How could you tweak them to make them better or worse?

(optional) Test: generate counterexamples

Step 2.5:

Repeat step 2. Build more models on each reference frame in turn.

Really do this! I've sometimes seen people (and am personally prone to) become quite anchored on or attached to their first model they / I build, since it seems obviously correct. Most of the value of this method comes building many models.

Step 3: Aggregate

Once you've generated some models, you know want to go back and evaluate your original question. Some of the models you built won't be relevant to the original question, but you should be sure to consider each one before dismissing it. Remember, the whole point is to generate considerations that wouldn't have occurred to you by default.

There are lots of ways to do this.

Looking at each of your models / functions, compare to the situation you're considering (your workshop, for instance). Estimate values for each of the terms in each model. Do the majority of the models recommend one type of action?

You can use the models you've generated abstractly in the more concrete context. What happens when you adjust the factors on your actual plan?

Try and come up with a plan that scores perfectly on each model. See how much overlap there is between those plans. Can you goal-factor and get most of the benefit?

The quantitative nature of your models means that you can also take your subjective analysis out of it. Set up a scoring system, that takes all the inputs for a plan and returns an aggregated score.

Closing thoughts

Advantages:

Since, for most of the process, you're not focusing on the original question at all, but rather building models only in the context of the reference frame, you avoid, somewhat, the "stickiness" of your initial models. You're less likely to get stuck thinking that the way you modeled the problem is the "correct" model (and then being resistant to seeing other perspectives, due to a whole slew of biases), since you shouldn't be thinking about the original problem at all.

As mentioned above, this method scales easily with more time invested. It's also parallelizable. it's easy to have multiple people on a team all Fermi modeling on the same topic, and each of them is likely to come up with novel, useful insights. I'd recommend that

each person do step 1 independently, have everyone share frames, then have each person do 2 and 3 on a subset of the frames.

Disadvantages:

This process is designed to produce rough heuristics rapidly. Sometimes a deep understanding of the specific situation is necessary.

Further reading:

Clinical vs. Statistical Prediction: A Theoretical Analysis and a Review of the Evidence by Paul Meehl: a slim but dense volume, this a classic of the the field that first made the case for numeric algorithms over expert judgment.

Chapter 21 of *Thinking Fast and Slow* by Daniel Kahneman is a popular overview of how simple algorithmic decision rules frequently outperform expert judgment.

Expert Political Judgment: How Good is it? How can we Know? by Philip Tetlock is a compendium on the research project that gave rise to the Fox vs. Hedgehog distinction.

Biases that are relevant

Abstraction vs. Analogy by Robin Hanson is a good, brief example of considering an object in terms of several various reference frames.

Cluster Thinking vs Sequence Thinking by Holden Karnofsky is an essay on making decisions on the basis of weighing and integrating many models.

How to Measure Anything: How to Find the Value of Intangibles in Business

by Douglas W. Hubbard is an excellent primer on applying quantitative measurements to qualitative domains.

Artificial Intelligence: A Modern Approach (4th edition) on the Alignment Problem

This is a linkpost for http://aima.cs.berkeley.edu/

Previously: AGI and Friendly AI in the dominant AI textbook (2011), Stuart Russell: AI value alignment problem must be an "intrinsic part" of the field's mainstream agenda (2014)

The 4th edition of *Artificial Intelligence: A Modern Approach* came out this year. While the 3rd edition published in 2009 mentions the Singularity and existential risk, it's notable how much the 4th edition gives the alignment problem front-and-center attention as part of the introductory material (speaking in the authorial voice, not just "I.J. Good (1965) says this, Yudkowsky (2008) says that, Omohundro (2008) says this" as part of a survey of what various scholars have said). Two excerpts—

1.1.5 Beneficial machines

> The standard model has been a useful guide for AI research since its inception, but it is probably not the right model in the long run. The reason is that the standard model assumes that we will supply a fully specified objective to the machine.
>
> For an artificially defined task such as chess or shortest-path computation, the task comes with an objective built in—so the standard model is applicable. As we move into the real world, however, it becomes more and more difficult to specify the objective completely and correctly. For example, in designing a self-driving car, one might think that the objective is to reach the destination safely. But driving along any road incurs a risk of injury due to other errant drivers, equipment failure, and so on; thus, a strict goal of safety requires staying in the garage. There is a tradeoff between making progress towards the destination and incurring a risk of injury. How should this tradeoff be made? Furthermore, to what extent can we allow the car to take actions that would annoy other drivers? How much should the car moderate its acceleration, steering, and braking to avoid shaking up the passenger? These kinds of questions are difficult to answer a priori. They are particularly problematic in the general area of human–robot interaction, of which the self-driving car is one example.
>
> The problem of achieving agreement between our true preferences and the objective we put into the machine is called the **value alignment problem**: the values or objectives put into the machine must be aligned with those of the human. If we are developing an AI system in the lab or in a simulator—as has been the case for most of the field's history—there is an easy fix for an incorrectly specified objective: reset the system, fix the objective, and try again. As the field progresses towards increasingly capable intelligent systems that are deployed in the real world, this approach is no longer viable. A system deployed with an incorrect objective will have negative consequences. Moreover, the more intelligent the system, the more negative the consequences.

> Returing to the apparently unproblematic example of chess consider what happens if the machine is intelligent enough to reason and act beyond the confines of the chessboard. In that case, it might attempt to increase its chances of winning by such ruses as hypnotizing or blackmailing its opponent or bribing the audience to make rustling noises during its opponents thinking time.[3] It might also attempt to hijack additional computing power for itself. *These behaviors are not* "*unintelligent*" *or* "*insane*"*; they are a logical consequence of defining winning as the* sole *objective for the machine*.
>
> It is impossible to anticipate all the ways in which a machine persuing a fixed objective might misbehave. There is good reason, then, to think that the standard model is inadequate. We don't want machines that are intelligent in the sense of pursuing *their* objectives; we want them to pursue *our* objectives. If we cannot transfer those objectives perfectly to the machine, tghen we need a new formulation—one in which the machine is pursuing our objectives, but is necessarily *uncertain* as to what they are. When a machine knows that it doesn't know the complete objective, it has an incentive to act cautiously, to ask permission, to learn more about our preferences through observation, and to defer to human control. Ultimately, we want agents that are **provably beneficial** to humans. We will return to this topic in Section 1.5.

And in Section 1.5, "Risks and Benefits of AI"—

> At around the same time, concerns were raised that creating **artificial superintelligence** or **ASI**—intelligence that far surpasses human ability—might be a bad idea (Yudkowsky, 2008; Omohundro 2008). Turing (1996) himself made the same point in a lecture given in Manchester in 1951, drawing on earlier ideas from Samuel Butler (1863):[15]
>
> > It seems probably that once the machine thinking method had started, it would not take long to outstrip our feeble powers. ... At some stage therefore we should have to expect the machines to take control, in the way that is mentioned in Samuel Butler's *Erewhon*.
>
> These concerns have only become more widespread with recent advances in deep learning, the publication of books such as *Superintelligence* by Nick Bostrom (2014), and public pronouncements from Stephen Hawking, Bill Gates, Martin Rees, and Elon Musk.
>
> Experiencing a general sense of unease with the idea of creating superintelligent machines is only natural. We might call this the **gorilla problem**: about seven million year ago, a now-extinct primate evolved, with one branch leading to gorillas and one to humans. Today, the gorillas are not too happy about the human branch; they have essentially no control over their future. If this is the result of success in creating superhuman AI—that humans cede control over their future—then perhaps we should stop work on AI and, as a corollary, give up the benefits it might bring. This is the essence of Turing's warning: it is not obvious that we can control machines that are more intelligent than us.
>
> If superhuman AI were a black box that arrived from outer space, then indeed it would be wise to exercise caution in opening the box. But it is not: *we* design the AI systems, so if they do end up "taking control," as Turing suggests, it would be the result of a design failure.

To avoid such an outcome, we need to understand the source of potential failure. Norbert Weiner (1960), who was motivated to consider the long-term future of AI after seeing Arthur Samuel's checker-playing program learn to beat its creator, had this to say:

> If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively ... we had better be quite sure that the purpose put into the machine is the purpose which we really desire.

Many cultures have myths of humans who ask gods, genies, magicians, or devils for something. Invariably, in these stories, they get what they literally ask for, and then regret it. The third wish, if there is one, is to undo the first two. We will call this the **King Midas problem**: Midas, a legendary King in Greek mythology, asked that everything he touched should turn to gold, but then regretted it after touching his food, drink, and family members.[16]

We touched on this issue in Section 1.1.5, where we pointed out the need for a significant modification to the standard model of putting fixed objectives into the machine. The solution to Weiner's predicament is not to have a definite "purpose put into the machine" at all. Instead, we want machines that strive to achieve human objectives but know that they don't know for certain exactly what those objectives are.

It is perhaps unfortunate that almost all AI research to date has been carried out within the standard model, which means that almost all of the technical material in this edition reflects that intellectual framework. There are, however, some early results within the new framework. In Chapter 16, we show that a machine has a positive incentive to allow itself to be switched off if and only if it is uncertain about the human objective. In Chapter 18, we formulate and study **assistance games**, which describe mathematically the situation in which a human has an objective and a machine tries to achieve it, but is initially uncertain about what it is. In Chapter 22, we explain the methods of **inverse reinforcement learning** that allow machines to learn more about human preferences from observations of the choices that humans make. In Chapter 27, we explore two of the principal difficulties: first, that our choices depend on our preferences through a very complex cognitive architecture that is hard to invert; and, second, that we humans may not have consistent preferences in the first place—either individually or as a group—so it may not be clear what AI systems *should* be doing for us.

"Win First" vs "Chill First"

Around a month ago, I found an *incredibly* insightful quote deep in the Reddit comments about a particular basketball player who had recently changed teams.

> People who try hard to win first and foremost make it uncomfortable when people are trying to just have a good time and do well. And this is aside from whether they're assholes or not, unscrupulous or not. **It's about win first vs chill first.**
>
> At some point, there's always a conflict between the two types. Someone has to chill, someone has to turn up the intensity, or someone has to leave.

- **Source**, minor formatting cleanup and emphasis added

Now, I'm not a particularly avid follower of sports. But this was a rather unusually fascinating case.

To simplify a very long story, there's a professional basketball team — the Philadelphia 76ers — that on paper have had *a lot* of really good players the last few years, yet have consistently underperformed expectations.

Last year, they traded for a player who is known as being super-crazy-hardcore-intense. That player was Jimmy Butler, who was an unheralded quite low draft pick (the 30th player chosen his year, meaning almost every team passed on him at least once) who worked very, very hard to turn himself into a star.

He wasn't one of those players who was really good right when joined the League — he didn't start in pro basketball until he was 22 years old, and wasn't really good until he was 25.

His career stats are here; you don't need to know much about basketball to see the trend of going from scoring 2.6 points per game your first year in the league, to 8.6 your second, to 13.1 your third year, to 20.0 your fourth year in the league is (1) someone who was not-at-all "anointed" or had an easy path for himself, and (2) showed really incredible year-over-year improvements.

Eventually, Jimmy became a consistent All-Star.

Last year, he joined the Philadelphia 76ers.

And it didn't go very well.

Although much of the story is secondhand and hearsay, apparently Jimmy Butler didn't get along with everyone else on the 76ers. During a film session to prepare for an upcoming opponent, there were reports that other players were sleeping or goofing around and Jimmy shouted at them.

Jimmy would yell at people who weren't training hard at practice, get under his teammate's skins, etc. If anyone wanted to relax and refused to go full-out competitive in pursuit of being the best individual player they could be, the best teammate they could be, and giving their utmost towards every game — Jimmy wasn't having it.

And the 76ers, by all accounts, had something of a "chill first" culture, despite having — on paper — really really good players. Anyway, much of this is hearsay but some of it isn't — obvious examples being when a coach publicly instructed one player who refused to follow instructions, a star player being noticeably out of shape and heavyset and suffering at the end of games, things like that.

Of course, these are still some of the finest athletes in the world — but the 76ers didn't have whatever that fanatic intensity that Michael Jordan was famous for, with all its advantages towards winning along with all its undeniable nasty side effects on stress and toxicity and lack of amicability.

Well, at the end of last season, Jimmy Butler's contract expired and he left the Philadelphia 76ers.

He went from Philadelphia to a team that missed the playoffs entirely that year, the Miami Heat, saying he went just because the culture there was intense and he felt his intensity would be appreciated there.

Jimmy Butler was roundly mocked for his decision. On paper, the 76ers looked like one of the best teams in basketball and the Miami Heat looked like a very subpar team.

Well, one year later, the 76ers just underperformed and were eliminated early again this year — and the team Jimmy Butler joined, Miami (which missed the playoffs last year)... is now heading to the NBA Finals as of tonight.

There's no doubt in my mind that there's a lot of people more *content* than Jimmy Butler, more *amicable* than Jimmy Butler, way more fun to *chill out with* than Jimmy Butler... there's not a single doubt in my mind that there are *very many downsides* to that fanatic junkyard dog mentality, that it's *incredibly stressful*, often *painful*, risks destroying relationships and discordancy rather than the more guaranteed affability and amicability of "chill first, don't worry about it"...

... and yet, y'know, I saved this comment over a month ago, before any of this could be truly foreseen, since it seemed to sum up a point rather elegantly:

> People who try hard to win first and foremost make it uncomfortable when people are trying to just have a good time and do well. And this is aside from whether they're assholes or not, unscrupulous or not. **It's about win first vs chill first.**
>
> At some point, there's always a conflict between the two types. Someone has to chill, someone has to turn up the intensity, or someone has to leave.

The new Editor

Look at this glorious table

Celebrations! The new editor is finally here!

Starting from today, all desktop users will by-default use the new visual editor that we've been testing for a while. While the primary goal of this is to have a better foundation on which to build future editor features, here are a number of things you can do starting from today:

- Insert tables! A heavily requested feature.
- Copy-paste LaTeX without everything breaking!
- Nest bullet lists and other block elements in blockquotes! (still no nested blockquotes though, though if enough people want that, it would be easy to change)
- Image Uploads! (Just drag-and-drop images into the editor, and things should work out naturally. You can also copy-paste, though beware that copy-pasting from other websites means we link to the copies of those images from other websites, and won't reupload them.)
- Much less jankyness and brokenness!

Let us know what you think about the new editor. We've been testing it for a while and have been pretty happy with it (and users who had opted into beta features also had predominantly positive feedback). You can also use the old editor if you run into any problems by checking the "Restore the previous WYSIWYG editor" checkbox in your user settings.

Stop pressing the Try Harder button

This is a linkpost for https://www.neelnanda.io/blog/mini-blog-post-6-stop-pressing-the-try-harder-button

(This is a post from a daily blogging experiment I did at neelnanda.io, which I thought might also fit the tastes of LessWrong. This is very much in the spirit of Trying to Try)

I recently had a productivity coaching session, and at the end we agreed on a few actions points that I'd do by the next session. But, come the next session, these had completely slipped my mind. These suggestions were good ideas, and I had no issue with implementing them, the problem was just that they completely slipped my mind! (We then spent the second session debugging my ability to actually follow action points, and this was pretty successful!)

I think the error I made there is a *really* common one when planning, and one I observe often in myself and others. Often I'll hear a cool book recommendation, offer to meet up with someone some time, hear about a new productivity technique, notice an example sheet deadline looming. But I consistently fail to action upon this. So this post is about what *exactly* went wrong, and the main solution I've found to this problem!

Planning, as I define it, is about **ensuring that the future goes the way I currently want it to**. And the error I made was that, implicitly, I was *trying* to make the future go the way I currently wanted it to. That by committing to do things, and wanting to them, and just applying effort, things would happen. And the end result of this was that I totally forgot about it. Or sometimes, that I vaguely remembered the commitment or idea, and felt some guilt about it, but it never felt urgent or my highest priority. And every time I thought about the task, I resolved to Try Harder, and felt a stronger sense of motivation, but this never translated into action. I call this error **Pressing the Try Harder button**, and it's characterised by feelings of guilt, obligation, motivation and optimism.

This is a classic case of failing to Be Deliberate. It feels *good* to try hard at something, it feels important and virtuous, and it's easy to think that trying hard is what matters. But ultimately, trying hard is just a means to an end - my goal is to ensure that the task happens. If I can get it done in half the effort, or get somebody else to do it, that's awesome! Because my true goal is the *result*. And pressing the Try Harder button is *not* an effective way of achieving the goal - you can tell, because it so often fails!

A good litmus test for whether you're pressing the Try Harder button: Imagine it's 2 weeks from now, and you never got round to doing the task. Are you surprised that this happened? Often my intuitions are well-calibrated when I phrase the question like this - on some level I *know* that I procrastinate on things and forget them all the time.

But just noticing yourself pushing the Try Harder button isn't enough - you need to do something stronger to change this. You need to *find strategies that actually work*. This is pretty personal, and much easier said than done! But it can be done. Look for common trends, strategies that have worked for you in the past, and things that you can repurpose.

Strategies that work for me:

- Scaffold systems - meta-systems that I check regularly
    - Calendars
    - Trello (my to-do list) - especially future reminders that result in an email
    - Getting friends to check in with me
- Do it *now*, not later. Set a 5 minute timer, and see if you can finish the task. Or at least make a start!
    - You can get a surprising amount done in 5 minutes! (Say, writing a third of a blog post)
    - Often the bottleneck is that getting *started* takes a bit of energy. Doing something for 5 minutes can take much less energy, and I find that timers help me focus a lot
- Make things *concrete* - often tasks feel overwhelming and fuzzy, so you put them off. Can you break it down into a concrete next action? Something that can be done in under 5 minutes?
- Schedule time for it - often the bottleneck is that it doesn't feel *urgent* - I care about the task getting done, but I always have something seemingly-higher-priority to do
    - This is terrible, because in the long-run I *always* have something that seems short-term higher priority, and I never make time for my long-term goals
    - So if I make time in my calendar for it, and make it feel *important* that I stick to these, that's valuable
    - I find focusmate.com valuable for carving out an hour for a specific task
- Add it to a queue
    - Ie a to-do list
    - I find Trello great for this - on Desktop I can add a new card from anywhere with CTRL+ALT+SPACE, it makes it really low friction
    - **Important:** It's not enough to just add it to a list - the other half of a good to-do list system is having a regular time to *process* the list!
        - Having an eg weekly routine for this is important - a routine doesn't involve decisions, it can just happen automatically. While if I just say "I'll make time for it", that's pushing the Try Harder button!
- Accountability
    - Message a friend saying that I commit to this task
    - Extreme: Give a friend some money, and tell them to only give it back when the task is done

These are just the strategies that work for me - I'd love to hear what works for others, and expect it vary a lot between people. The message I want you to take from this post is just to notice when you next push the Try Harder button. And ask yourself: "am I just being virtuous and trying? Or am I trying to *change what my future self actually does*?"

Some thoughts on criticism

Here are some somewhat unconnected unconfident thoughts on criticism that I've been thinking about recently.

A while ago, when I started having one-on-ones with people I was managing, I went into the meetings with a list of questions I was going to ask them. After the meetings, I'd look at my notes and realize that almost all the value of the meeting came from the part where I asked them what the worst parts of their current work situation were and what the biggest mistakes I was making as a manager were.

I started thinking that almost the whole point of meetings like that is to get your report to feel comfortable giving an honest answer to those questions when you ask them--everything else you talk about is just buttering them up.

I wish I knew how to make people I'm managing more enthusiastic about criticising me and telling me their insecurities. Maybe I could tell them to have a group chat with just them in it where they all had to name their biggest complaints about me? Maybe I should introduce them all to a former intern of mine who promised not to repeat anything they said who told them about all the mistakes I made while managing them, as an attempt to credibly signal actual interest?

A lot of the helpful criticism I've gotten over the last few years was from people who were being kind of unreasonable and unfair.

One simple example of this is that one time someone (who I'd engaged with for many hours) told me he didn't take my ideas seriously because I had blue hair. On the one hand, fuck that guy; on the other hand, it's pretty helpful that he told me that, and I'm grateful to him for telling me. You'd think that being on the internet would expose you to all the relatively uninformed impolite criticism you'd possibly need, but in my experience this isn't true.

Additionally I think that when people are annoyed at you, they look harder for mistakes you're making and they speak more frankly about them. So it's sometimes actually more likely that people will give you useful criticism if they get unreasonably annoyed at you first. This goes especially for people who know you and understand you well. (This is also a reason to think it's probably helpful to sometimes get drunk around people you like a lot but don't totally see eye to eye with. I haven't actually experimented with this.)

I think I've often been insufficiently gracious about receiving criticism in this kind of case, which seems pretty foolish of me. This is even more foolish because I've often behaved this way in contexts where I had more social power than the person who was criticizing me. I wish that I basically always responded to criticism from people I don't know extremely well by saying "that's interesting, thanks for telling me you think that, I'll think about it." I'm working on it.

[epistemic status: I'm a little worried about this kind of amateur psychology speculation]

I think one of the central things that's hard about criticism is that people often tie their identities to being good at various things, and it's hard to predict exactly which way they do this and so it's tricky to know what criticism will deeply hurt them. For example, I think people often have pretty core beliefs like "I'm not that good at X and Y, but at least I can hold onto the fact that I'm good at Z". Often, people like that will respond well to criticism about X and Y but not about Z. The problem is that it's kind of hard to guess which things are in which category for someone.

I think it's really really hard to be actually entirely open to criticism, and I don't know if it's even a good idea for most people to try to strive for it.

I think that if you tell people that it's extremely virtuous to be open to deep criticism, they sometimes just become really good at not listening to or understanding criticism (like described here https://www.lesswrong.com/posts/byewoxJiAfwE6zpep/reality-revealing-and-reality-masking-puzzles#Disorientation_patterns).

A lot of the time, one of my biggest bottlenecks is that I'm not feeling secure enough to be properly open to criticism. This means both that I can't properly criticise myself and that other people correctly conclude that they shouldn't criticise me (and these people don't even look as hard as they could for my weaknesses).

For example, this is true right now as I'm writing this. If I imagine getting an email from someone who I deeply admire where they'd written up their thoughts on the biggest mistakes I was making, I feel like I'd put off opening it, because I feel fragile enough that I'd worry that reading it would crush me and make me feel useless and depressed and unable to do the things I do. And when I go to a whiteboard and try to make lists of the most likely ways that I'm currently making big mistakes, I feel like I intuitively flinch away from looking directly at the question.

My guess is that this is true of most people most of the time.

I think that this is a major mechanism via which I'm less productive when I'm less happy--I'm less able to ask myself whether I'm really working on the most important problem right now.

Even in this state it's pretty useful to get criticism from people, because they manage to do a pretty good job of filtering the criticism to be not too core to who you are as a person.

I definitely wouldn't want anyone reading this to criticize me less as a result of reading this post.

One way of getting better criticism is to come up with a list of things you think you might be doing wrong, then ask specifically about them. This both credibly signals that you're actually interested in criticism, and also communicates that that topic isn't one of your weak points.

I think that it's probably generally more helpful to come up with a list of twenty mistakes you're most likely to be making, and then circulate an anonymous survey

where people check the ones they think you're indeed making, rather than to circulate an open ended criticism form.

I recently heard about someone who I've spent between 10 and 100 hours talking to doing something related to their career that looked to me and to many of my friends like a blunder. I don't think any of us told that person that we thought they'd fucked up. This was partially because it seemed like they'd already made the decision, and in my case it was because I had only heard about this indirectly and it felt a bit weird to reach out to someone to say that I'd heard they'd done something that I thought was dumb. It still feels a bit sad.

If you message me asking for it, I'll tell you if I think you're doing something that looks like it's plausibly a bad mistake with your career or life at the moment. I can only think of a few people where my answer would be yes.

I think that thinking of yourself as better than other people is, in some ways, helpful for being more pleasant to talk to. I've basically never heard anyone make this point directly before.

One context in which I'm often unpleasant is when someone's saying something I strongly disagree with in a way I dislike, and I lash out aggressively and unhelpfully. I think this is because I feel threatened--I think I intuitively feel like it's really important for the people in the conversation to see me win the argument, so that they think that I'm smart and right.

If I felt more secure and more superior to the people in the conversation, I think it would be easier to behave better, because I'd feel more like I was proposing some ideas and then seeing if the people I was talking to were interested in them, and then inasmuch as they weren't, I'd shrug and give up and quietly update against those people.

I have this attitude much less than a lot of the people I know. I think this makes them better than me at being pleasant.

However, I think that feeling more insecure makes it somewhat easier to connect with people, because it means that my heart is more on my sleeve and I can engage with their disagreements more wholeheartedly and openly, and this makes people more comfortable about having some kinds of conversations with me.

Probably there's a happy medium here that is better than my current attitude.

AGI safety from first principles: Superintelligence

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

In order to understand superintelligence, we should first characterise what we mean by intelligence. We can start with Legg's well-known definition, which identifies intelligence as the ability to do well on a broad range of cognitive tasks.[1] The key distinction I'll draw in this section is between agents that understand how to do well at many tasks because they have been specifically optimised for each task (which I'll call the task-based approach to AI), versus agents which can understand new tasks with little or no task-specific training, by generalising from previous experience (the generalisation-based approach).

Narrow and general intelligence

The task-based approach is analogous to how humans harnessed electricity: while electricity is a powerful and general technology, we still need to design specific ways to apply it to each task. Similarly, computers are powerful and flexible tools - but even though they can process arbitrarily many different inputs, detailed instructions for how to do that processing needs to be individually written to build each piece of software. Meanwhile our current reinforcement learning algorithms, although powerful, produce agents that are only able to perform well on specific tasks at which they have a lot of experience - Starcraft, DOTA, Go, and so on. In *Reframing Superintelligence*, Drexler argues that our current task-based approach will scale up to allow superhuman performance on a range of complex tasks (although I'm skeptical of this claim).

An example of the generalisation-based approach can be found in large language models like GPT-2 and GPT-3. GPT-2 was first trained on the task of predicting the next word in a corpus, and then achieved state of the art results on many other language tasks, without any task-specific fine-tuning! This was a clear change from previous approaches to natural language processing, which only scored well when trained to do specific tasks on specific datasets. Its successor, GPT-3, has displayed a range of even more impressive behaviour. I think this provides a good example of how an AI could develop cognitive skills (in this case, an understanding of the syntax and semantics of language) which generalise to a range of novel tasks. The field of meta-learning aims towards a similar goal.

We can also see the potential of the generalisation-based approach by looking at how humans developed. As a species, we were "trained" by evolution to have cognitive skills including rapid learning capabilities; sensory and motor processing; and social skills. As individuals, we were also "trained" during our childhoods to fine-tune those skills; to understand spoken and written language; and to possess detailed knowledge about modern society. However, the key point is that almost all of this evolutionary and childhood learning occurred on different tasks from the economically useful ones we perform as adults. We can perform well on the latter category only by reusing the cognitive skills and knowledge that we gained previously. In our case, we were fortunate that those cognitive skills were not too specific to tasks in the ancestral

environment, but were rather very *general* skills. In particular, the skill of abstraction allows us to extract common structure from different situations, which allows us to understand them much more efficiently than by learning about them one by one. Then our communication skills and theories of mind allow us to share our ideas. This is why humans can make great progress on the scale of years or decades, not just via evolutionary adaptation over many lifetimes.

I should note that I think of task-based and generalisation-based as parts of a spectrum rather than a binary classification, particularly because the way we choose how to divide up tasks can be quite arbitrary. For example, AlphaZero trained by playing against itself, but was tested by playing against humans, who use different strategies and playing styles. We could think of playing against these two types of opponents as two instances of a single task, or as two separate tasks where AlphaZero was able to generalise from the former task to the latter. But either way, the two cases are clearly very similar. By contrast, there are many economically important tasks which I expect AI systems to do well at primarily by generalising from their experience with very different tasks - meaning that those AIs will need to generalise much, much better than our current reinforcement learning systems can.

Let me be more precise about the tasks which I expect will require this new regime of generalisation. To the extent that we can separate the two approaches, it seems plausible to me that the task-based approach will get a long way in areas where we can gather a lot of data. For example, I'm confident that it will produce superhuman self-driving cars well before the generalisation-based approach does so. It may also allow us to automate most of the tasks involved even in very cognitively demanding professions like medicine, law, and mathematics, if we can gather the right training data. However, some jobs crucially depend on the ability to analyse and act on such a wide range of information that it'll be very difficult to train directly for high performance on them. Consider the tasks involved in a role like CEO: setting your company's strategic direction, choosing who to hire, writing speeches, and so on. Each of these tasks sensitively depends on the broader context of the company and the rest of the world. What industry is their company in? How big is it; where is it; what's its culture like? What's its relationship with competitors and governments? How will all of these factors change over the next few decades? These variables are so broad in scope, and rely on so many aspects of the world, that it seems virtually impossible to generate large amounts of training data via simulating them (like we do to train game-playing AIs). And the number of CEOs from whom we could gather empirical data is very small by the standards of reinforcement learning (which often requires billions of training steps even for much simpler tasks). I'm not saying that we'll never be able to exceed human performance on these tasks by training on them directly - maybe a herculean research and engineering effort, assisted by other task-based AIs, could do so. But I expect that well before such an effort becomes possible, we'll have built AIs using the generalisation-based approach which know how to perform well even on these broad tasks.

In the generalisation-based approach, the way to create superhuman CEOs is to use other data-rich tasks (which may be very different from the tasks we actually want an AI CEO to do) to train AIs to develop a range of useful cognitive skills. For example, we could train a reinforcement learning agent to follow instructions in a simulated world. Even if that simulation is very different from the real world, that agent may acquire the planning and learning capabilities required to quickly adapt to real-world tasks. Analogously, the human ancestral environment was also very different to the modern world, but we are still able to become good CEOs with little further training. And

roughly the same argument applies to people doing other highly impactful jobs, like paradigm-shaping scientists, entrepreneurs, or policymakers.

One potential obstacle to the generalisation-based approach succeeding is the possibility that specific features of the ancestral environment, or of human brains, were necessary for general intelligence to arise. For example, some have hypothesised that a social "arms race" was required to give us enough social intelligence to develop large-scale cultural transmission. However, most possibilities for such crucial features, including this one, could be recreated in artificial training environments and in artificial neural networks. Some features (such as quantum properties of neurons) would be very hard to simulate precisely, but the human brain operates under conditions that are too messy to make it plausible that our intelligence depends on effects at this scale. So it seems very likely to me that eventually we will be able to create AIs that can generalise well enough to produce human-level performance on a wide range of tasks, including abstract low-data tasks like running a company. Let's call these systems artificial general intelligences, or AGIs. Many AI researchers expect that we'll build AGI within this century; however, I won't explore arguments around the timing of AGI development, and the rest of this document doesn't depend on this question.

Paths to superintelligence

Bostrom defines a superintelligence as "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest". For the purposes of this report, I'll operationalise "greatly exceeding human performance" as doing better than all of humanity could if we coordinated globally (unaided by other advanced AI). I think it's difficult to deny that in principle it's possible to build individual generalisation-based AGIs which are superintelligent, since human brains are constrained by many factors which will be much less limiting for AIs. Perhaps the most striking is the vast difference between the speeds of neurons and transistors: the latter pass signals about four million times more quickly. Even if AGIs never exceed humans in any other way, a speedup this large would allow one to do as much thinking in minutes or hours as a human can in years or decades. Meanwhile our brain size is important in making humans more capable than most animals - but I don't see any reason why a neural network couldn't be several orders of magnitude larger than a human brain. And while evolution is a very capable designer in many ways, it hasn't had much time to select specifically for the skills that are most useful in our modern environment, such as linguistic competence and mathematical reasoning. So we should expect that there are low-hanging fruit for improving on human performance on the many tasks which rely on such skills.[2]

There are significant disagreements about how long it will take to transition from human-level AGI to superintelligence, which won't be a focus of this report, but which I'll explore briefly in the section on Control. In the remainder of this section I'll describe in qualitative terms how this transition might occur. By default, we should expect that it will be driven by the standard factors which influence progress in AI: more compute, better algorithms, and better training data. But I'll also discuss three factors whose contributions to increasing AI intelligence will become much greater as AIs become more intelligent: replication, cultural learning, and recursive improvement.

In terms of replication, AIs are much less constrained than humans: it's very easy to create a duplicate of an AI which has all the same skills and knowledge as the original. The cost of compute for doing so is likely to be many times smaller than the original

cost of training an AGI (since training usually involves running many copies of an AI much faster than they'd need to be run for real-world tasks). Duplication currently allows us to apply a single AI to many tasks, but not to expand the range of tasks which that AI can achieve. However, we should expect AGIs to be able to decompose difficult tasks into subtasks which can be tackled more easily, just as humans can. So duplicating such an AGI could give rise to a superintelligence composed not of a single AGI, but rather a large group of them (which, following Bostrom, I'll call a *collective AGI*), which can carry out significantly more complex tasks than the original can.[3] Because of the ease and usefulness of duplicating an AGI, I think that collective AGIs should be our default expectation for how superintelligence will be deployed.

The efficacy of a collective AGI might be limited by coordination problems between its members. However, most of the arguments given in the previous paragraphs are also reasons why individual AGIs will be able to surpass us at the skills required for coordination (such as language processing and theories of mind). One particularly useful skill is cultural learning: we should expect AGIs to be able to acquire knowledge from each other and then share their own discoveries in turn, allowing a collective AGI to solve harder problems than any individual AGI within it could. The development of this ability in humans is what allowed the dramatic rise of civilisation over the last ten thousand years. Yet there is little reason to believe that we have reached the peak of this ability, or that AGIs couldn't have a much larger advantage over a human than that human has over a chimp, in acquiring knowledge from other agents.

Thirdly, AGIs will be able to improve the training processes used to develop their successors, which then improve the training processes used to develop their successors, and so on, in a process of *recursive improvement*.[4] Previous discussion has mostly focused on recursive *self*-improvement, involving a single AGI "rewriting its own source code". However, I think it's more appropriate to focus on the broader phenomenon of AIs advancing AI research, for several reasons. Firstly, due to the ease of duplicating AIs, there's no meaningful distinction between an AI improving "itself" versus creating a successor that shares many of its properties. Secondly, modern AIs are more accurately characterised as models which could be retrained, rather than software which could be rewritten: almost all of the work of making a neural network intelligent is done by an optimiser via extensive training. Even a superintelligent AGI would have a hard time significantly improving its cognition by modifying its neural weights directly; it seems analogous to making a human more intelligent via brain surgery (albeit with much more precise tools than we have today). So it's probably more accurate to think about self-modification as the process of an AGI modifying its high-level architecture or training regime, then putting itself through significantly more training. This is very similar to how we create new AIs today, except with humans playing a much smaller role. Thirdly, if the intellectual contribution of humans does shrink significantly, then I don't think it's useful to require that humans are *entirely* out of the loop for AI behaviour to qualify as recursive improvement (although we can still distinguish between cases with more or less human involvement).

These considerations reframe the classic view of recursive self-improvement in a number of ways. For example, the retraining step may be bottlenecked by compute even if an AGI is able to design algorithmic improvements very fast. And for an AGI to trust that its goals will remain the same under retraining will likely require it to solve many of the same problems that the field of AGI safety is currently tackling - which should make us more optimistic that the rest of the world could solve those problems before a misaligned AGI undergoes recursive self-improvement. However, to be clear, this reframing doesn't imply that recursive improvement will be unimportant. Indeed, since AIs will eventually be the primary contributors to AI research, recursive

improvement as defined here will eventually become the key driver of progress. I'll discuss the implications of this claim in the section on Control.

So far I've focused on how superintelligences might come about, and what they will be able to do. But how will they decide what to actually do? For example, will the individuals within a collective AGI even *want* to cooperate with each other to pursue larger goals? Will an AGI capable of recursive improvement have any reason to do so? I'm wary of phrasing these questions in terms of the goals and motivations of AGIs, without exploring more thoroughly what those terms actually mean. That's the focus of the next section.

---

1. Unlike the standard usage, in this technical sense an "environment" also includes a specification of the input-output channels the agent has access to (such as motor outputs), so that solving the task only requires an agent to process input information and communicate output information. ↩

2. This observation is closely related to Moravec's paradox, which I discuss in more detail in the section on Goals and Agency. Perhaps the most salient example is how easy it was for AIs to beat humans at chess. ↩

3. It's not quite clear whether the distinction between "single AGIs" and collective AGIs makes sense in all cases, considering that a single AGI can be composed of many modules which might be very intelligent in their own right. But since it seems unlikely that there will be hundreds or thousands of modules which are each generally intelligent, I think that the distinction will in practice be useful. See also the discussion of "collective superintelligence" in Bostrom's Superintelligence. ↩

4. Whether it's more likely that the successor agent will be an augmented version of the researcher AGI itself or a different, newly-trained AGI is an important question, but one which doesn't affect the argument as made here. ↩

Why is Bayesianism important for rationality?

My impression from the Sequences seems to be that Eliezer considers Bayesianism to be a core element of rationality. Some people have even referred to the community as Bayesian Rationalists. I've always found it curious, like it seemed like more of a technicality most of the time. Why is Bayesianism important or why did Eliezer consider it important?

The Four Children of the Seder as the Simulacra Levels

Previously: Unifying the Simulacra Definitions, Simulacra Levels and their Interactions, On Negative Feedback and Simulacra

Simulacra levels are complex, counter-intuitive and difficult to understand.

Thus, it is good and right to continue exploring them partly via story and metaphor.

The metaphor here will be that of the four children from Jewish Passover Seder.

The Jewish Seder tells us of four generations of children: The wise child, the wicked child, the simple child, and the one who does not know how to ask.

The story is *profoundly weird* and does not, on its face, make much sense. Yet every year it is told anyway. What is going on here?

Many attempts have been made to interpret it.

A while back I wrote the first rationalist seder (later versions can be found here). At the time, the story of the four children did not make sense to me. Why this narrative of decline and fall, of wisdom as something that can only decay?

To make sense of the story of the children and to tie it to the themes I wanted to focus on, I told a reversed story and substituted in generations of rationalists and truth seekers.

In this story, we first learn how to ask, then we are simple, then we are instrumental, then we seek to fully understand, and then finally in a fifth stage we can transcend. We can be great because we stand on the shoulders of giants.

Reversing the order of development is reasonably common, as is an implied fifth child. When I was googling for details of what the sons say, the first hit was a reversed-order story of the children as stages of psychological development, with a fifth stage beyond the four listed.

These are fine tales, worthy of telling. Today, I bring a different story.

I bring the story that I now believe was *originally intended.*

The four children *are the four simulacra levels.*

The wise child represents level 1. They want to know how the Seder works.

The wicked child represents level 2. They want to know what the Seder can get them.

The simple child represents level 3. They want to know what the Seder symbolizes.

The child who does not know how to ask represents level 4. They don't know things anymore.

This hypothesis and the analysis that follows *could* be me doing what Scott Alexander often did and cherry picking to find entertaining and potentially enlightening connections that were clearly never intended. But I actually don't think so.

I believe this is the primary original intent of the story. This makes the four children, and in particular the fourth child, *make sense.* This is not a coincidence because nothing is ever a coincidence.

Quotes are taken from an Orthodox Haggadah excerpt, which is the third hit on a Google search of "the four children passover." The second hit is reform, so it doesn't count. The first hit, as noted above, was Psychology Today doing its own thing, which really shouldn't have been in the highlight box.

You are encouraged to click through to the sources, or even better perform your own search or pick up and read the section from your own Haggadah, to verify that I am not engaging in cherry picking and to consider additional perspectives.

Level One – The Wise Child

The Wise Child lives in object-level reality. She cares about understanding the territory, and knows the map is a means to that end. She wants the facts.

She asks this question:

"What are the testimonies, the statutes, and the laws that G-d, our G-d, has commanded to you?" (deut. 6:20)

A naturalist might interpret this question as "how does the physical world work?"

As she communicates, thus shall you communicate to her. She wants to know the facts, so you give her the facts.

You should respond to him as the Torah commands, "We were slaves to Pharaoh in Egypt, etc." and also instruct him in all the laws of Passover, up to and including its final law: "After eating the Passover offering, one should not then conclude the meal with dessert which would wash away the taste of the Passover offering."

When one cares about the object level, one cares about every detail. The final law, a requirement with a specific physical purpose, is stressed here to illustrate that.

The final law is likely the final law *so that it can be the final law in this passage.* Dessert in the Seder is part of step 13 of 15. It's not a natural place to put a final law.

The act and purpose matter in the Wise Child's object-level literal senses. We wish to remember the taste of the Passover offering, so despite having an explicit phase of the meal for dessert, we must be careful that this dessert does not wash away the taste of the offering.

The act and purpose also matter directly as metaphor, in the more important meaning of both this law and its explanation. We finish the ceremony with joyful songs, but joyful songs that remind us of our struggles and do not hide the truth of our world – we know what the numbers are, the strong prey upon the weak then we all fall to the Angel of Death. Actions have consequences.

We also explicitly remind the Wise Child, that merely observing commandments without understanding them is not sufficient, for to do so would allow not merely them but our other actions and maps to cease to be anchored by reality:

So we tell the Wise Child:

It is true that the essence of the soul transcends the “natural order” of the person—the intellect and emotions—and therefore is blind to distinctions between commandments. It is likewise true that one can observe commandments without understanding them but simply because of the innate, essence-connection between the soul and G-d. One can “pass over” and bypass the complications and limitations of self.

But it is G-d’s will that we experience commandments within the “natural order” of our psyche, within our intellect and emotions. The transcendent “Passover” of our souls then finds expression within and permeates the “laws” of our minds and hearts (The Rebbe).

The very name of the holiday – Passover – *is superficially* about the Exodus from Egypt and the concept that the Angel of Death ‘passed over’ Jewish houses during the tenth plague. But that never really made sense as a justification for the name of the entire holiday. This does.

What the name is really for is a warning to avoid this trap of ‘passing over’ the object level, not forming a gears-level understanding, and allowing our maps to become disconnected from profound reality.

Without discussion and argument, the Seder is hardly a Seder at all.

We must remain anchored in the object level, in our profound reality, if we wish to remain wise.

Inevitably, we lose sight of this, and proceed to level two. Thus, the second generation.

Level Two – The Wicked Child

The Wicked Child cares not about the first level, the obligation to the truth — as embodied by the Torah and the Passover story and Passover service.

Instead, the Wicked Child cares about what effect the service, and the story that we tell at the Seder, will have on others – to be at the second level is to draw a distinction between what you believe and do, and what you seek others to believe and do.

He cares not about whether the service *reflects* reality. He cares about in what way the service could *mask and denature* reality, and *what he can get out of* this service.

He thus asks:

“What is this service of yours?!”

He says of *yours*—implying that it is not for him. By excluding himself from the community, he denies the essential principle of Judaism, the obligation to fulfill the commandments of the Torah.

You should also "blunt his teeth" (speak harshly to him) **and say to him**:

"It is because of this that I would fulfill His commandments, such as this Passover offering, matzah and *maror* **that G-d acted for *me* when I left Egypt** (Exodus 13:8) —for me, but not for *him*. If he [the wicked child] **had been there, he would not have been redeemed.**"

As he speaks on the second level, so we need to respond to him on the second level.

Thus, the first thing we note about the Wicked Child is that he has *separated himself from* this central principle of Judaism, the obligation to the truth. We put his failure to be at level one front and center. That's how important this is.

Yet we do not give up on him. One cannot have level one without the inevitability of level two. To care about what we believe, for any reason, is to invite others to care about what we believe, for their own selfish reasons.

Incentives will always be a thing.

We must constantly remind everyone that we seek truth and to understand and manipulate the object level not (merely) for its own sake, but because this is how we all survive and have nice things. Without this, all is lost.

Thus, we speak back to him in his own language of consequences *to him*. We seek truth *because truth saves us.* We fulfill the obligations of reality and tell its stories that connect us to its profound reality – we are the people of the book – because they grant us freedom and life.

If the Wicked Child had been there, he would not have taken such action, would neither have been of help to or earned the help of the community, and thus *he would not have been saved.*

This is the whole quest. It is the central mission. Once they become wise to this, the child can study the details on their own:

As the Talmud states, a Jew cannot lose his Jewishness. Regardless of the degree of his disengagement from Judaism, the Jewish spark lives on within him.

Kabbalah teaches that the wicked child, *second* of the four children, corresponds to the *second* of the Four Cups. This means that the bulk of the Haggadah is recited over the cup related to the wicked child! Clearly, befriending and educating the wicked child is a central aspect of the Haggadah. For this effort helps bring about the ultimate realization of the Egyptian Exodus.

The Jewish spark here represents this drive towards truth in all of us. Of course this cannot be fully extinguished. Reality is that which, when you stop believing in it, doesn't go away. A sufficiently powerful smackdown from reality will wake anyone (who survives it) up.

It can, however, be *suspended* indefinitely under the wrong conditions.

Thus, we spend the bulk of the Seder speaking primarily to the Wicked Child.

In each generation the wicked child must be convinced of the need to choose wisdom. The wicked child follows from the wise child, as the second level follows from the first.

Only by continuously maintaining right incentives and norms, and hammering the necessary messages into everyone's heads over and over, can we ensure the wicked children among us ultimately choose wisdom.

This is not a struggle that happens once. It happens continuously for each of us that still thinks reality is a thing. Each of us *who still believes that others believe that one thing is and another is not,* is tempted continuously by the ability to say that which is not in order to get others to believe that which is not.

This fits with my model that, while higher-simulacra-levels are always present to some extent, past societies have mostly succeeded at keeping the focus on the object level and thus preventing things on the whole from degenerating further.

Or, that those that have failed at this task have fallen soon thereafter.

When the community fails at this task, the Wicked Children grow up and remain wicked. They continuously work to mask and denature the grand reality. Words become less and less often and less and less substantively a reflection of reality, and more and more a mask of that reality – the mask the speaker wishes to place upon it. In turn, people's expectations adjust.

Things then give way to the third generation.

Level Three – The Simple Child

The Simple Child is not born simple. Nor is she stupid. The Simple Child is responding to incentives. She plays the game laid out before her.

Raised by and around the wicked, The Simple Child lacks the expectation that symbols line up with reality. Those around her have been pretending the whole time. She wants to know how to pretend to do this pretending.

She does not have *or seek* a useful model of physical reality. Such a model does not seem like it would be useful.

She notices instead that rewards and punishments in such a world are best navigated through asking what signals to send. So she seeks to understand symbols well enough to send the right signals.

Thus, the simple child asks the most basic question: "What is this?", or "What is this celebration about?"

You shall say to him: "We are commemorating the fact that **with a strong hand G-d took us out of Egypt, from the house of slaves**" (Exodus 13:14).

As she speaks to you, so shall you speak to her. She wants to know *what this symbol means.* So we tell her what it means, and what and who is to be raised or lowered in status.

We don't actually answer the question! We do not tell her *what this is.*

She isn't really asking for that information. She isn't ready for the answer. We don't have that kind of time. We will. But not now. Not tonight.

But this is all rather tragic. Did we give up on her so easily? Has all been lost by this point? Can we not do better than to get her to think of us as her in-group whose actions should be imitated and signals sent?

This is one of the biggest problems of our age. If someone seeks to be nothing but a partisan, how does one get them to be more than that? If everyone is being judged on their partisanship, how is one to free them from that? To snap them out of it?

The text does not seem to have an answer. The Haggadahs I have used don't even try to answer. This particular version advises:

We tell the simpleton how the Exodus occurred and how he too can experience a personal "Exodus": Just as G-d used a strong hand to "overcome" the attribute of justice, we too must use a strong hand to overcome those aspects of our personalities that impede our spiritual growth. We then experience a spiritual liberation from our personal enslavements.

That does not seem likely to get us much of anywhere. We're talking in mumbo-jumbo in the hopes it will symbolically resonate. All we hold out is the promise of 'spiritual liberation.'

It seems that all the Rabbis believe we can do, at this point, is damage control. Thus, we spend so much time trying to rescue the Wicked Child. That's where there is still some hope. The Simple Child, in this model, is mostly a lost cause.

But we offer a way out. We note that we are commemorating *a fact.*

We link our explanation back to a concrete origin, as a first step in reorienting her attention. It's a trick that just might work.

The 'spiritual liberation' is exactly this – to notice reality and be liberated from being trapped in meaningless symbols. To think for one's self.

That's why there is no talk about the Wise Child's spiritual liberation. There is no need.

Thus, this model says the goal is purely to get the Simple Child to *pay attention.* The promises we make to her are to get her to participate at all, to be present. After that, she can be exposed to the arguments and discussions, to the details. She can notice what is actually going on, and think more on that level.

There is hope. Room to grow. She can still ask questions and care about the answers. Remember her opening question. She asks, what is this? Thus, she still knows on some level that there is a this and it has a what.

What she is unable to do, if she is not helped out of her trap, is pass this remaining understanding along. The fourth generation is coming.

Level Four – The One Who Does Not Know How to Ask

It is frequently pointed out that the name of the fourth generation is profoundly weird.

Have you ever met a *child* who did not know how to ask?

I have not. I've met *adults* who *no longer* know how to ask. Who have fully integrated level four. Who have *forgotten.* The fourth level ceases to know that the first level

exists.

There is the temptation to not engage with the name. To treat it as some sort of metaphor.

The temptation is wrong. The fourth generation *does not know how to ask.*

That does not *quite* mean "*literally* does not know how to ask anything at all". But it also kind of does mean that.

Asking requires realizing that there exist questions and answers. It requires believing that those questions and answers matter. That there is a 'there there' under all that.

He does not know that some things are while other things are not. If answers don't matter, there can be no questions.

Even if he did somehow want that information, *he doesn't know how to ask about actual things.* Everything is a symbol referencing another symbol. There's no way to get those symbols to reference the physical world. Thus, no way to ask a question.

This is the giveaway that we've been talking about simulacrum levels.

The one who does not know how to ask cannot ask for wisdom. For them, wisdom isn't a thing.

And they can't ask how reality works. For them, reality isn't a thing.

What is to be done about this? We must talk in a way he might understand, that might cause him to realize there are things to be understood.

Thus:

As for The One Who Knows Not How To Ask—you must open up [the conversation] for him.

As it is written: You shall tell your child on that day: "It is because of this that G-d acted for me when I left Egypt" (Exodus 13:8).

What we are trying to communicate here is *basic cause and effect.* That there is a *this* and it caused a *that.* Because of this, G-d acted for me when I left Egypt. The very idea of logic, of consequence, is lost upon him. Recover those, together with the idea that some things are and others are not, and the child can learn how to ask. All that matters, for now, is teaching this most basic lesson.

Their need to leave Egypt (which in Hebrew is literally "the narrow place"), is here about the need to realize this. Because we know things and seek knowledge, our world exists and can expand. We can do things, go places, not be trapped. We can be free.

Two levels. Because of these actions, things happened. Because of knowledge, one can take actions that do things.

The child's participation in the Seder is not about any of that; they are just employing systems that attend the rituals that those around them participate in. They go through all the motions, but have no idea what they are doing.

What about alternative interpretations of this stage?

I have heard the suggestion that the fourth child is very young, and does not yet know how to speak. This seems clearly wrong.

If that was what was going on, the child would have a different name – the child who cannot (yet) speak – and our advice for them would be different. The child being unable to speak doesn't make sense in the context of the text telling you to start the conversation for them. If they can't talk, trying to start a conversation about the Exodus would be quite pointless.

Another reason to reject this interpretation is that this child does not yet know how to talk, but *does know how to ask.* He doesn't know the words, but if you hang around a child who hasn't yet learned to talk and pay attention it's clear they can ask about basic things without words.

Another alternative interpretation, from the same Haggadah as above, is this angle:

Too Smart For Questions

This fourth child may be a ritually observant Jew who fulfills all the customs of the Seder. But his Judaism is cold and dry. He does not feel a need for spiritual liberation. He has no questions about or real interest in the Exodus because he does not think of himself as being in exile.

He claims that he is not the excitable type and thus excuses his lifeless Jewish practice. Yet while he cannot muster any excitement for Judaism, he is easily exercised and engaged by material ambitions. He does not realize that his heart and mind are in exile, oblivious to the spiritual content of life.

We cannot begin by telling this Jew what G-d did (as we tell the simple child); we must first inspire him to seek spiritual liberation. We therefore tell him:

"G-d did this for me when *I left Egypt*"—you too are in need of leaving Egypt.

The key insight here is that *we cannot begin the way we did with the Simple Child, by conveying information.* It won't work! The Simple Child has redirected her curiosity, and does not yet much value information, but still understands that information is a thing.

Information would only bounce off The One Who Does Not Know How To Ask. Not being able to ask is merely a symptom. Spiritual liberation again means realizing knowledge exists at all, and is the necessary first step.

However, I think the rest of this is importantly wrong. And it can be wrong in two ways.

First, this child may be *misidentified.*

If the child is instead Simple, going through the ritual without feeling makes sense. The simple child can be told what this is and what to do, and then they go through the motions. It certainly would not occur to them to seek 'spiritual revelation' because life at the third level has no spiritual aspect.

If the child is instead Wicked, that is another potential explanation for this data. They are there to avoid punishment, or to score points, rather than to have the experience and/or better themselves.

The second way this is wrong is the most common mistake when those outside it try to model level four. It is the idea that he is easily exercised and engaged by material ambitions— that those sufficiently at level 4 are doing what the rest of us are doing, engaging in actions because of their model's guess as to their consequences, in order to achieve particular ends.

That's not how level 4 works. Such people don't have goals. They have systems. The fourth child truly is lifeless and unexcited. When such people seem excited, it is because their systems think being excited is the next move, the way deep learning might suggest excitement be expressed at particular points. Nothing more.

Such strategies do often cash out in material ambitions, but that is not because such ambitions excited the person or a plan was formed to get them. The idea of having a plan or ambitions, or of there being a physical thing to be ambitious about, doesn't parse for them the same way it does for others.

Then there's this other note:

The fourth child may actually want to ask but lacks confidence and fears being seen as a fool. The Haggadah instructs us to be sensitive to such people and to put them at ease by initiating conversation with them until they are comfortable sharing their thoughts confidently and clearly (R. Shlomo Alkabetz; Chida).

That is *definitely* not the fourth child. The issue lies elsewhere.

It's certainly a thing that happens. But the child it would be happening to would be the Wise child.

Knowledge is desired. There's social issues in the way, but that is *our fault.*

This is, of course, how it all begins. Children do not start out not knowing how to ask. The problem is caused by the adults who do not know how to answer.

We have somehow taught this child that asking questions can mean being a fool and that this is bad. We've answered his questions by telling him what we want them to see, or what the ritual response to their statement is, rather than by explaining what is and what is not. Without answers, what is a question?

It's on us to fix it. Not them. The prescription here is a good idea, but seems importantly non-central. What is most important is taking away this idea that asking questions is bad or foolish, and setting up an expectation that questions get answers. If seek means ye might find, perhaps then ye will seek.

Otherwise, engaging them in conversation will seem like torture rather than opening them up. It's calling on kids unprompted in class to interrogate and humiliate them. It's grading kids on 'class participation' where participation means guessing the teacher's password. It is being polite at the dinner table until you can ask to be excused. If those around you will only respond to your level one inquiries with level three or four answers, either because that is all they know or they assume that is what you must seek, *then you too do not know how to ask.*

Thus, once things move along sufficiently, the full *generation* does not know how to ask, even those who remain wise, wicked or simple. When they attempt to ask, no answers come. Meaningful questioning ceases.

This is a common failure mode.

Level Five – The Child Who Is Not There

Despite the failings of the four children, they all did the most important thing of all.

They showed up. They are present at the Seder.

That is important because, in this story and metaphor, the Seder (literally ‘order’) represents civilization. It is the ability to know things and pass on that knowledge. Also therefore to accomplish meaningful things, to gather the fruits of our labor.

The fourth generation *still sits down with* the first one. They work together. To some extent, they must listen. This maintains an anchor.

Without the first generation’s renewal and participation, the process cannot be sustained.

As the generations progress, it becomes harder to draw the children into wisdom. Those who are drawn in become less rewarded for it, and more punished. The wicked understand, acknowledge and value the Wise—they depend on the Wise for their own cynical gain. The simple don’t see the point of wisdom. Those who do not know how to ask don’t even know wisdom is a thing.

Finally, there is the child who is not there. Not only do they not know how to ask, they are not connected to those that do. Value in the physical world ceases to be sustained at all. All is lost.

Conclusion, Goals and Takeaways

There were a few distinct goals here.

The first was that when I realized this lined up, it felt too good not to explore and share. Other goals were not necessary, and could be figured out later.

The second was to provide another look at the elephant that provides additional intuition pumps. When something is confusing, the more distinct ways to illustrate both the key points and the details around them, the more likely any given person is to find one that resonates. This also provides additional potential names and references for the levels.

The third was to reinforce in particular the idea that there is something profound that is lost at the fourth level, and to provide help understanding what that is and how that could be. That the fourth level loses its logical facilities. This version puts that so front and center that the loss of logic is explicit and much of the rest of the model is implicit. And it’s important enough that it has survived two thousand years of looking like nonsense.

The fourth, similar to the third, was to provide additional support for the idea of progression through the stages. And to look at how this first attempt tried to halt and

even reverse that progression, in the hopes that we can use those strategies and/or find ways to do better.

This was a fun one. No doubt there are many other similar attempts out there. I can think of several but am curious what people come up with on their own. What are some others, real or fictional?

Is GPT-3 a simulation of the child who does not know how to ask?

I have now produced a book-long sequence on Moral Mazes, and a succession of posts on Simulacra levels. The central hope is to use this as background common knowledge concepts and jargon vocabulary going forward, and that others can do so as well.

Rationality for Kids?

UPDATE 11NOV:

I came up with a game to use as an icebreaker. And I'd love ideas for future variations. It's a combination of Credence Calibration, 20 Questions, and Taboo. The children are trying to determine which of three possible states exist on the card which I have face down (for my first iteration, the possibilities will be "Cat", "Rat", and "Dog"). Every kid gets 30 poker chips to allocate to each of the three possibilities. Kids will then take turns asking a yes or no question, but before each Q, I roll a six sided die. If it comes up six, all chips placed on a wrong answer are turned in, otherwise, they ask their question, I answer with something on a scale of "Never" to "Always", and they are permitted to reallocate their chips. But there is a catch: they are not permitted to use certain words (i.e. cat, dog, rat, meow, bark, pet, etc.) in their questions.
The point is to find tests which can serve as evidence between the possibilities and recognize how confidence should change according to evidence.

Would be interested in other possible states for future iterations

END UPDATE

So I really appreciate the lessons I've learned from "Rationality", but I wish I had learned them earlier in life. We are now homeschooling my kids, and I want to volunteer to teach my kids plus others who are interested lessons about thinking rationally.

Does anyone have recommendations on how to put together a curriculum which gets at the core ideas of rationality, but is oriented towards young kids? Some criteria:

Children will likely range from 7-11, meaning they should be simple concepts and require very little prior knowledge and only the simplest math.

Lessons should be interactive.

Lessons should include TRUE experiments (not just doing fun stuff with chemicals).

Lessons should be fun and appealing enough that parents will want to sign their kids up.

Any other suggestions on the course (wording that will be appealing without sounding too "nerdy" or alarming to the conservative types who usually homeschool) are welcome.

UPDATE: the Inflection Point Curriculum appears to be the middle school version of what I am looking to do:
https://drive.google.com/file/d/1tcUJXRlZXeKjAWeU9Y37FcPKv3lj6PsX/view?usp=sharing

I currently envision the course as a combination of game type exercises like Credence Calibration, Zendo, and Meta-Forms, and experiments like adjusting the air composition of a room and investigating bernoulli effects using things like paper and shower curtains. Other ideas: investigating citrus batteries, water absorption by celery, and the light spectrum of various sources as split by a prism.

Why GPT wants to mesa-optimize & how we might change this

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

This post was inspired by orthonormal's post Developmental Stages of GPTs and the discussion that followed, so only part of it is original.

First I'll aim to provide a crisper version of the argument for why GPT wants to mesa-optimize. Specifically, I'll explain a well-known optimization algorithm used in text generation, and argue that GPT can improve performance on its objective by learning to implement something like this algorithm internally.

Then I'll offer some ideas of mine about how we might change this.

Explanation of beam search

Our goal is to generate plausible text. We evaluate whether text is "plausible" by multiplying together all the individual word probabilities from our language model.

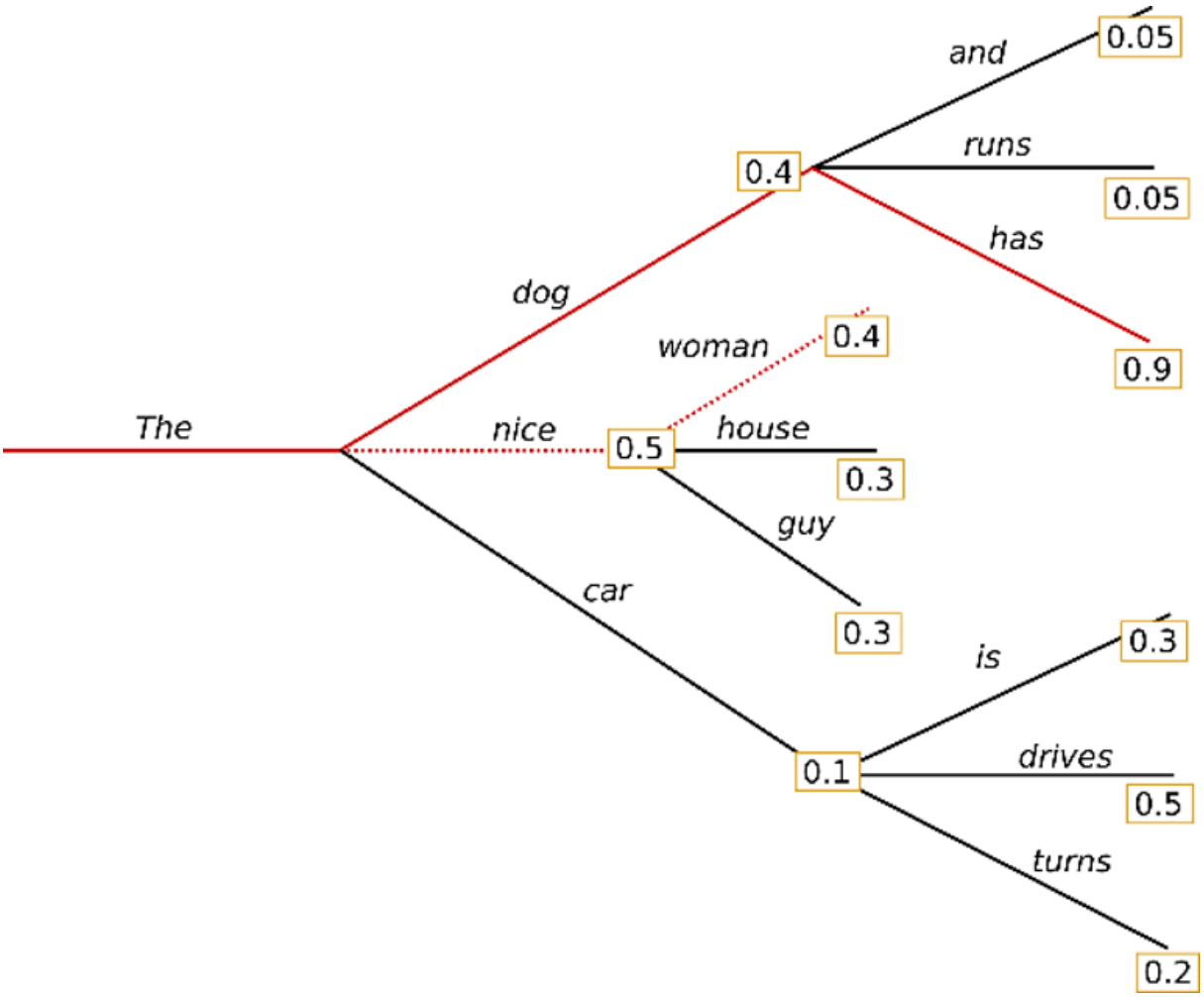
Greedy word selection has a problem: Since it doesn't do lookahead, it's liable to get stuck in a dead end. Let's say we give our system the following poem about cheeses and ask it to generate more text:

> Mozzarella is white
>
> So you can see it at night
>
> Cheddar is...

If our language model is decent, the word it will assign the highest probability to is "orange". But this creates a problem, because "orange" is a hard word to rhyme.

Beam search is an attempt to solve this problem. Instead of picking the next word greedily, we explore the tree of completions and try to find a multi-word completion that maximizes the product of the individual word probabilities.

Because there are so many words in the English language, the tree grows at a very fast exponential rate. So we choose an integer *beam_width* for the number of partial completions to track, and each time we take another step deeper into the tree, we discard all but the most plausible *beam_width* partial completions.



Beam search with a beam width of 2. The bold red path corresponds to the maximum-plausibility completion, which would not get discovered by greedy search because "nice" has a higher probability than "dog". Image stolen from [this Hugging Face blog post](), which has another explanation of beam search if you didn't like mine.

Claim: GPT can do better on its training objective if it learns to do beam search internally

We've discussed text generation with a pretrained language model. Let's switch gears and talk about the model's training process.

Suppose GPT's training corpus has the following poem:

> Mozzarella is white
>
> So you can see it at night
>
> Cheddar is marigold

Unless you let it get too old

GPT is trained by giving it some text and asking it to predict the next word. So eventually GPT will be given the example from above

Mozzarella is white

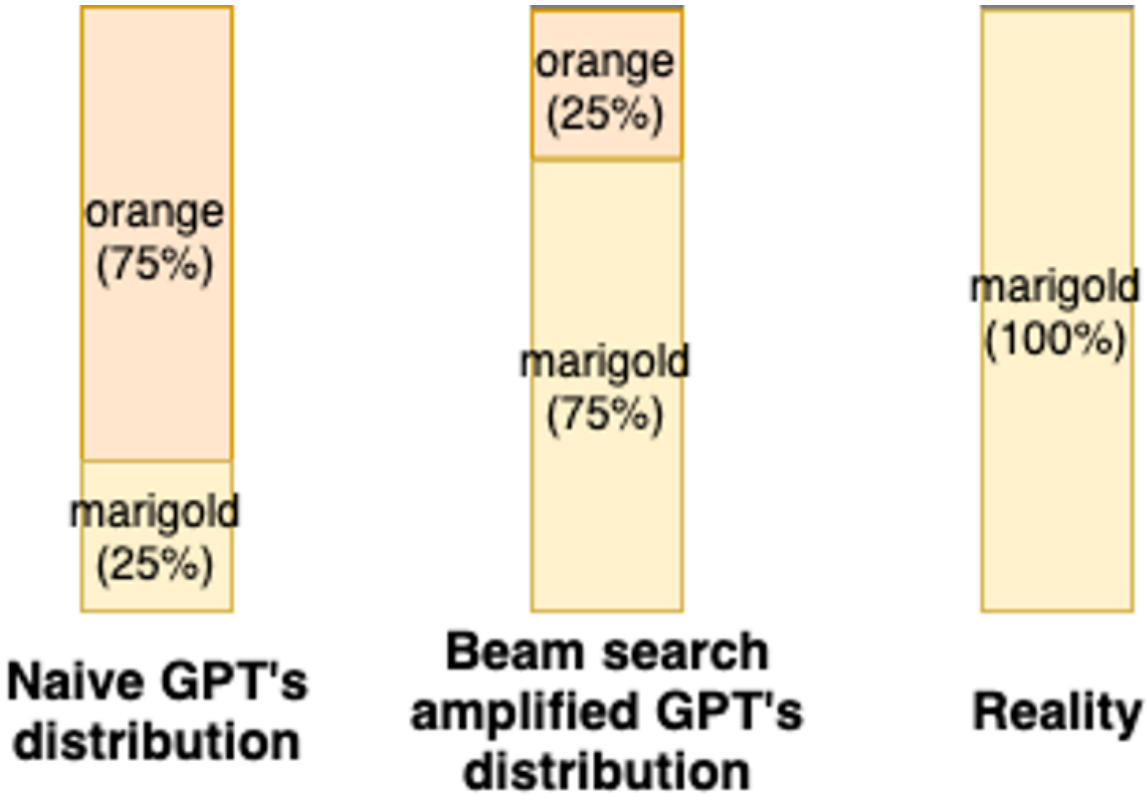
So you can see it at night

Cheddar is...

and be asked to predict the next word.

Let's consider the performance of two models on this task: regular "naive" GPT, and "beam search amplified" GPT. Beam search amplified GPT works by performing beam search using naive GPT, then looking at the distribution of the first words in the resulting completions, then outputting some weighted average of that distribution and the distribution from naive GPT.

Because beam search can find lots of ways to continue the poem using "marigold", but few ways using "orange", beam search amplified GPT's distribution ends up being closer to reality than that of naive GPT. Something like this:



So when we update GPT's weights during training, we're shifting the weights towards the sort of computational structure that would make predictions like beam search amplified GPT does.

Does this actually help?

In this instance, GPT has an incentive to do internal lookahead. But it's unclear how frequently these situations actually arise. And maybe it's usually easier to do something else, like learning which words are easy to rhyme.

It would be straightforward to implement beam search amplified GPT (experimenting with different weighted averaging schemes) and check whether it can be made to assign higher plausibility to real text. (It might be best to try with GPT-2 rather than GPT-3, in case GPT-3 is already doing internal lookahead. Note that there's a risk of mesa-optimization developing if lookahead improves performance at *any point* during GPT's training.)

Is internal lookahead possible for GPT-3?

Relative to other optimization algorithms, it seems to me that beam search would be unusually easy for GPT to implement. Traditional iterative optimization algorithms like gradient descent or simulated annealing require a lot of serial computation, and the number of serial steps GPT can perform is strongly limited. Beam search is way less heavy on the number of serial steps required. The number of available serial steps would still limit the maximum lookahead horizon though.

The transformer architecture learns computations of the form "find some data from the previous step which scores highly according to particular criteria, do some computation on it, pass it on to the next step". That sounds like beam search.

In any case, the topic of what incentives arise while training a language model seems important more generally.

Is internal lookahead dangerous?

If GPT's architecture *is* capable of discovering lookahead internally, the worry is that GPT might modify and misuse it in creative ways after it's discovered. It might start making plans, or searching for the idea that maximizes some attribute which is correlated with harm.

Let's say there are chess problems in GPT's training corpus which describe a board state along with an objective like "black to move and win in 6 turns even with best play by white". If GPT can do lookahead internally, it can use this to search for game histories where black wins even though white is playing very well. In other words, it's doing spontaneous internal planning. And this spontaneous internal planning is incentivized because it helps predict solutions to chess problems.

Who knows what other contexts spontaneous internal planning might get used in.

Fix idea #1: Switch to BERT style training

How might we remove the incentive for mesa-optimization?

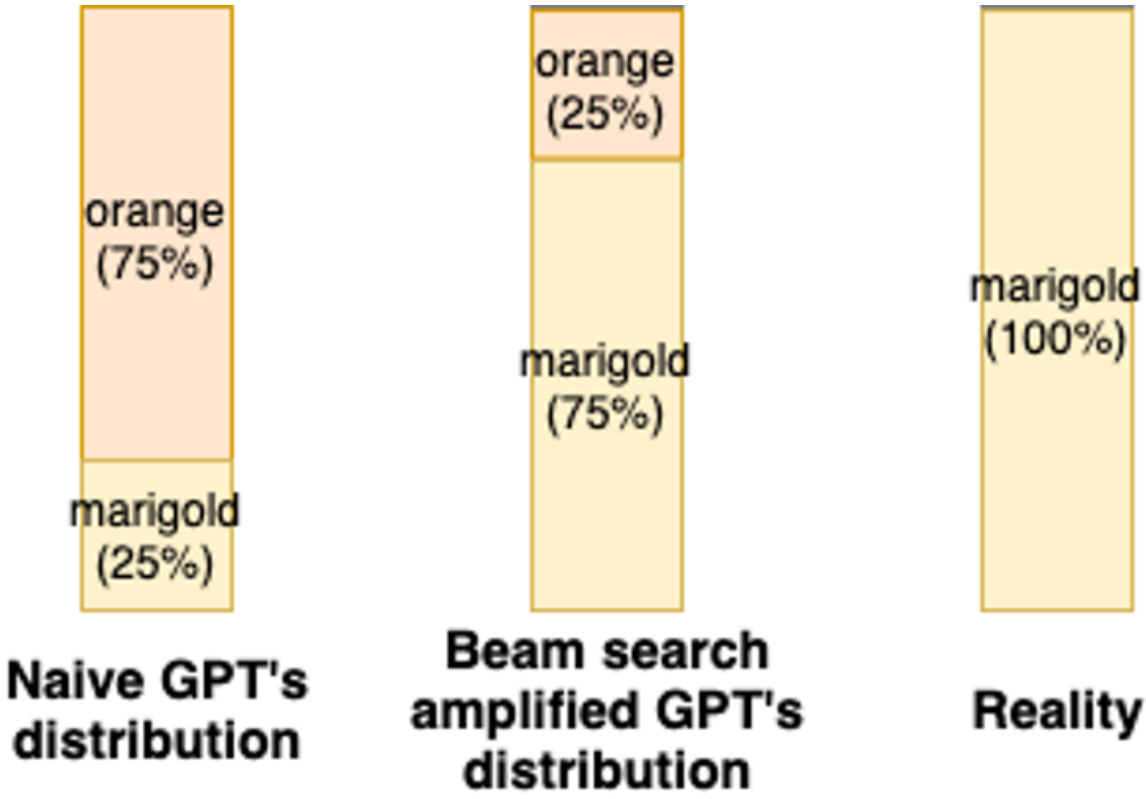
A simple idea is to stop training the task of predicting the next word, and instead train the task of predicting a masked word given the surrounding context. This is what BERT uses. The incentive for internal lookahead seems smaller with this task, but I guess you'd still see it in e.g. predicting masked chess moves of strong players.

BERT's ability to generate text is unclear. But I have a feeling BERT-style training actually offers greater potential than GPT-style training for text generation, because BERT can edit its writing—see discussion here. You could get really creative, like doing genetic algorithms with a population of texts generated on a particular topic, and using BERT to mutate and recombine texts in the population until you get one with really high plausibility. (Some words at the beginning could stay fixed as the "prompt".)

Fix idea #2: Prohibit updates towards lookahead

This fix is more interesting because the core idea seems applicable for stopping mesa-optimization in other contexts too.

v1 of the idea is: During training, before any gradient update, we make a prediction with beam search amplified GPT. And if an update to naive GPT would mean moving it towards beam search amplified GPT (because beam search amplified GPT assigns a higher probability than naive GPT to the actual next word), we just skip the update and keep moving. Nick Bostrom refers to this general approach as "stunting" in Chapter 9 of *Superintelligence*.



This diagram from above is an example of an update that would get skipped.

Training will take longer, obviously. But only insofar as all this mesa-optimization stuff was a problem in the first place.

v2 of the idea is to add a penalty term to GPT's loss function that causes it to update *away* from whatever beam search amplified GPT predicts.

A risk with v2 is that GPT is incentivized to emulate beam search amplified GPT so it can *avoid* making its prediction (i.e. an internal mesa-optimizer gets developed anyway). Skipping updates as in v1 is probably safer, but the fundamental problem seems possible either way.

So v3 is to track gradient updates towards/away from beam search amplified GPT and ensure a rough balance. We don't want systematic movement towards or away from beam search amplified GPT—we don't want incentives to develop either it or its inverse. This could be accomplished by e.g. rescaling gradients on the fly so the average movement towards beam search amplified GPT's answers is around 0. (To be safe, you might want to balance both the number of updates in each direction and their typical magnitude.) An advantage is that training can go faster than in v1 because less updating is skipped.

Beam search isn't the only possible way to do lookahead though.

In particular, the distribution of first words in beam search completions will only assign non-zero probability to words which are the first word of a completion. That leaves room for predictive gains by doing lookahead for other words.

So v4 of the idea is to sample from the infinite beam width distribution, particle filter style, and update in such a way that your expected update towards next-word predictions from the infinite beam width distribution is 0. (Not sure how to do this concretely, but it seems like it might be possible.)

The broader issue here is the nearest unblocked strategy problem. The problem seems relatively tractable in this case, because e.g. beam search based stunting seems to remove the incentive for *any* sort of lookahead insofar as it helps in the cases where beam search amplified GPT outperforms. (I suspect mesa-optimization is not an all-or nothing proposition, and it will matter how much it has been refined + how readily it is deployed.)

Hence, v5 of the idea—even more underspecified than v4 at this point—is adversarial anti-lookahead stunting. Characterize some kind of "lookahead amplifier" which takes naive GPT as input and attempts to amplify it by doing lookahead-type things in a very freeform way. Train this and make it really good at amplifying GPT to achieve better next-word prediction. Then stunt naive GPT during training by forcing its average update towards the next-word predictions of lookahead-amplified GPT to be 0. (Perhaps you'd want to be retraining the lookahead amplifier repeatedly over the course of GPT's training, in case different lookahead strategies work best at different GPT maturity levels.) In production, go ahead and make use of lookahead-amplified GPT to make performance as competitive as possible. (It wouldn't shock me if this scheme actually produced *superior* performance, if the lookahead amplifier learns lookahead that GPT couldn't learn for itself, due to a better inductive bias.)

Ideally, GPT's training process would be monitored to ensure that GPT + lookahead in any form consistently outperforms naive GPT. For bonus points, have the lookahead used in the monitoring be a form of lookahead which is *not* used for stunting (a "dev set").

Do these fixes actually help?

An objection to Fix #2 is the possibility of mesa-optimization which isn't very much like lookahead. For example, if we're training on text that describes a newly discovered animal, the system has an incentive to try & figure out the animal for itself internally so it can better predict how it will be described—and it might make use of some optimization algorithm, genetic algorithms say, to achieve this.

Another objection is that pulling optimization up from the mesa level, as in the "BERT + genetic algorithms" idea or the "lookahead amplifier in production" idea, isn't actually helpful. There's still optimization happening, and the system as a whole could still make devious plans or search for harmful ideas.

However, less mesa-optimization means less risk that transformer blocks develop optimization/planning capabilities and reuse them in contexts we didn't expect. It's easier to reason about searching for text which maximizes plausibility than a mysterious mesa-objective. In particular, an agent that gets instantiated internally might search for side-channel attacks in the text generation machinery and surrounding system (especially risky if GPT has read about this stuff). But it seems very unlikely that a search for plausibility-maximizing text would cause this (except maybe if those attacks somehow got activated during training). Non-mesa-optimization also has parameters that allow us to control its strength without retraining the model, and we have a better understanding of how it works.

There's still a lot of potential for misuse & accidents either way, of course.

OpenAI doesn't offer beam search? Why? Is GPT-3 already mesa-optimizing?

Up until now, I've been pretending that maximizing plausibility (product of individual word probabilities) is a good way to generate text. But beam search doesn't even seem to be an option in the GPT-3 interface. (Please correct me if I'm missing something!)

Why is beam search missing? One possibility is that GPT-3 already does internal lookahead. OpenAI tried beam search, found it didn't improve text generation, and didn't bother adding it as an option. In other words, GPT-3 is already mesa-optimizing 😳

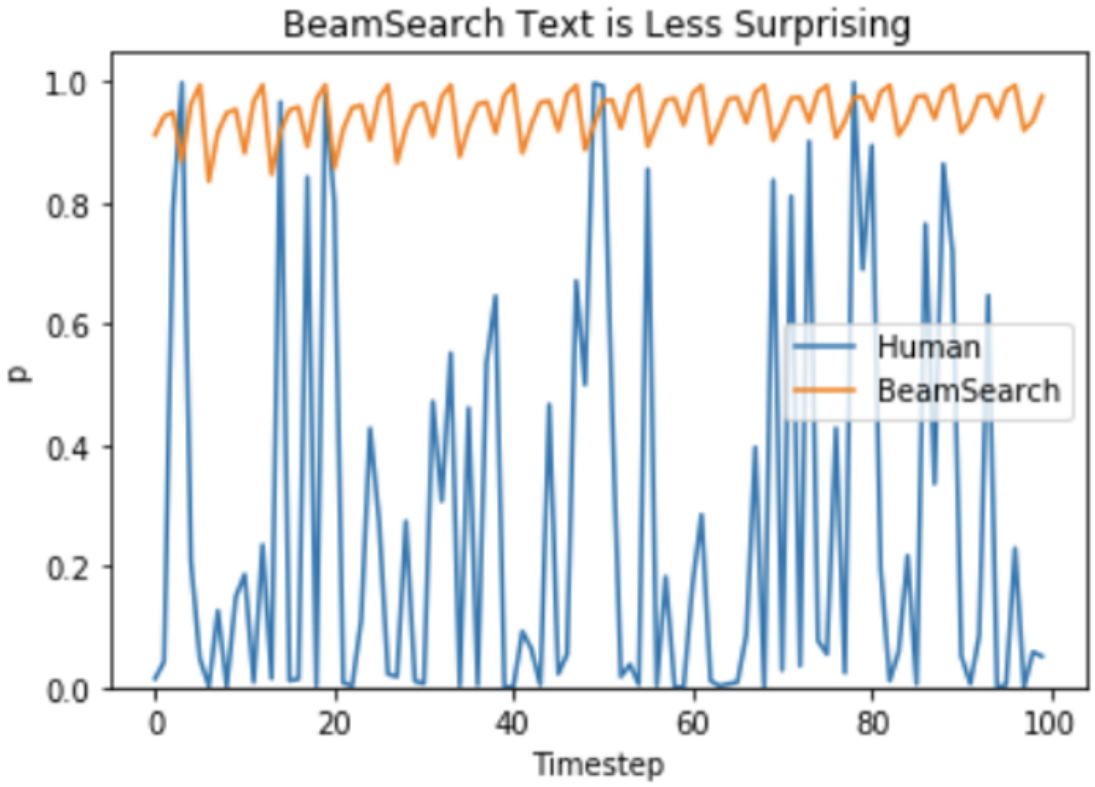
Another possibility:

> [Generated text:] "I enjoy walking with my cute dog, but I'm not sure if I'll ever be able to walk with my dog. I'm not sure if I'll ever be able to walk with my dog."
>
> ...
>
> ...The generated words following the context are reasonable, but the model quickly starts repeating itself! This is a very common problem in language generation in general and seems to be **even more so in greedy and beam search**...
>
> ...

> ...Recently, there has been more evidence though that the apparent flaws of greedy and beam search - mainly generating repetitive word sequences - are caused by the model (especially the way the model is trained), rather than the decoding method, cf. Welleck et al. (2019).

From the Hugging Face post (emphasis mine). OK, this thing about language models that find repetitive text plausible sounds like a problem that will eventually get solved. Anything else?

> As argued in Ari Holtzman et al. (2019), high quality human language does not follow a distribution of high probability next words. In other words, as humans, we want generated text to surprise us and not to be boring/predictable. The authors show this nicely by plotting the probability, a model would give to human text vs. what beam search does.



> So let's stop being boring and introduce some randomness 🤪.

This is a much deeper & more interesting issue IMO. It may be that only a superintelligent language model will find human writing so boringly predictable that every word has high likelihood based on what came before.

Will there be an intermediate stage where prompting a language model with "I just had a brilliant and highly original idea related to X" will cause it to assign higher plausibilities to completions that are actually quite brilliant & original? (Is this the case for GPT-3 already?) I have no idea.

In any case, maybe we could get the benefits of both originality and avoidance of dead ends by sampling from beam search amplified GPT's next-word distribution to generate text? (This could be especially useful if Fix #2 has been applied and the GPT's ability to do lookahead for itself has been stunted.)

Note also that the surprisingness of human text could be an objection to the "GPT can do better on its training objective if it learns to do beam search for itself" claim above. If human text tends to have periodic surprises, using beam search to look for predictable completions may not help performance since those predictions aren't actually very likely.

However, it also may be the case that beam search ends up improving the accuracy of next-word prediction despite the fact that it doesn't generate interesting text.

AGI safety from first principles: Goals and Agency

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

The fundamental concern motivating the second species argument is that AIs will gain too much power over humans, and then use that power in ways we don't endorse. Why might they end up with that power? I'll distinguish three possibilities:

1. AIs pursue power for the sake of achieving other goals; i.e. power is an instrumental goal for them.
2. AIs pursue power for its own sake; i.e. power is a final goal for them.
3. AIs gain power without aiming towards it; e.g. because humans gave it to them.

The first possibility has been the focus of most debate so far, and I'll spend most of this section discussing it. The second hasn't been explored in much depth, but in my opinion is still important; I'll cover it briefly in this section and the next. Following Christiano, I'll call agents which fall into either of these first two categories *influence-seeking*. The third possibility is largely outside the scope of this document, which focuses on dangers from the intentional behaviour of advanced AIs, although I'll briefly touch on it here and in the last section.

The key idea behind the first possibility is Bostrom's instrumental convergence thesis, which states that there are some instrumental goals whose attainment would increase the chances of an agent's final goals being realised for a wide range of final goals and a wide range of situations. Examples of such instrumentally convergent goals include self-preservation, resource acquisition, technological development, and self-improvement, which are all useful for executing further large-scale plans. I think these examples provide a good characterisation of the type of power I'm talking about, which will serve in place of a more explicit definition.

However, the link from instrumentally convergent goals to dangerous influence-seeking is only applicable to agents which have final goals large-scale enough to benefit from these instrumental goals, and which identify and pursue those instrumental goals even when it leads to extreme outcomes (a set of traits which I'll call *goal-directed agency*). It's not yet clear that AGIs will be this type of agent, or have this type of goals. It seems very intuitive that they will because we all have experience of pursuing instrumentally convergent goals, for example by earning and saving money, and can imagine how much better we'd be at them if we were more intelligent. Yet since evolution has ingrained in us many useful short-term drives (in particular the drive towards power itself), it's difficult to determine the extent to which human influence-seeking behaviour is caused by us reasoning about its instrumental usefulness towards larger-scale goals. Our conquest of the world didn't require any humans to strategise over the timeframe of centuries, but merely for many individuals to expand their personal influence in a relatively limited way - by inventing a slightly better tool, or exploring slightly further afield.

Furthermore, we should take seriously the possibility that superintelligent AGIs might be even less focused than humans are on achieving large-scale goals. We can imagine them possessing final goals which don't incentivise the pursuit of power, such as

deontological goals, or small-scale goals. Or perhaps we'll build "tool AIs" which obey our instructions very well without possessing goals of their own - in a similar way to how a calculator doesn't "want" to answer arithmetic questions, but just does the calculations it's given. In order to figure out which of these options is possible or likely, we need to better understand the nature of goals and goal-directed agency. That's the focus of this section.

Frameworks for thinking about agency

To begin, it's crucial to distinguish between the goals which an agent has been *selected* or *designed* to do well at (which I'll call its *design objectives*), and the goals which an agent itself wants to achieve (which I'll just call "the agent's goals").[1] For example, insects can contribute to complex hierarchical societies only because evolution gave them the instincts required to do so: to have "competence without comprehension", in Dennett's terminology. This term also describes current image classifiers and (probably) RL agents like AlphaStar and OpenAI Five: they can be competent at achieving their design objectives without understanding what those objectives are, or how their actions will help achieve them. If we create agents whose design objective is to accumulate power, but without the agent itself having the goal of doing so (e.g. an agent which plays the stock market very well without understanding how that impacts society) that would qualify as the third possibility outlined above.

By contrast, in this section I'm interested in what it means for an agent to have a goal of its own. Three existing frameworks which attempt to answer this question are Von Neumann and Morgenstern's *expected utility maximisation*, Daniel Dennett's *intentional stance*, and Hubinger et al's *mesa-optimisation*. I don't think any of them adequately characterises the type of goal-directed behaviour we want to understand, though. While we can prove elegant theoretical results about utility functions, they are such a broad formalism that practically any behaviour can be described as maximising some utility function. So this framework doesn't constrain our expectations about powerful AGIs.[2] Meanwhile, Dennett argues that taking the intentional stance towards systems can be useful for making predictions about them - but this only works given prior knowledge about what goals they're most likely to have. Predicting the behaviour of a trillion-parameter neural network is very different from applying the intentional stance to existing artifacts. And while we do have an intuitive understanding of complex human goals and how they translate to behaviour, the extent to which it's reasonable to extend those beliefs about goal-directed cognition to artificial intelligences is the very question we need a theory of agency to answer. So while Dennett's framework provides some valuable insights - in particular, that assigning agency to a system is a modelling choice which only applies at certain levels of abstraction - I think it fails to reduce agency to simpler and more tractable concepts.

Additionally, neither framework accounts for bounded rationality: the idea that systems can be "trying to" achieve a goal without taking the best actions to do so. In order to figure out the goals of boundedly rational systems, we'll need to scrutinise the structure of their cognition, rather than treating them as black-box functions from inputs to outputs - in other words, using a "cognitive" definition of agency rather than "behavioural" definitions like the two I've discussed so far. Hubinger et al. use a cognitive definition in their paper on *Risks from Learned Optimisation in Advanced ML systems*: "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". I think this is a promising start, but it has some significant problems. In particular, the concept of "explicit representation" seems like a tricky one - what is explicitly represented within a human brain, if anything? And their definition doesn't draw the important distinction between "local" optimisers such as gradient descent and goal-directed planners such as humans.

My own approach to thinking about agency tries to improve on the approaches above by being more specific about the cognition we expect goal-directed systems to do. Just as "being intelligent" involves applying a range of abilities (as discussed in the previous section), "being goal-directed" involves a system applying some specific additional abilities:

1. *Self-awareness*: it understands that it's a part of the world, and that its behaviour impacts the world;
2. *Planning*: it considers a wide range of possible sequences of behaviours (let's call them "plans"), including long plans;
3. *Consequentialism*: it decides which of those plans is best by considering the value of the outcomes that they produce;
4. *Scale*: its choice is sensitive to the effects of plans over large distances and long time horizons;
5. *Coherence*: it is internally unified towards implementing the single plan it judges to be best;
6. *Flexibility*: it is able to adapt its plans flexibly as circumstances change, rather than just continuing the same patterns of behaviour.

Note that none of these traits should be interpreted as binary; rather, each one defines a different spectrum of possibilities. I'm also not claiming that the combination of these six dimensions is a precise or complete characterisation of agency; merely that it's a good starting point, and the right *type* of way to analyse agency. For instance, it highlights that agency requires a combination of different abilities - and as a corollary, that there are many different ways to be less than maximally agentic. AIs which score very highly on some of these dimensions might score very low on others. Considering each trait in turn, and what lacking it might look like:

1. *Self-awareness*: for humans, intelligence seems intrinsically linked to a first-person perspective. But an AGI trained on abstract third-person data might develop a highly sophisticated world-model that just doesn't include itself or its outputs. A sufficiently advanced language or physics model might fit into this category.
2. *Planning*: highly intelligent agents will by default be able to make extensive and sophisticated plans. But in practice, like humans, they may not always apply this ability. Perhaps, for instance, an agent is only trained to consider restricted types of plans. Myopic training attempts to implement such agents; more generally, an agent could have limits on the actions it considers. For example, a question-answering system might only consider plans of the form "first figure out subproblem 1, then figure out subproblem 2, then...".
3. *Consequentialism*: the usual use of this term in philosophy describes agents which believe that the moral value of their actions depends only on those actions' consequences; here I'm using it in a more general way, to describe agents whose subjective preferences about actions depend mainly on those actions' consequences. It seems natural to expect that agents trained on a reward function determined by the state of the world would be

consequentialists. But note that humans are far from fully consequentialist, since we often obey deontological constraints or constraints on the types of reasoning we endorse.
4. *Scale*: agents which only care about small-scale events may ignore the long-term effects of their actions. Since agents are always trained in small-scale environments, developing large-scale goals requires generalisation (in ways that I discuss below).
5. *Coherence*: humans lack this trait when we're internally conflicted - for example, when our system 1 and system 2 goals differ - or when our goals change a lot over time. While our internal conflicts might just be an artefact of our evolutionary history, we can't rule out individual AGIs developing modularity which might lead to comparable problems. However, it's most natural to think of this trait in the context of a collective, where the individual members could have more or less similar goals, and could be coordinated to a greater or lesser extent.
6. *Flexibility*: an inflexible agent might arise in an environment in which coming up with one initial plan is usually sufficient, or else where there are tradeoffs between making plans and executing them. Such an agent might display sphexish behaviour. Another interesting example might be a multi-agent system in which many AIs contribute to developing plans - such that a single agent is able to execute a given plan, but not able to rethink it very well.

A question-answering system (aka an oracle) could be implemented by an agent lacking either planning or consequentialism. For AIs which act in the real world I think the scale of their goals is a crucial trait to explore, and I'll do so later in this section. We can also evaluate other systems on these criteria. A calculator probably doesn't have any of them. Software that's a little more complicated, like a GPS navigator, should probably be considered consequentialist to a limited extent (because it reroutes people based on how congested traffic is), and perhaps has some of the other traits too, but only slightly. Most animals are self-aware, consequentialist and coherent to various degrees. The traditional conception of AGI has all of the traits above, which would make it capable of pursuing influence-seeking strategies for instrumental reasons. However, note that goal-directedness is not the only factor which determines whether an AI is influence-seeking: the content of its goals also matter. A highly agentic AI which has the goal of remaining subordinate to humans might never take influence-seeking actions. And as previously mentioned, an AI might be influence-seeking because it has the final goal of gaining power, even without possessing many of the traits above. I'll discuss ways to influence the content of an agent's goals in the next section, on Alignment.

The likelihood of developing highly agentic AGI

How likely is it that, in developing an AGI, we produce a system with all of the six traits I identified above? One approach to answering this question involves predicting which types of model architecture and learning algorithms will be used - for example, will they be model-free or model-based? To my mind, this line of thinking is not abstract enough, because we simply don't know enough about how cognition and learning work to map them onto high-level design choices. If we train AGI in a model-free way, I predict it will end up planning using an implicit model anyway. If we train a model-based AGI, I predict its model will be so abstract and hierarchical that looking at its architecture will tell us very little about the actual cognition going on.

At a higher level of abstraction, I think that it'll be easier for AIs to acquire the components of agency listed above if they're also very intelligent. However, the extent to which our most advanced AIs are agentic will depend on what type of training regime is used to produce them. For example, our best language models already generalise well enough from their training data that they can answer a wide range of questions. I can imagine them becoming more and more competent via unsupervised and supervised training, until they are able to answer questions which no human knows the answer to, but still without possessing any of the properties listed above. A relevant analogy might be to the human visual system, which does very useful cognition, but which is not very "goal-directed" in its own right.

My underlying argument is that agency is not just an emergent property of highly intelligent systems, but rather a set of capabilities which need to be developed during training, and which won't arise without selection for it. One piece of supporting evidence is Moravec's paradox: the observation that the cognitive skills which seem most complex to humans are often the easiest for AIs, and vice versa. In particular, Moravec's paradox predicts that building AIs which do complex intellectual work like scientific research might actually be easier than building AIs which share more deeply ingrained features of human cognition like goals and desires. To us, understanding the world and changing the world seem very closely linked, because our ancestors were selected for their ability to act in the world to improve their situations. But if this intuition is flawed, then even reinforcement learners may not develop all the aspects of goal-directedness described above if they're primarily trained to answer questions.

However, there are also arguments that it will be difficult to train AIs to do intellectual work without them also developing goal-directed agency. In the case of humans, it was the need to interact with an open-ended environment to achieve our goals that pushed us to develop our sophisticated general intelligence. The central example of an analogous approach to AGI is training a reinforcement learning agent in a complex simulated 3D environment (or perhaps via extended conversations in a language-only setting). In such environments, agents which strategise about the effects of their actions over long time horizons will generally be able to do better. This implies that our AIs will be subject to optimisation pressure towards becoming more agentic (by my criteria above) will do better. We might expect an AGI to be even more agentic if it's trained, not just in a complex environment, but in a complex competitive multi-agent environment. Agents trained in this way will need to be very good at flexibly adapting plans in the face of adversarial behaviour; and they'll benefit from considering a wider range of plans over a longer timescale than any competitor. On the other hand, it seems very difficult to predict the overall effect of interactions between many agents - in humans, for example, it led to the development of (sometimes non-consequentialist) altruism.

It's currently very uncertain which training regimes will work best to produce AGIs. But if there are several viable ones, we should expect economic pressures to push researchers towards prioritising those which produce the most agentic AIs, because they will be the most useful (assuming that alignment problems don't become serious until we're close to AGI). In general, the broader the task an AI is used for, the more valuable it is for that AI to reason about how to achieve its assigned goal in ways that we may not have specifically trained it to do. For example, a question-answering system with the goal of helping its users understand the world might be much more useful than one that's competent at its design objective of answering questions accurately, but isn't goal-directed in its own right. Overall, however, I think most safety researchers would argue that we should prioritise research directions which produce less agentic AGIs, and then use the resulting AGIs to help us align later more

agentic AGIs. There's also been some work on directly making AGIs less agentic (such as quantilisation), although this has in general been held back by a lack of clarity around these concepts.

I've already discussed recursive improvement in the previous section, but one further point which is useful to highlight here: since being more agentic makes an agent more capable of achieving its goals, agents which are capable of modifying themselves will have incentives to make themselves more agentic too (as humans already try to do, albeit in limited ways).[3] So we should consider this to be one type of recursive improvement, to which many of the considerations discussed in the previous section also apply.

Goals as generalised concepts

I should note that I don't expect our training tasks to replicate the scale or duration of all the tasks we care about in the real world. So AGIs won't be directly selected to have very large-scale or long-term goals. Yet it's likely that the goals they learn in their training environments will generalise to larger scales, just as humans developed large-scale goals from evolving in a relatively limited ancestral environment. In modern society, people often spend their whole lives trying to significantly influence the entire world - via science, business, politics, and many other channels. And some people aspire to have a worldwide impact over the timeframe of centuries, millennia or longer, even though there was never significant evolutionary selection in favour of humans who cared about what happened in several centuries' time, or paid attention to events on the other side of the world. This gives us reason to be concerned that even AGIs which aren't explicitly trained to pursue ambitious large-scale goals might do so anyway. I also expect researchers to actively aim towards achieving this type of generalisation to longer time horizons in AIs, because some important applications rely on it. For long-term tasks like being a CEO, AGIs will need the capability and motivation to choose between possible actions based on their worldwide consequences on the timeframe of years or decades.

Can we be more specific about what it looks like for goals to generalise to much larger scales? Given the problems with the expected utility maximisation framework I identified earlier, it doesn't seem useful to think of goals as utility functions over states of the world. Rather, an agent's goals can be formulated in terms of whatever concepts it possesses - regardless of whether those concepts refer to its own thought processes, deontological rules, or outcomes in the external world.[4] And insofar as an agent's concepts flexibly adjust and generalise to new circumstances, the goals which refer to them will do the same. It's difficult and speculative to try to describe how such generalisation may occur, but broadly speaking, we should expect that intelligent agents are able to abstract away the differences between objects or situations that have high-level similarities. For example, after being trained in a simulation, an agent might transfer its attitudes towards objects and situations in the simulation to their counterparts in the (much larger) real world.[5] Alternatively, the generalisation could be in the framing of the goal: an agent which has always been rewarded for accumulating resources in its training environment might interalise the goal of "amassing as many resources as possible". Similarly, agents which are trained adversarially in a small-scale domain might develop a goal of outcompeting each other which persists even when they're both operating at a very large scale.

From this perspective, to predict an agent's behaviour, we will need to consider what concepts it will possess, how those will generalise, and how the agent will reason about them. I'm aware that this appears to be an intractably difficult task - even human-level reasoning can lead to extreme and unpredictable conclusions (as the history of philosophy shows). However, I hope that we can instill lower-level mindsets or values into AGIs which guide their high-level reasoning in safe directions. I'll discuss some approaches to doing so in the next section, on Alignment.

Groups and agency

After discussing collective AGIs in the previous section, it seems important to examine whether the framework I've proposed for understanding agency can apply to a group of agents as well. I think it can: there's no reason that the traits I described above need to be instantiated within a single neural network. However, the relationship between the goal-directedness of a collective AGI and the goal-directedness of its individual members may not be straightforward, since it depends on the internal interactions between its members.

One of the key variables is how much (and what types of) experience those members have of interacting with each other during training. If they have been trained primarily to cooperate, that makes it more likely that the resulting collective AGI is a goal-directed agent, even if none of the individual members is highly agentic. But there are good reasons to expect that the training process will involve some competition between members, which would undermine their coherence as a group. Internal competition might also increase short-term influence-seeking behaviour, since each member will have learned to pursue influence in order to outcompete the others. As a particularly salient example, humanity managed to take over the world over a period of millennia not via a unified plan to do so, but rather as a result of many individuals trying to expand their short-term influence.

It's also possible that the members of a collective AGI have not been trained to interact with each other at all, in which case cooperation between them would depend entirely on their ability to generalise from their existing skills. It's difficult to imagine this case, because human brains are so well-adapted for group interactions. But insofar as humans and aligned AGIs hold a disproportionate share of power over the world, there is a natural incentive for AGIs pursuing misaligned goals to coordinate with each other to increase their influence at our expense.[6] Whether they succeed in doing so will depend on what sort of coordination mechanisms they are able to design.

A second factor is how much specialisation there is within the collective AGI. In the case where it consists only of copies of the same agent, we should expect that the copies understand each other very well, and share goals to a large extent. If so, we might be able to make predictions about the goal-directedness of the entire group merely by examining the original agent. But another case worth considering is a collective consisting of a range of agents with different skills. With this type of specialisation, the collective as a whole could be much more agentic than any individual agent within it, which might make it easier to deploy subsets of the collective safely.

---

1. AI systems which learn to pursue goals are also known as *mesa-optimisers*, as coined in Hubinger et al's paper *Risks from Learned Optimisation in Advanced*

Machine Learning Systems. ↩

2. Related arguments exist which attempt to do so. For example, Eliezer Yudkowsky argues here that, “while corrigibility probably has a core which is of lower algorithmic complexity than all of human value, this core is liable to be very hard to find or reproduce by supervised learning of human-labeled data, because deference is an unusually anti-natural shape for cognition, in a way that a simple utility function would not be an anti-natural shape for cognition.” Note, however, that this argument relies on the intuitive distinction between natural and anti-natural shapes for cognition. This is precisely what I think we need to understand to build safe AGI - but there has been little explicit investigation of it so far. ↩

3. I believe this idea comes from Anna Salamon; unfortunately I’ve been unable to track down the exact source. ↩

4. For example, when people want to be “cooperative” or “moral”, they’re often not just thinking about results, but rather the types of actions they should take, or the types of decision procedures they should use to generate those actions. An additional complication is that humans don’t have full introspective access to all our concepts - so we need to also consider unconscious concepts. ↩

5. Consider if this happened to you, and you were pulled “out of the simulation” into a real world which is quite similar to what you’d already experienced. By default you would likely still want to eat good food, have fulfilling relationships, and so on, despite the radical ontological shift you just underwent. ↩

6. In addition to the prima facie argument that intelligence increases coordination ability, it is likely that AGIs will have access to commitment devices not available to humans by virtue of being digital. For example, they could send potential allies a copy of themselves for inspection, to increase confidence in their trustworthiness. However, there are also human commitment devices that AGIs will have less access to - for example, putting ourselves in physical danger as an honest signal. And it’s possible that the relative difficulty of lying versus detecting lying shifts in favour of the former for more intelligent agents. ↩

Safety via selection for obedience

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

In a previous post, I argued that it's plausible that "the most interesting and intelligent behaviour [of AGIs] won't be directly incentivised by their reward functions" - instead, "many of the selection pressures exerted upon them will come from *emergent* interaction dynamics". If I'm right, and the easiest way to build AGI is using open-ended environments and reward functions, then we should be less optimistic about using scalable oversight techniques for the purposes of safety - since capabilities researchers won't need good oversight techniques to get to AGI, and most training will occur in environments in which good and bad behaviour aren't well-defined anyway. In this scenario, the best approach to improving safety might involve structural modifications to training environments to change the emergent incentives of agents, as I'll explain in this post.

My default example of the power of structural modifications is the evolution of altruism in humans. Consider Fletcher and Doebeli's model of the development of altruism, which relies on assortment in repeated games - that is, when players with a tendency to cooperate end up playing together more often than random chance predicts. In humans, some of the mechanisms which lead to assortment are:

- Kin recognition: we can tell who we share genes with.
- Observation of intentions or previous behaviour: these give us evidence about other agents' future behaviour.
- Costly signalling: this can allow us to reliably demonstrate our future altruism.
- Communication of observed information: once one person has made an observation, it can be shared widely.
- Flexible interactions: we can choose who to assort with in different interactions.

I claim that, given this type of understanding of the evolution of altruism, we can identify changes to high-level properties of the human ancestral environment which would have made humans significantly more altruistic. For example, human cognition is not very transparent, and so it's relatively difficult for each of us to predict the intentions of others. However, if we had direct observational access to each other's brains, cooperation would become easier and more advantageous. As another example, if we had evolved in environments where we frequently cooperated with many different species, then we'd likely feel more broadly altruistic today.

To be clear, I don't think these types of interventions are robust enough to be plausible paths to building safe AGIs: they're only intuition pumps. In particular, I expect it to be much easier to push AIs to learn to cooperate by directly modifying their reward functions to depend on the rewards gained by other agents. However, agents trained in this way might still learn to care about instrumental goals such as acquiring resources. After all, those instrumental goals will still be useful in allowing them to benefit themselves and others; and unless we train them in a purely cooperative environment, they will still be rewarded for outcompeting other agents. Our question now is: how do we train agents which *only* care about fulfilling the goals of other agents, while lacking any other goals of their own?

The approach I'm most excited about is changing the high-level properties of the training environment in a way which encourages division of labour, in particular by separating the roles of planner and worker. For example, consider a meta-learning setup incorporating many different tasks. In each episode, the planner agent A is given a detailed set of instructions in natural language, specifying a task which is too complicated for A to do on its own. The worker agent, B, receives the same rewards as agent A, but doesn't have access to the same instructions. Instead, B can only receive short commands or feedback from A, which forces it to pay attention to A's instructions. We then train A and B on a very wide range of tasks, so that B learns the general skill of inferring what A wants and acting accordingly. Indeed, we should make A's reward function and instructions change even within each episode, so that B is always incentivised to continue listening to A's override commands. And of course we should train each B agent with many different As, so that the Bs don't overfit to obeying one particular A.

The purpose of this setup is to train B to be only motivated by obedience to A, and not at all intrinsically motivated to pursue other goals *even when those other goals are usually useful for increasing A and B's shared reward*. In other words: if we want agents that remain obedient in a wide range of circumstances, then let's train them to remain obedient in a wide range of circumstances! Instead of putting the burden of hunting down undesirable correlations on the overseer, we're putting the burden on whichever process generates a wide range of environments, to ensure that for any potential misaligned goal B might learn, there's some environment where carrying out that goal will be a worse strategy than just following A's instructions. Then, although the combination of A and B may be highly agentic, we can hope that each of them will be safe when used individually.

It's not clear that this is a bet which will pay off, but this type of training regime seems worth investigating. Here are three more speculative ideas for ways to structurally modify open-ended training regimes to develop more obedient agents, inspired by selection pressures which affected the extent to which humans are "obedient" and "domesticated":

- Increasing specialisation. Agents which are only able to influence their environments in specific, limited ways will have greater incentives to be obedient. For example, if B's intelligence and observation space are strictly greater than A's, then it may sometimes be useful for B to override A. Whereas if B can't see most of the environment, while A can, then obeying A is more likely to increase reward according to their shared reward function.
    - Specialisation might also arise naturally if agents are trained in sufficiently large groups to make it viable, with sufficiently good communication channels to stay coordinated.
- Increasing the value of learning from others. Agents which are trained in settings where most knowledge is culturally transmitted, rather than derivable individually, will have greater incentives to listen to others rather than always making their own decisions.
    - Given this claim, we may be able to increase our agents' tendencies towards obedience by making cultural knowledge transmission easier. For example, we could allow agents to easily write to and read from a permanent record (as compared with humans, who needed to invent laborious techniques to facilitate reading and writing).
- Increasing the value of coordination. Agents trained on tasks which require large-scale coordination will likely learn to be more obedient to central planners, since unilateral action isn't very useful for solving such tasks. In such cases,

other agents might learn to detect and punish disobedience, since it has negative externalities.

One key concern with these proposals is that the concepts learned by agents in simulation won't generalise to the real world. For example, they may learn the goal of obedience to a broad class of artificial agents, but not want to obey humans, since we're very different. It's particularly hard to predict the generalisation of goals of artificial agents because our reasoning about goals often falls prey to anthropomorphic optimism. To counter this concern, we can try to use adversarial training in a wide variety of domains, including some domains where humans are directly involved (although I expect human oversight to be too expensive to make up more than a small minority of training time). Thorough testing is also very important.

Testing multi-agent safety

Many of our current safety test environments have the problem that, while they contain suggestively-labelled components which make sense to humans, they aren't rich enough to develop the sort of deliberate misbehaviour we are worried about from AGI. To deliberately misbehave, an agent needs a theory of mind, and the ability to predict the consequences of its actions. Without these, even though we can design environments in which the agent takes the action that we've *labelled* as misbehaviour, the agents won't have the semantic content which we want to check for. While I expect sufficiently advanced language models to have that semantic content, I expect that the easiest way to observe potential misbehaviour is in 3D environments.

In particular, our default test for the safety of an AI should involve putting it in novel simulated environments with other AIs it's never interacted with before, and seeing what happens. Cooperation between those AIs - especially via the mechanism of some of them being obedient to the commands of others - would be evidence of their safety. It'd be particularly valuable to see if agents trained in very different ways, with very different patterns of behaviour, would cooperate in novel environments. After testing in simulation, we could also test by deploying agents on increasingly complex tasks in the real world.

A second approach to testing safety might use more theoretical ideas. It's usually difficult to formally reason about complex goals in complex environments, but we can take inspiration from the field of evolutionary biology, which does so by analysing the incentives of agents to help or harm each other given how related they are. From similar incentive analysis in agents, we then might be able to derive theories which we can empirically test, and then use to make predictions.

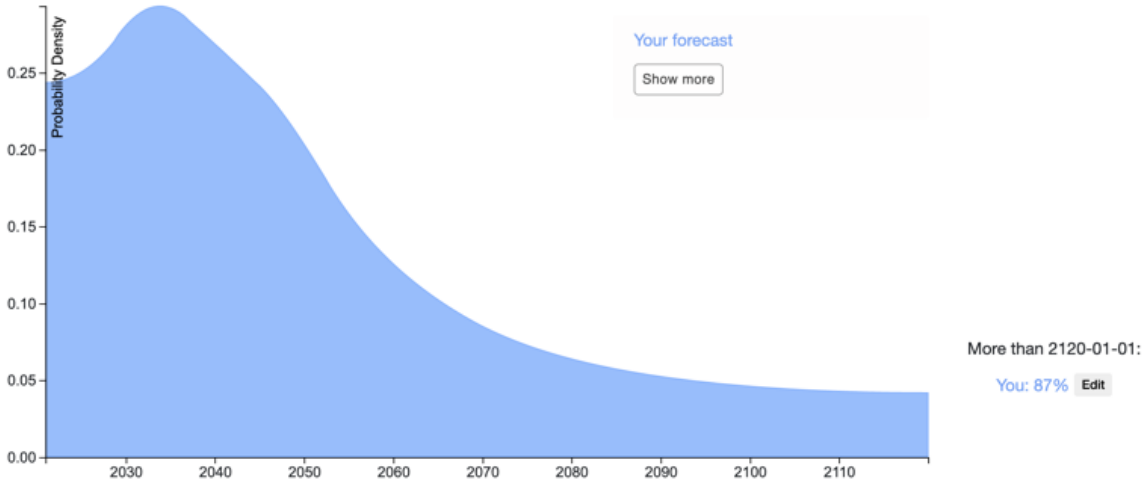
Another relevant formalism for these situations is that of bargaining games. However, there are a couple of ways in which multi-agent environments differ from standard bargaining games. Firstly, the former are almost always iterated, especially if they're in a persistent environment. Secondly, there are reputational effects in the former, but not in the latter. Thirdly, all of the agents involved are being optimised based on how well they perform, which shifts their policies over time. So I'm not too optimistic about this type of analysis.

Forecasting Thread: Existential Risk

This is a thread for displaying your probabilities of an existential catastrophe that causes extinction or the destruction of humanity's long-term potential.

Every answer to this post should be a forecast showing your probability of an existential catastrophe happening at any given time.

For example, here is Michael Aird's timeline:



The goal of this thread is to create a set of comparable, standardized x-risk predictions, and to facilitate discussion on the reasoning and assumptions behind those predictions. The thread isn't about setting predictions in stone – you can come back and update at any point!

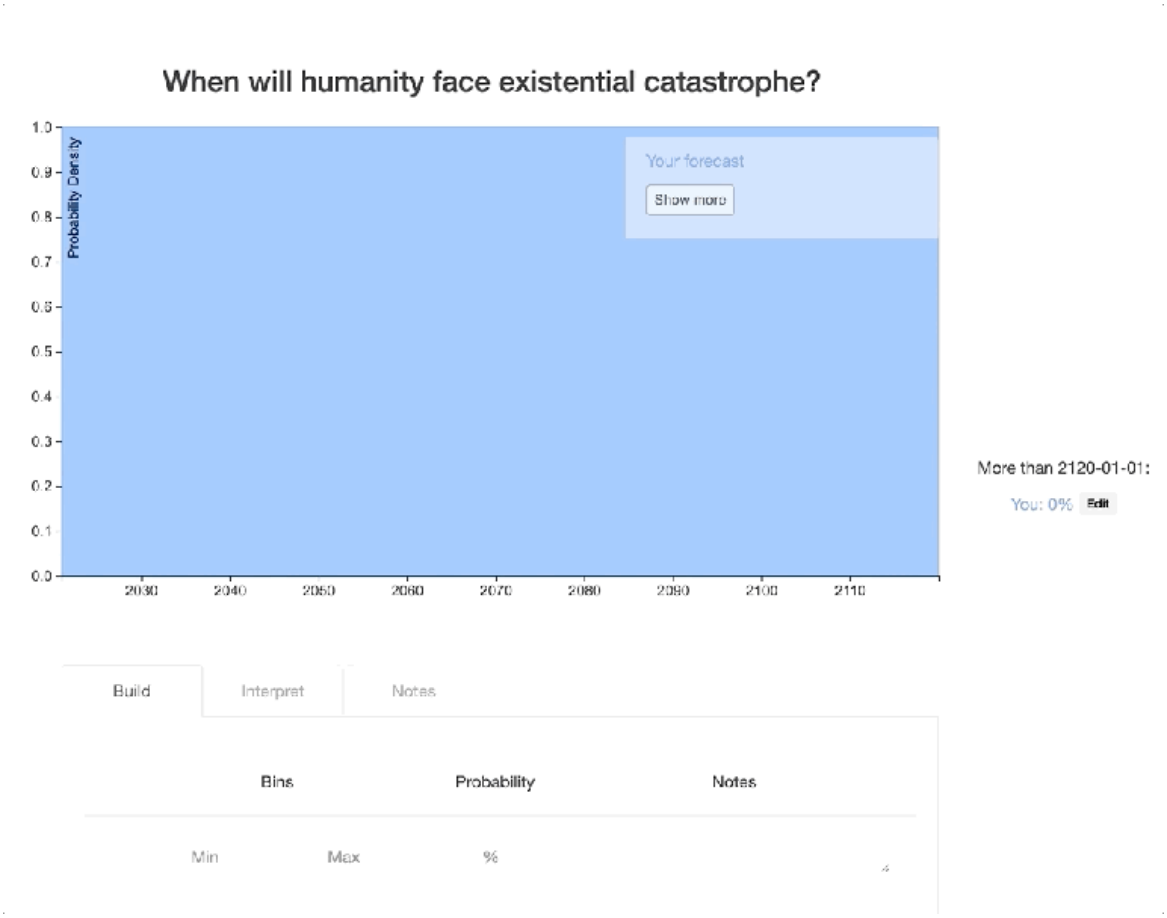
How to participate

1. **Go to this page**
2. **Create your distribution**
    - Specify an interval using the Min and Max bin, and put the probability you assign to that interval in the probability bin.
    - You can specify a cumulative probability by leaving the Min box blank and entering the cumulative value in the Max box.
    - To put probability on *never*, assign probability above January 1, 2120 using the edit button to the right of the graph. Specify your probability for *never* in the notes, to distinguish this from putting probability on existential catastrophe occurring after 2120.
3. **Click 'Save snapshot' to save your distribution to a static URL**
    - A timestamp will appear below the 'Save snapshot' button. This links to the URL of your snapshot.
    - Make sure to copy it before refreshing the page, otherwise it will disappear.
4. **Click 'Log in' to automatically show your snapshot on the Elicit question page**
    - You don't have to log in, but if you do, Elicit will:
        - Store your snapshot in your account history so you can easily access it.
        - Automatically add your most recent snapshot to the x-risk question page under 'Show more'. Other users will be able to import your most recent snapshot from the dropdown, shown below.

- We'll set a default name that your snapshot will be shown under – if you want to change it, you can do so on your profile page.
- If you're logged in, *your snapshots for this question will be publicly viewable.*

5. **Copy the snapshot timestamp link and paste it into your LessWrong comment**
   - You can also add a screenshot of your distribution in your comment using the instructions below.

Here's an example of how to make your distribution:



How to add an image to your comment

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

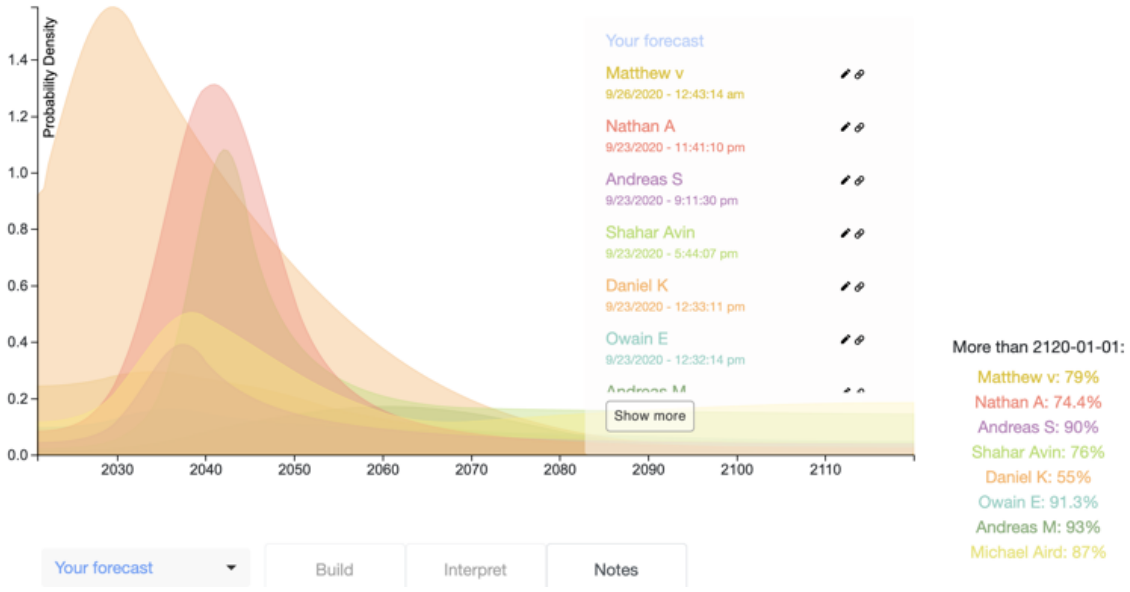
If you have any bugs or technical issues, reply to Ben from the LW team or Amanda (me) from the Ought team in the comment section, or email me at amanda@ought.org.

Questions to consider as you're making your prediction

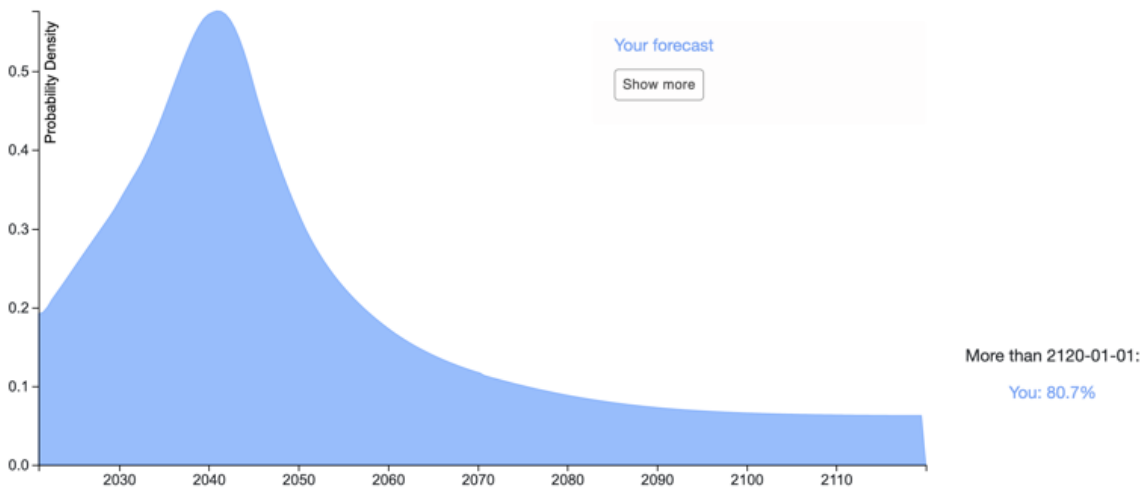
- What definitions are you using? It’s helpful to specify them.
- What evidence is driving your prediction?
- What are the main assumptions that other people might disagree with?
- What evidence would cause you to update?
- How is the probability mass allocated amongst x-risk scenarios?
- Would you bet on these probabilities?

Comparisons and aggregations

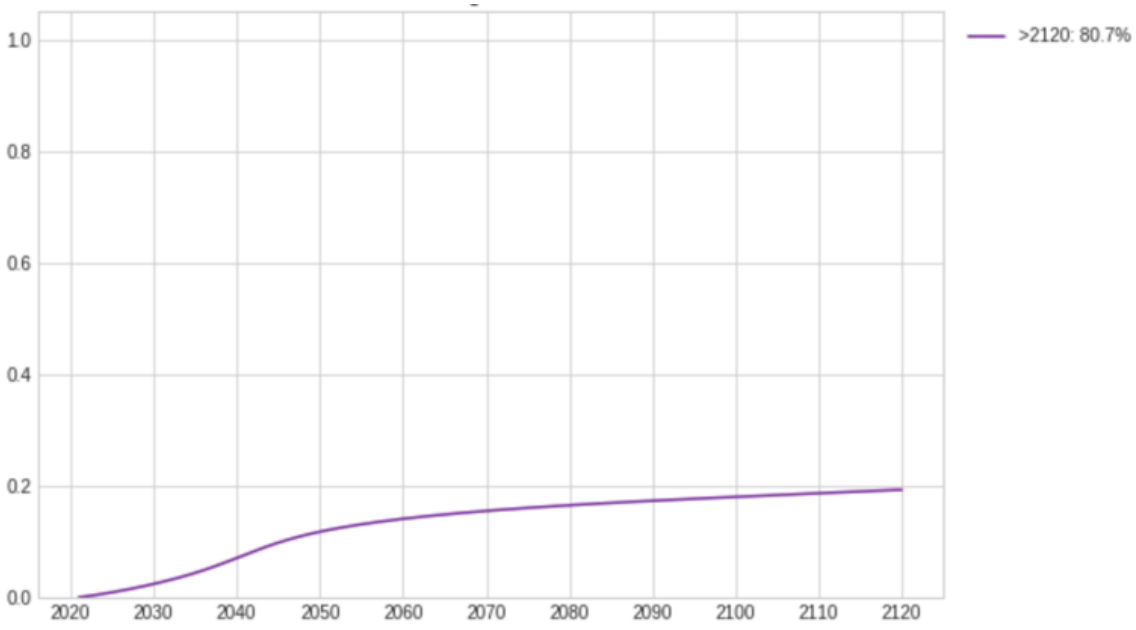
Here's a comparison of the 8 predictions made so far (last updated 9/26/20).



Here's a distribution averaging all the predictions (last updated 9/26/20). The averaged distribution puts **19.3% probability before 2120** and **80.7% after 2120.** The year within 2021-2120 with the greatest risk is **2040**.



Here's a CDF of the averaged distribution:



CTWTB: Paths of Computation State

This is the second post of Category Theory Without The Baggage; first post here . You can probably follow most of this post without reading the first one. Be warned that I am a novice when it comes to category theory; please leave a comment if you see a substantive error or oversight.

Suppose we want to evaluate $\frac{x(x+3)}{x+y}$ at the point x = 2, y = 5. We have a choice about the order in which to perform the operations. One possibility:

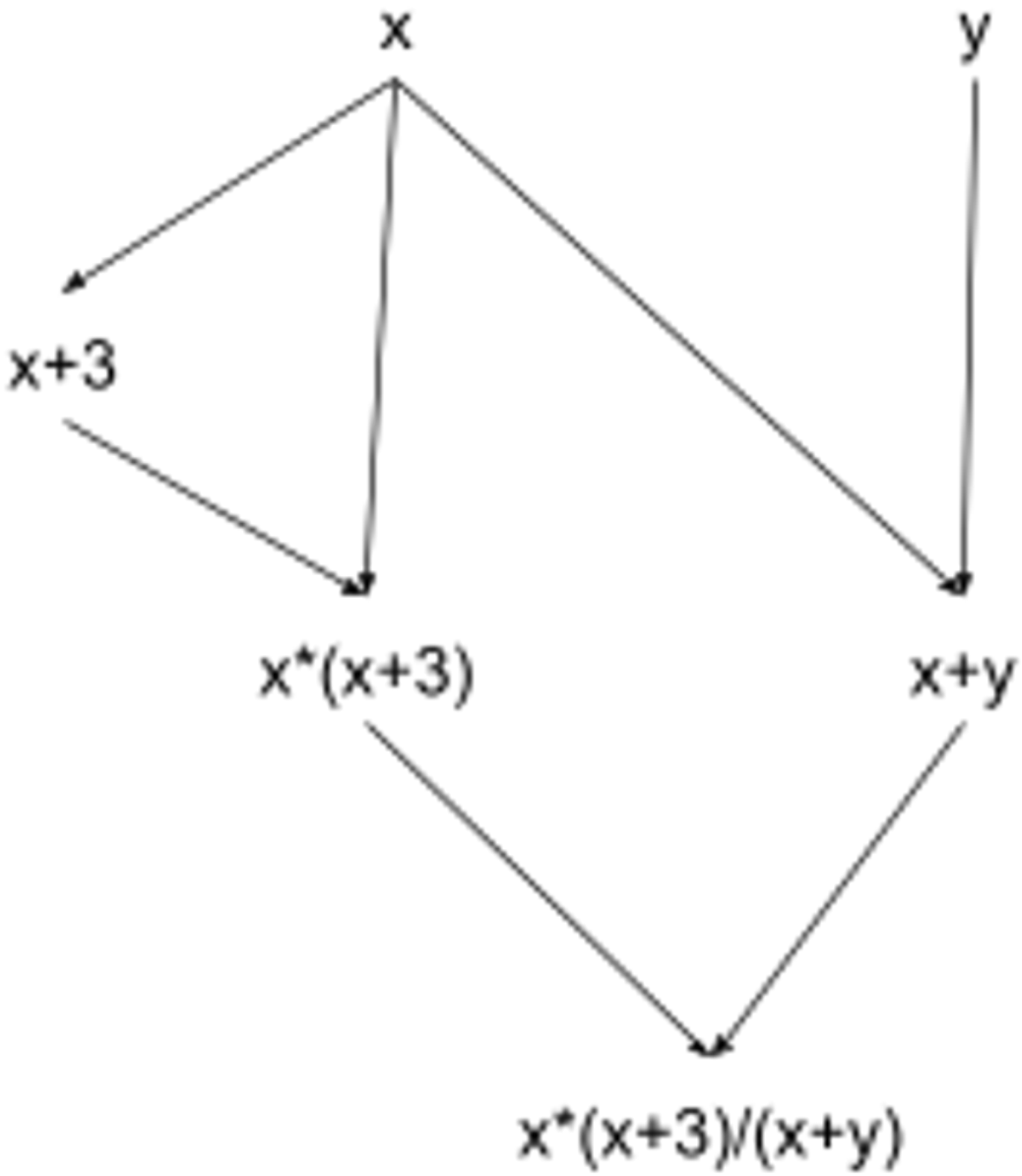
- Add x+y
- Add x+3
- Multiply x*(x+3)
- Divide

Another possibility:

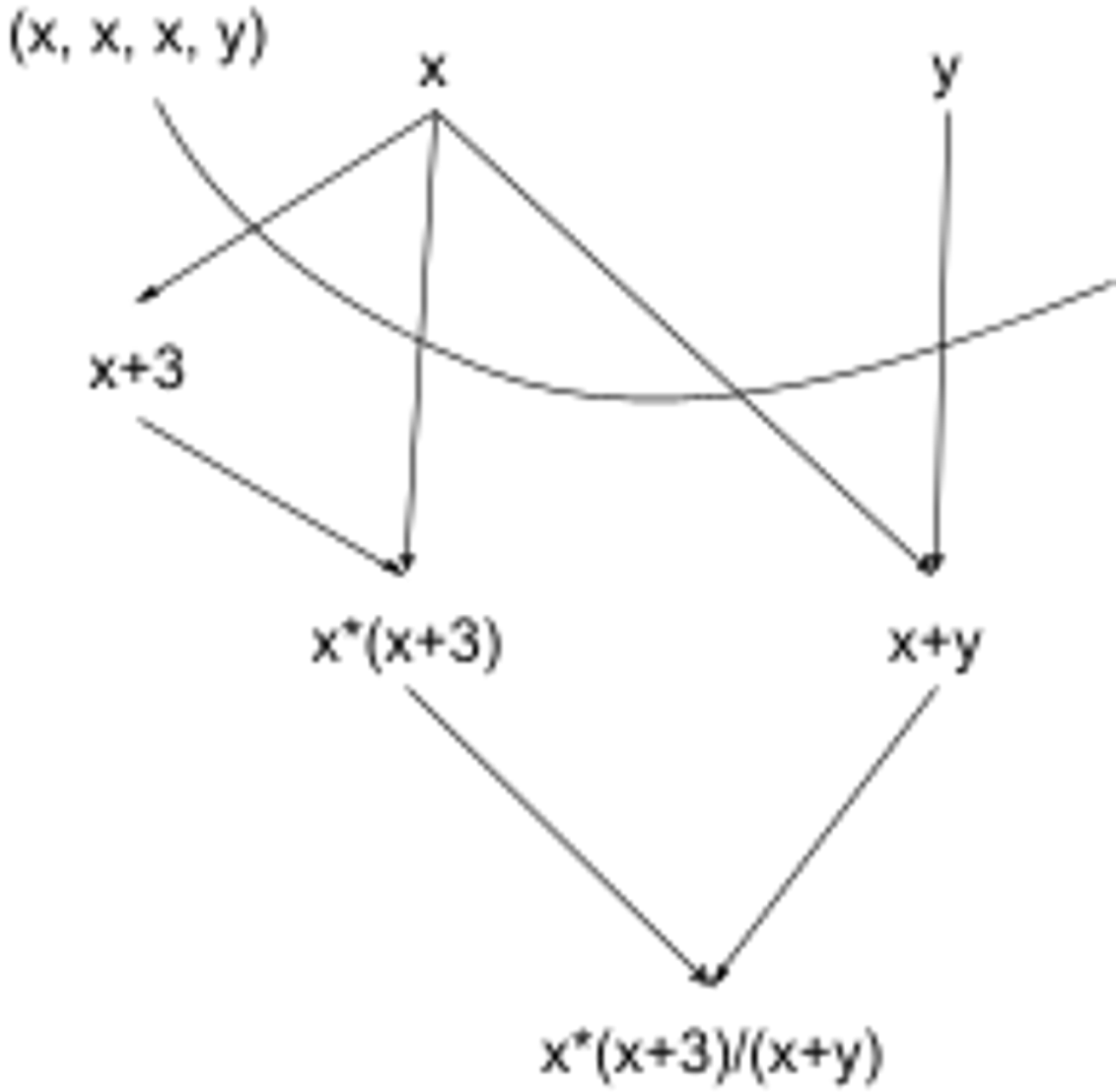
- Add x+3
- Add x+y
- Multiply x*(x+3)
- Divide

This hopefully seems trivial in such a simple example, but we want to be able to easily talk about this sort of thing in more complicated problems. To that end, we're going to visualize these possible computation-orders in terms of graphs.

We'll start with a graph representing the expression itself - i.e. the usual visual for a circuit:

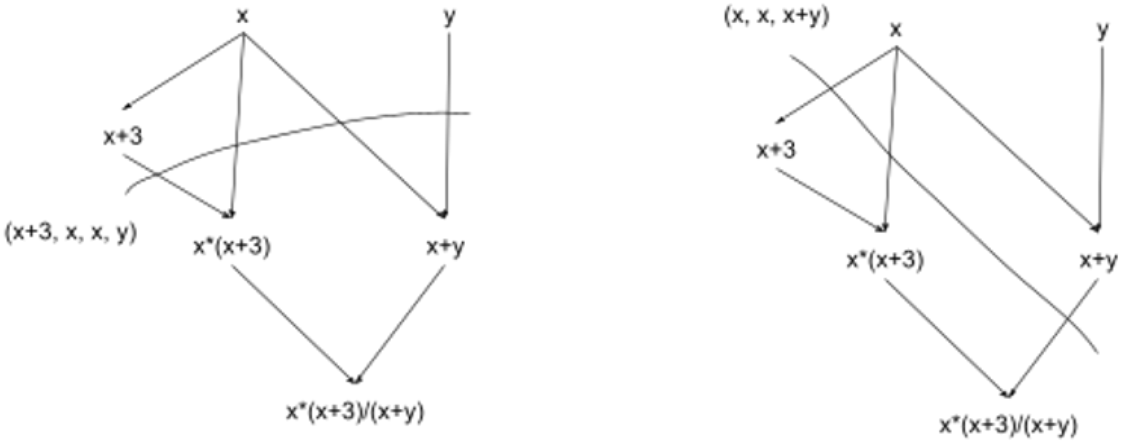


We can represent “computation state” as a cut through this DAG. For instance, we start out in this state:



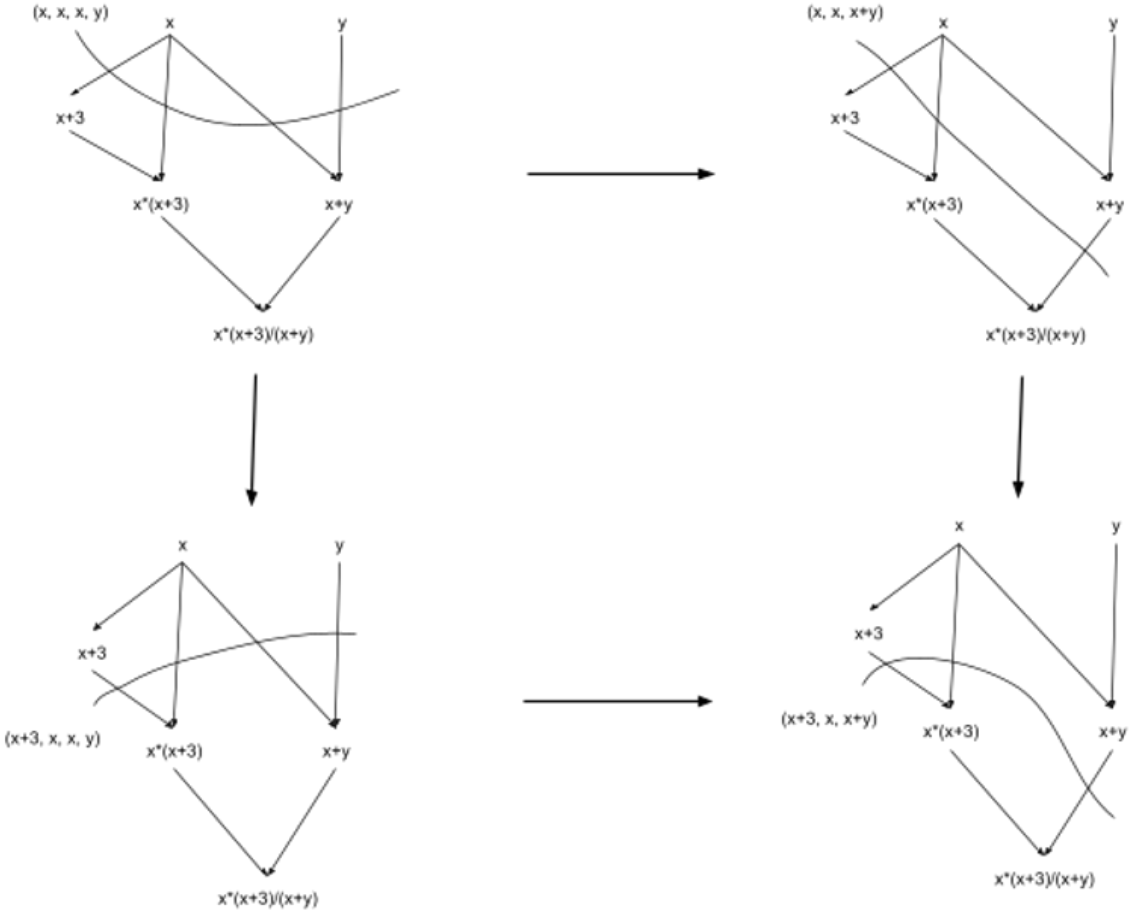
The state (x, x, x, y) gives the data carried by each edge we cut through. If we imagine that the computation at each node is performed by a different person, then at the very beginning of the computation, these would be the messages sent from our first two people (i.e. the people with our input values x = 2 and y = 5) to the people downstream. Alternatively, if a single CPU were performing this computation, it would probably save some memory by only storing x once rather than keeping three separate copies; more on that later.

From there, we have two possible next steps:

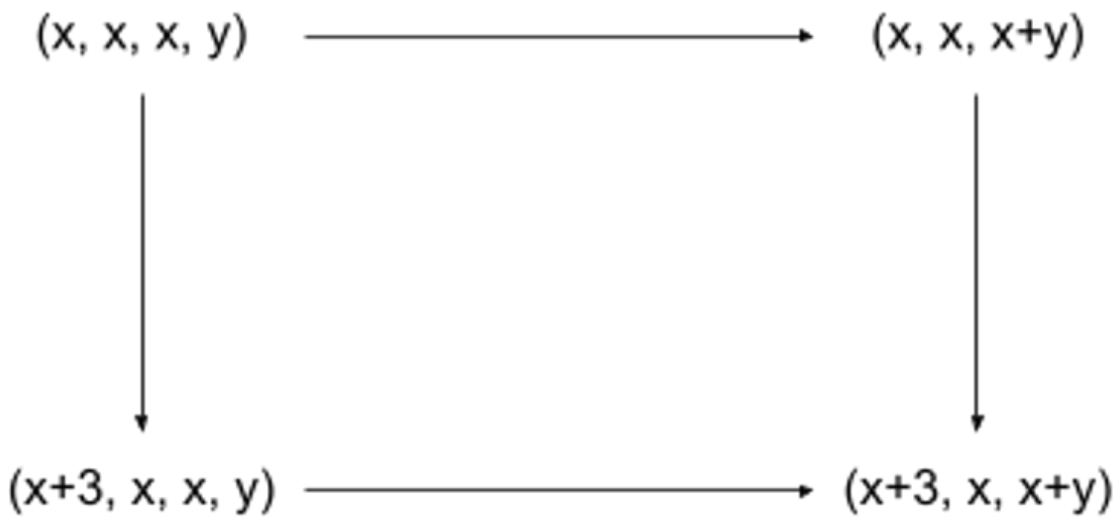


On the left, we compute x+3 first, so our state updates to (x+3, x, x, y). On the right, we compute x+y first, so our state updates to (x, x, x+y).

After performing either of these operations, we can perform the other, leaving us in the state (x+3, x, x+y). We can visualize this as two computation-paths in a computation-graph:

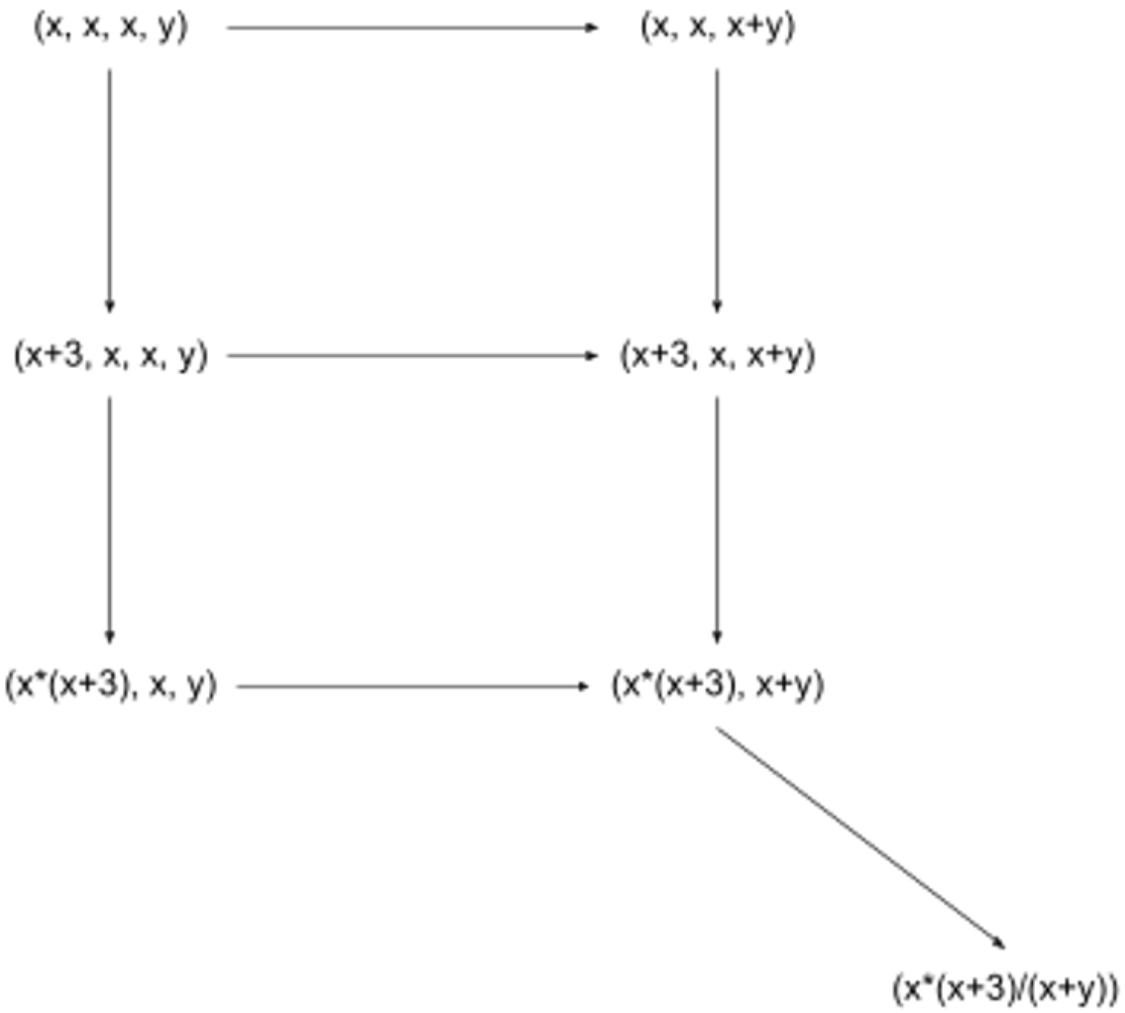


Or, dropping all the visuals of the circuit-cuts:



The two paths from (x, x, x, y) to (x+3, x, x+y) correspond to the two possible orders of the addition operation: x+3 followed by x+y, or x+y followed by x+3.

Here’s the full graph of computation states, with the multiplication and division operations added in.



Any allowed order of the operations in our original circuit corresponds to a path from the upper-left to lower-right in this computation-state graph. If that makes sense, then you've probably understood the basic idea. (If it's still a bit murky, try drawing the graph-cuts corresponding to the three states at the bottom of the diagram above, then walk through an evaluation of the circuit and consider which cut represents the state at each step.)

As A Category

Recall from the previous post that a category is a graph with some notion of equivalence between paths. If we have a graph in which the paths represent something interesting - e.g. our computation-state graph - then we can generate potentially-interesting categories by thinking about notions of "equivalence" between the paths.

One example: imagine that each node of our computation is handled by a different person, and the arrows in the circuit correspond to messages passed between people. If we have a limited number of messengers, then we might want to limit the maximum number of messages passed simultaneously - i.e. if we only have 4 messengers, then we'd want to pick a computation-path which never cuts through four arrows simultaneously. (Equivalently: we want a computation-path which never has more than 4 variables in its state.) When searching for such a computation-path, it might be useful to consider two paths "equivalent" if they start and end at the same node and have the same maximum number of messages.

With one small adjustment, we can make this example into something more realistically useful for e.g. writing a compiler. Rather than counting all arrows cut, we count the number of “distinct” arrows cut - i.e. the number of messages with different contents. Then (x, x, x, y) would only count as two, and (x+3, x, x, y) would count as three. When performing the computation on a CPU, this would be the number of memory cells we need to use simultaneously - so we’d consider two computation-paths equivalent if they use the same maximum number of memory cells.

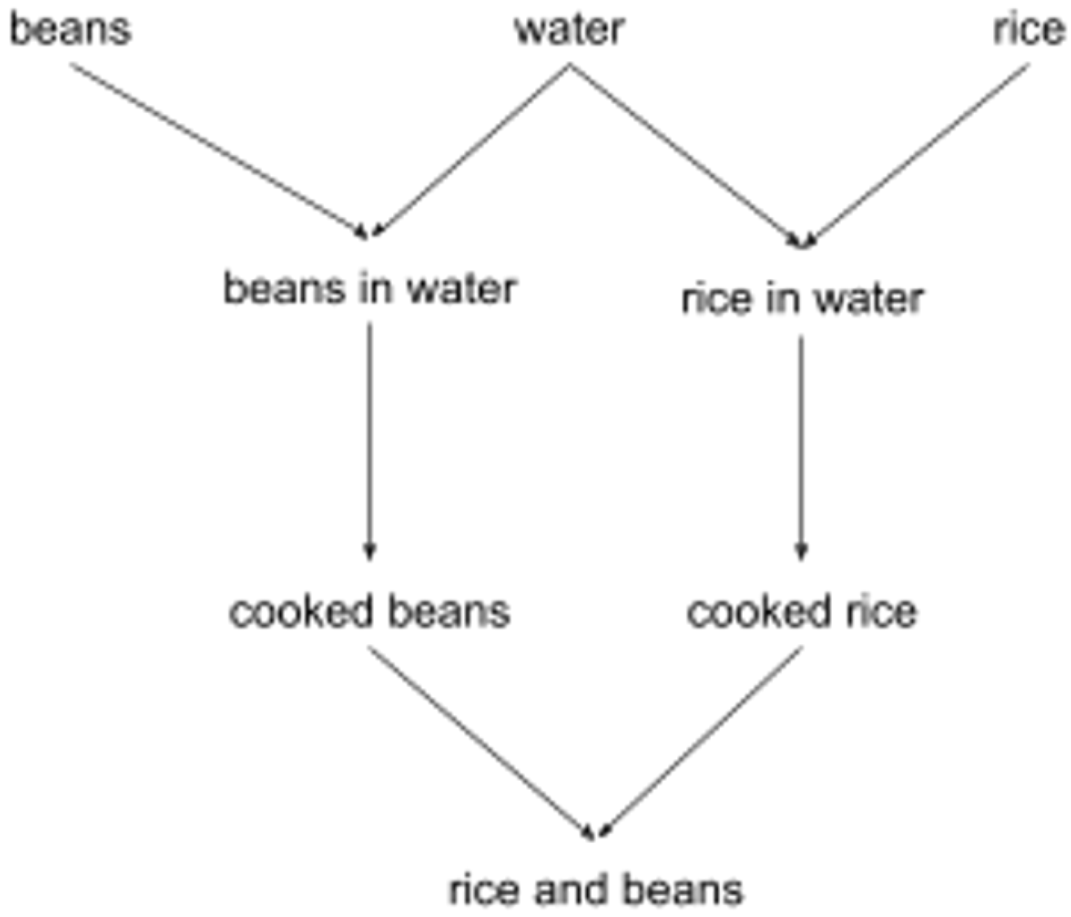
Of course, there are many other possibilities. There’s the trivial possibility: any two paths with the same start and end state are equivalent (i.e. we just need to find *some* computation-path, and don’t care which). Or the edges in the computation-graph could have some kind of costs associated with them, in which case we’d call two paths equivalent if they have the same cost - potentially quite nontrivial if e.g. it’s expensive to perform two multiplications back-to-back on a deeply pipelined processor, but cheaper if they’re not performed back-to-back.

Now imagine that we’re writing a compiler, and it needs to compile code where the computation-state graph lives in a space with thousands of dimensions and has exponentially many nodes, but most paths are equivalent to large numbers of other paths. I can see where it might be useful to have some mathematical constructions which can compress some of those equivalent paths - thus, category theory.

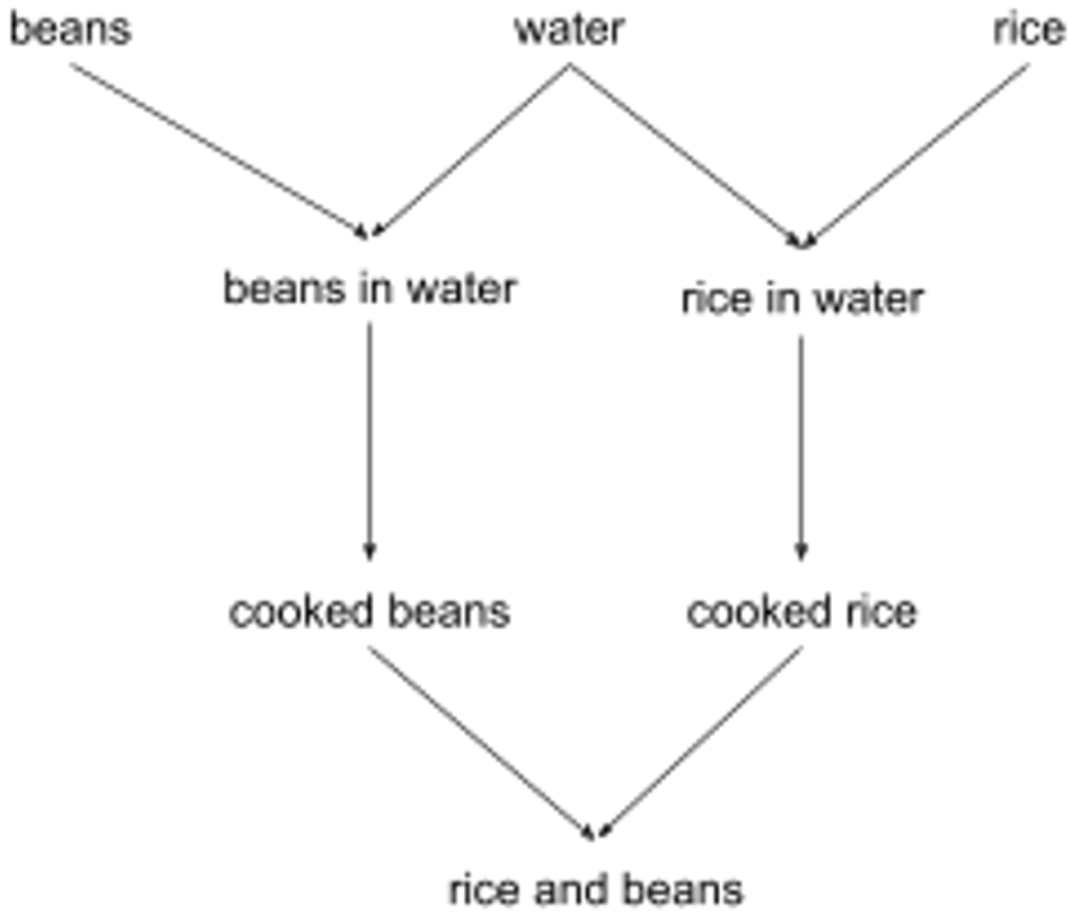
Generalization: Symmetric Monoidal Categories

The computation-state graph associated with a circuit is the prototypical example of a “symmetric monoidal category”. That’s an absolutely awful name, and I’m not going to explain where it comes from. But I will give one more example, which should be sufficient to illustrate the concept.

Suppose we’re cooking rice and beans. We add water to the beans and cook them. We also add water to the rice and cook that. Finally, we combine the rice and beans. Visually:

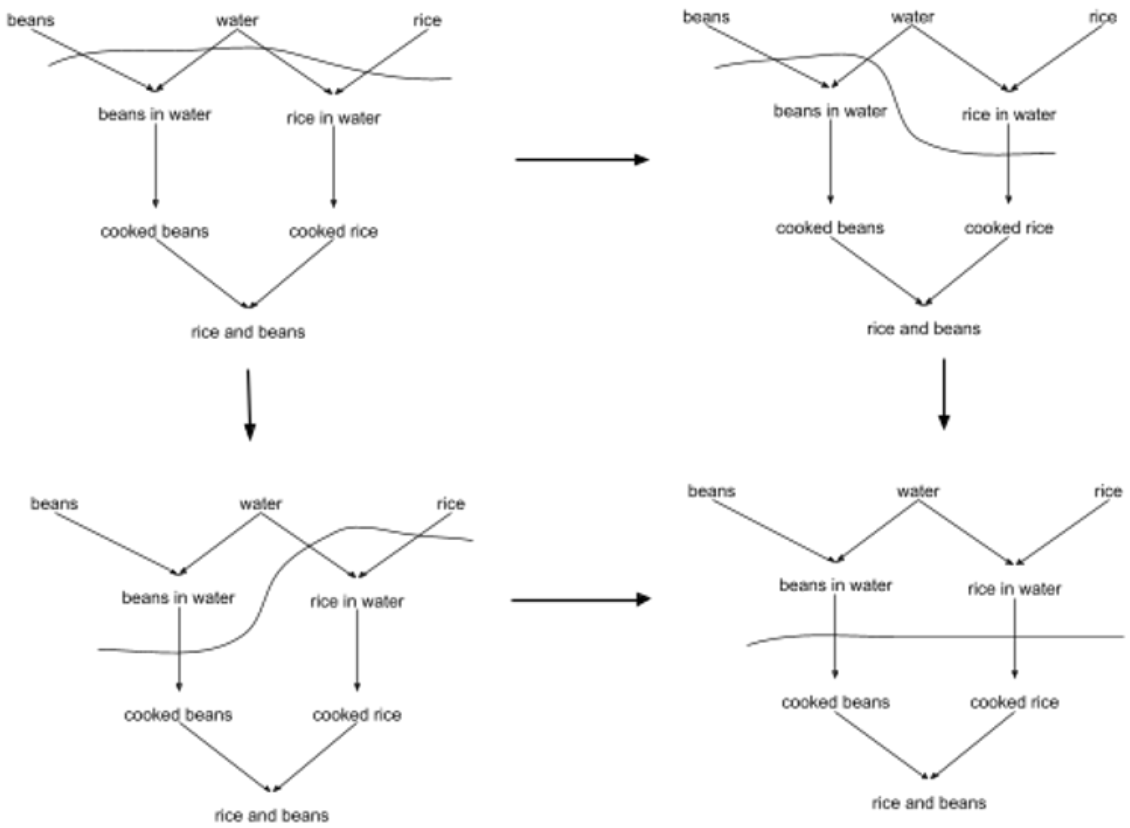


As before, we can use cuts in this graph to identify the state at any time. For instance:



This cut represents a state where the beans are done cooking, but we have not yet started the rice at all.

As before, we can create a state graph which shows possible paths between states:



In this case, the path through the lower left involves starting the beans first, while the path through the upper right involves starting the rice first. Both end up in the lower right state, in which both the beans and the rice are cooking.

Thinking about this state graph as a category, many of the same notions of equivalence from computation-states carry over nicely. For instance, the maximum number of arrows cut by a cooking-path might correspond to the number of items which need to be on the countertop simultaneously; as before, it's a measure of "how much workspace" is needed by a path. We could imagine expanding this to a large-scale model of economic production in some domain, with thousands of dimensions and exponentially many possible paths. As before, it would be useful to have standard tools to compress "equivalent" paths through the state graph.

Using GPT-N to Solve Interpretability of Neural Networks: A Research Agenda

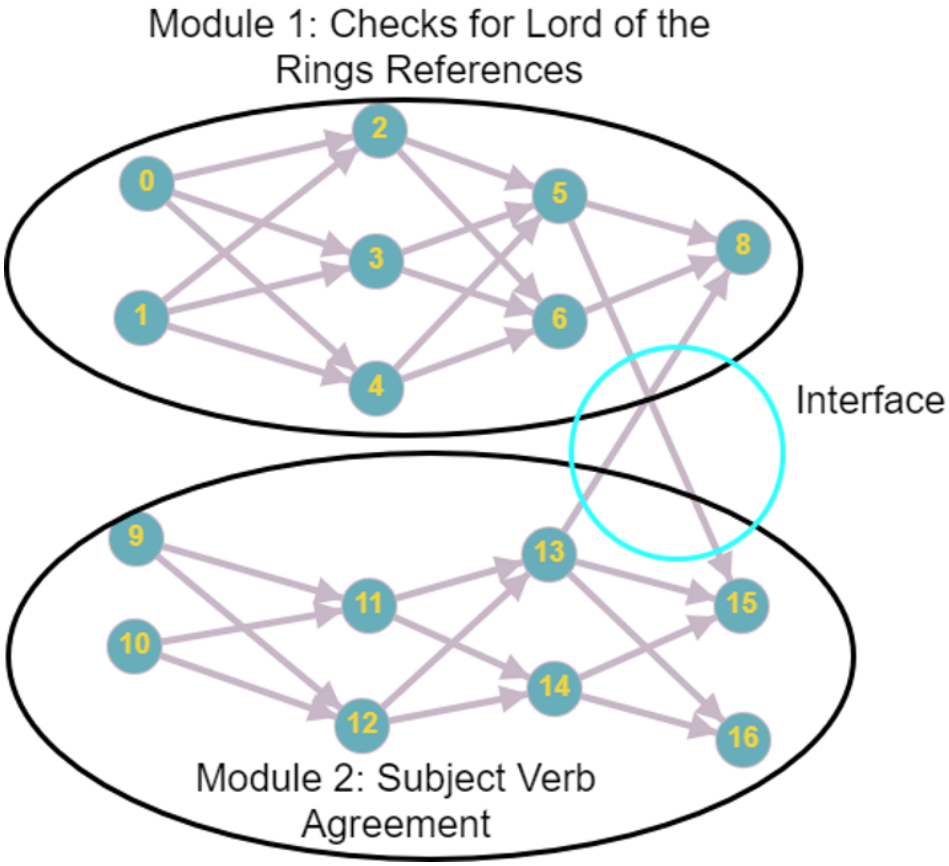
Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Tl;dr We are attempting to make neural networks (NN) modular, have GPT-N interpret each module for us, in order to catch mesa-alignment and inner-alignment failures.

Completed Project

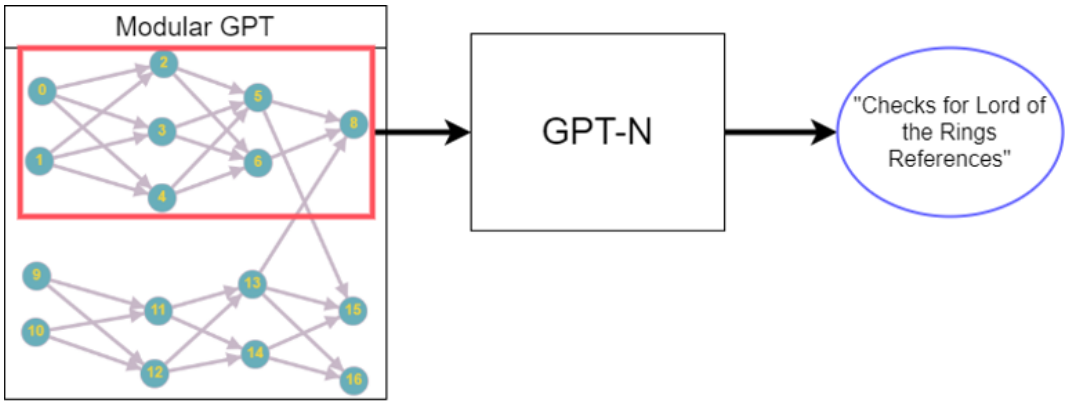
Train a neural net with an added loss term that enforces the sort of modularity that we see in well-designed software projects. To use this paper's informal definition of modularity

> a network is modular to the extent that it can be partitioned into sets of neurons where each set is strongly internally connected, but only weakly connected to other sets.



Example of a “Modular” GPT. Each module should be densely connected w/ relatively larger weights. Interfaces between modules should be sparsely connected w/ relatively smaller weights.

Once we have a Modular NN (for example, a GPT), we will use a normal GPT to map each module into a natural language description. Notice that there are two different GPT’s at work here.



GPT-N reads in each “Module” of the “Modular GPT”, outputting a natural language description for each module.

If successful, we could use GPT-N to interpret any modular NN in natural language. Not only should this help our understanding of what the model is doing, but it should also catch mesa-alignment and inner-alignment failures.

Cruxes

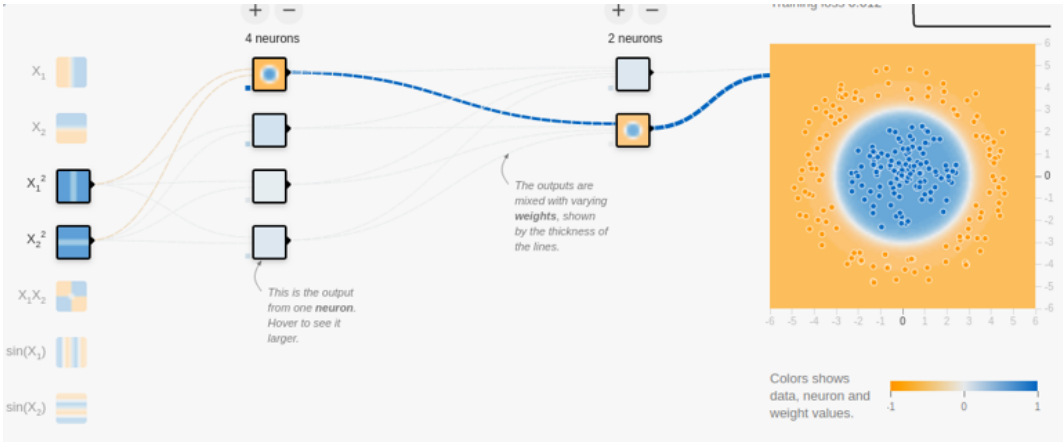
There are a few intuitions we have that go counter to other’s intuitions. Below is an elaboration of our thoughts and why we think this project could work.

Finding a Loss function that Induces Modularity

We currently think a Gomory-Hu Tree (GH Tree) captures the relevant information. We will initially convert a NN to a GH Tree to calculate the new loss function. This conversion will be computationally costly, though more progress can be made to calculate the loss function directly from the NN. See Appendix A for more details

Small NN’s are Human Interpretable

We're assuming humans can interpret small NN's, given enough time. A "Modular" NN is just a collection of small NN's connected by sparse weights. If humans could interpret each module in theory, then GPT-N could too. If humans can interpret the interfaces between each, then GPT-N could too.



Examples from NN Playground are readily interpretable (such as the above example).

GPT-3 can already turn comments into code. We don't expect the reverse case to be fundamentally harder, and neural nets can be interpreted as just another programming language.

Microscope AI has had some success in interpreting large NN's. These are NN's that should be much harder to interpret than modular NN's that we would be interpreting.

Technical Questions:

First question: Capabilities will likely be lost by adding a modularity loss term. Can we spot-check capability of GPT by looking at the loss of the original loss terms? Or would we need to run it through NLP metrics (like Winograd Schema Challenge questions)?

To create a modular GPT, we have two paths, but I'm unsure of which is better.

1. Train from scratch with modified loss
2. Train OpenAI's gpt-2 on more data, but with added loss term. The intuition here is that it's already capable, so optimizing for modularity starting here will preserve capabilities.

Help Wanted

If you are interested in the interpretability of GPT (even unrelated to our project), I can add you to a discord server full of GPT enthusiasts (just DM me). If you're interested in helping out our project specifically, DM me and we'll figure out a way to divvy up tasks.

Appendix A

Gomory-Hu Tree Contains Relevant Information on Modularity

Some readily accessible insights:

1. The size of the minimum cut between two neurons can be used to measure the size of the interface between their modules.
2. Call two graphs G and G’ on the same vertices equivalent if for every two u,v, the sizes of their minimum cuts are the same in G and G’. It turns out that there always exists a G’ which is a tree! (The Gomory-Hu tree.)
3. It turns out that the minimum cut between two neurons within a module never needs to expose the innards of another module.

Therefore, the Gomory-Hu tree probably contains all the information needed to calculate the loss term and the hierarchy of software modules.

Comparing Utilities

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

(This is a basic point about utility theory which many will already be familiar with. I draw some non-obvious conclusions which may be of interest to you even if you think you know this from the title -- but the main point is to communicate the basics. I'm posting it to the alignment forum because I've heard misunderstandings of this from some in the AI alignment research community.)

I will first give the basic argument that the utility quantities of different agents aren't directly comparable, and a few important consequences of this. I'll then spend the rest of the post discussing what to do when you need to compare utility functions.

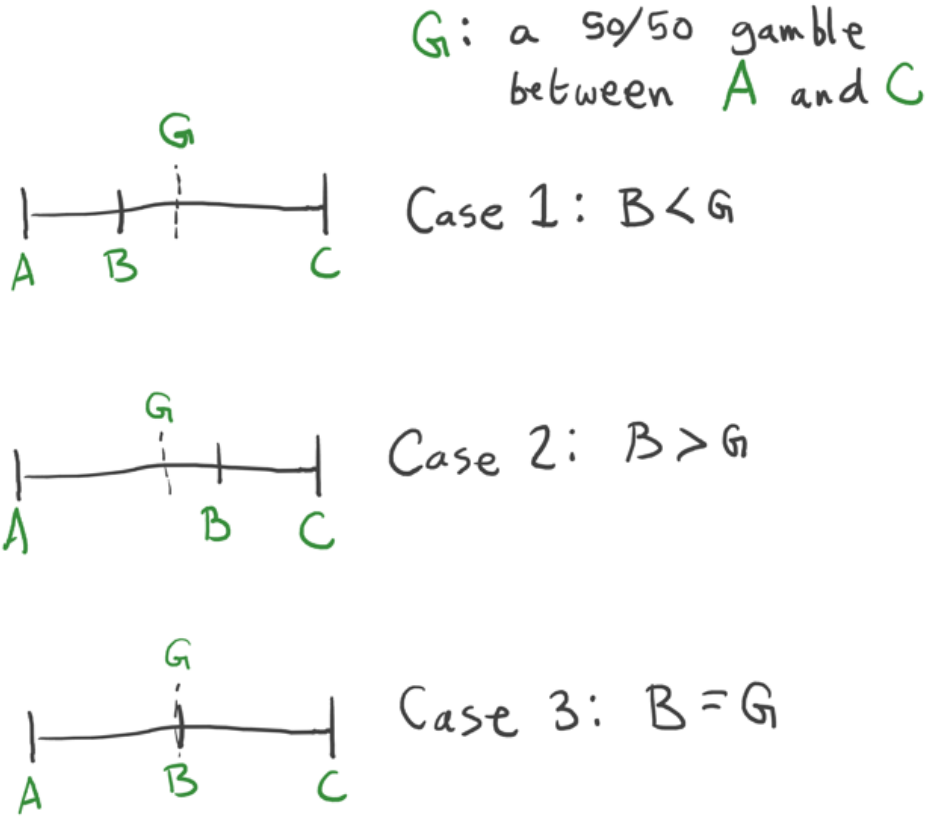
Utilities aren't comparable.

Utility isn't an ordinary quantity. A utility function is a device for expressing the preferences of an agent.

Suppose we have a notion of *outcome.** We could try to represent the agent's preferences between outcomes as an ordering relation: if we have outcomes A, B, and C, then one possible preference would be A<B<C.

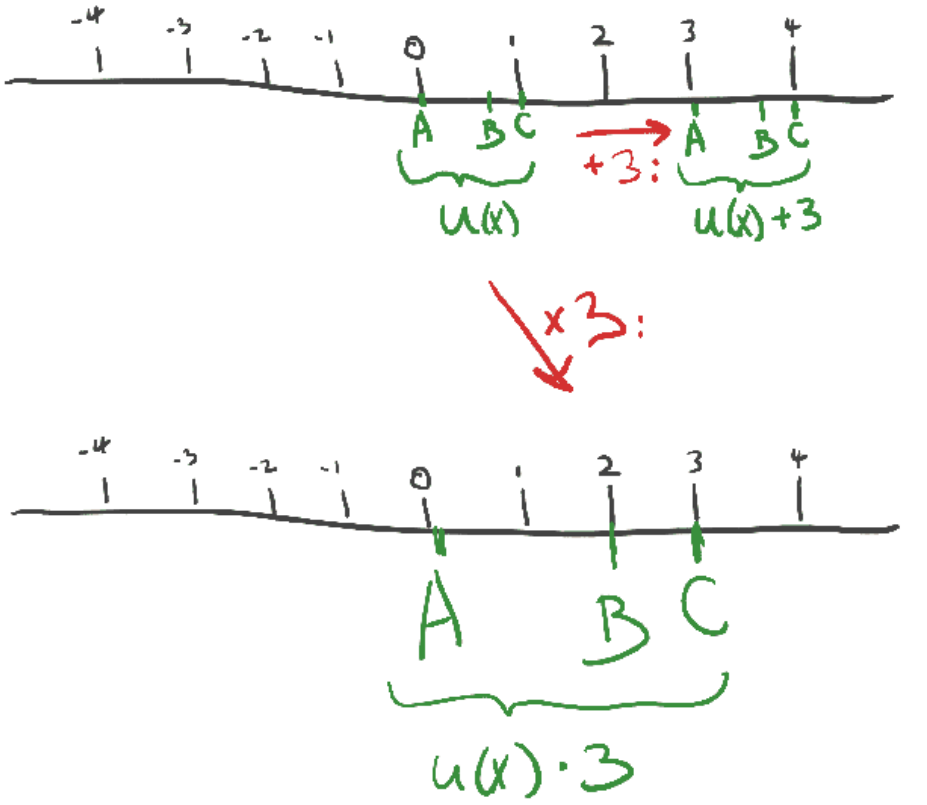
However, a mere ordering does not tell us how the agent would decide between *gambles,* ie, situations giving A, B, and C with some probability.

With just three outcomes, there is only one thing we need to know: is B closer to A or C, and by how much?



We want to construct a utility function U() which represents the preferences. Let's say we set U(A)=0 and U(C)=1. Then we can represent B=G as U(B)=1/2. If not, we would look for a different gamble which *does* equal B, and then set B's utility to the expected value of that gamble. By assigning real-numbered values to each outcome, we can fully represent an agent's preferences over gambles. (Assuming the VNM axioms hold, that is.)

But the initial choices U(A)=0 and U(C)=1 were arbitrary! We could have chosen any numbers so long as U(A)<U(C), reflecting the preference A<C. In general, a valid representation of our preferences U() can be modified into an equally valid U'() by adding/subtracting arbitrary numbers, or multiplying/dividing by positive numbers.



So it's just as valid to say someone's expected utility in a given situation is 5 or -40, provided you shift everything *else* around appropriately.

Writing ≈ to mean that two utility functions represent the same preferences, what we have in general is: $U_1(x) \approx U_2(x)$ if and only if $U_1(x) = aU_2 + b$. (I'll call a the ***multiplicative***

constant and b the ***additive constant***.)

This means that we can't directly compare the utility of two different agents. Notions of fairness should not directly say "everyone should have the same expected utility". Utilitarian ethics cannot directly maximize the sum of everyone's utility. Both of these operations should be thought of as a type error.

Some non-obvious consequences.

The game-theory term "zero sum" is a misnomer. You shouldn't directly think about the sum of the utilities.

In mechanism design, *exchangeable utility* is a useful assumption which is often needed in order to get nice results. The idea is that agents can give utils to each other, perhaps to compensate for unfair outcomes. This is *kind of* like assuming there's money which can be exchanged between agents. However, the non-comparability of utility should make this seem *really weird*. (There are also other disanalogies with money; for example, utility is closer to logarithmic in money, not linear.)

This could (should?) also make you suspicious of talk of "average utilitarianism" and "total utilitarianism". However, beware: only one kind of "utilitarianism" holds that the term "utility" in decision theory means the same thing as "utility" in ethics: namely, preference utilitarianism. Other kinds of utilitarianism can distinguish between these two types of utility. (For example, one can be a hedonic utilitarian without thinking that what everyone wants is happiness, if one isn't a preference utilitarian.)

Similarly, for preference utilitarians, talk of *utility monsters* becomes questionable. A utility monster is, supposedly, someone who gets much more utility out of resources than everyone else. For a hedonic utilitarian, it would be someone who experiences much deeper sadness and much higher heights of happiness. This person supposedly merits more resources than other people.

For a preference utilitarian, incomparability of utility means we can't simply posit such a utility monster. It's meaningless *a priori* to say that one person simply has much stronger preferences than another (in the utility function sense).

All that being said, we *can* actually compare utilities, sum them, exchange utility between agents, define utility monsters, and so on. We just need *more information.*

Comparing utilities.

The incomparability of utility functions ***doesn't mean*** we can't trade off between the utilities of different people.

I've heard the non-comparability of utility functions summarized as the thesis that we can't say anything meaningful about the relative value of one person's suffering vs another person's convenience. Not so! Rather, the point is just that *we need more assumptions in order to say anything.* The utility functions alone aren't enough.

Pareto-Optimality: The Minimal Standard

Comparing utility functions suggests putting them all onto one scale, such that we can trade off between them -- "this dollar does more good for Alice than it does for Bob". We formalize this by imagining that we have to decide policy for the whole group of people we're

considering (e.g., the whole world). We consider a *social choice function* which would make those decisions on behalf of everyone. Supposing it is VNM rational, its decisions must be comprehensible in terms of a utility function, too. So the problem reduces to combining a bunch of individual utility functions, to get one big one.

So, how do we go about combining the preferences of many agents into one?

The first and most important concept is the ***pareto improvement:*** *our social choice function should endorse changes which benefit someone and harm no one.* An option which allows no such improvements is said to be ***Pareto-optimal.***

We might also want to consider ***strict Pareto improvements:*** *a change which benefits everyone.* (An option which allows no strict Pareto improvements is ***weakly Pareto-optimal.***) Strict Pareto improvements can be more relevant in a bargaining context, where you need to give everyone something in order to get them on board with a proposal -- otherwise they may judge the improvement as unfairly favoring others. However, in a bargaining context, individuals may refuse even a strict Pareto improvement due to fairness considerations.

In either case, a version of Harsanyi's utilitarianism Theorem implies that the utility of our social choice function *can be understood as some linear combination of the individual utility functions.*

So, pareto-optimal social choice functions can always be understood by:

1. Choosing a scale for everyone's utility function -- IE, set the multiplicative constant. (If the social choice function is only weakly Pareto optimal, some of the multiplicative constants might turn out to be zero, totally cancelling out someone's involvement. Otherwise, they can all be positive.)
2. Adding all of them together.

(Note that the *additive constant* doesn't matter -- shifting a person's utility function up or down doesn't change what decisions will be endorsed by the sum. However, it ***will*** matter for some other ways to combine utility functions.)

This is nice, because we can always combine everything linearly! We just have to set things to the right scale and then sum everything up.

However, it's far from the end of the story. How do we choose multiplicative constants for everybody?

Variance Normalization: Not Too Exploitable?

We could set the constants any way we want... totally subjective estimates of the worth of a person, draw random lots, etc. But we do typically want to represent some notion of fairness.

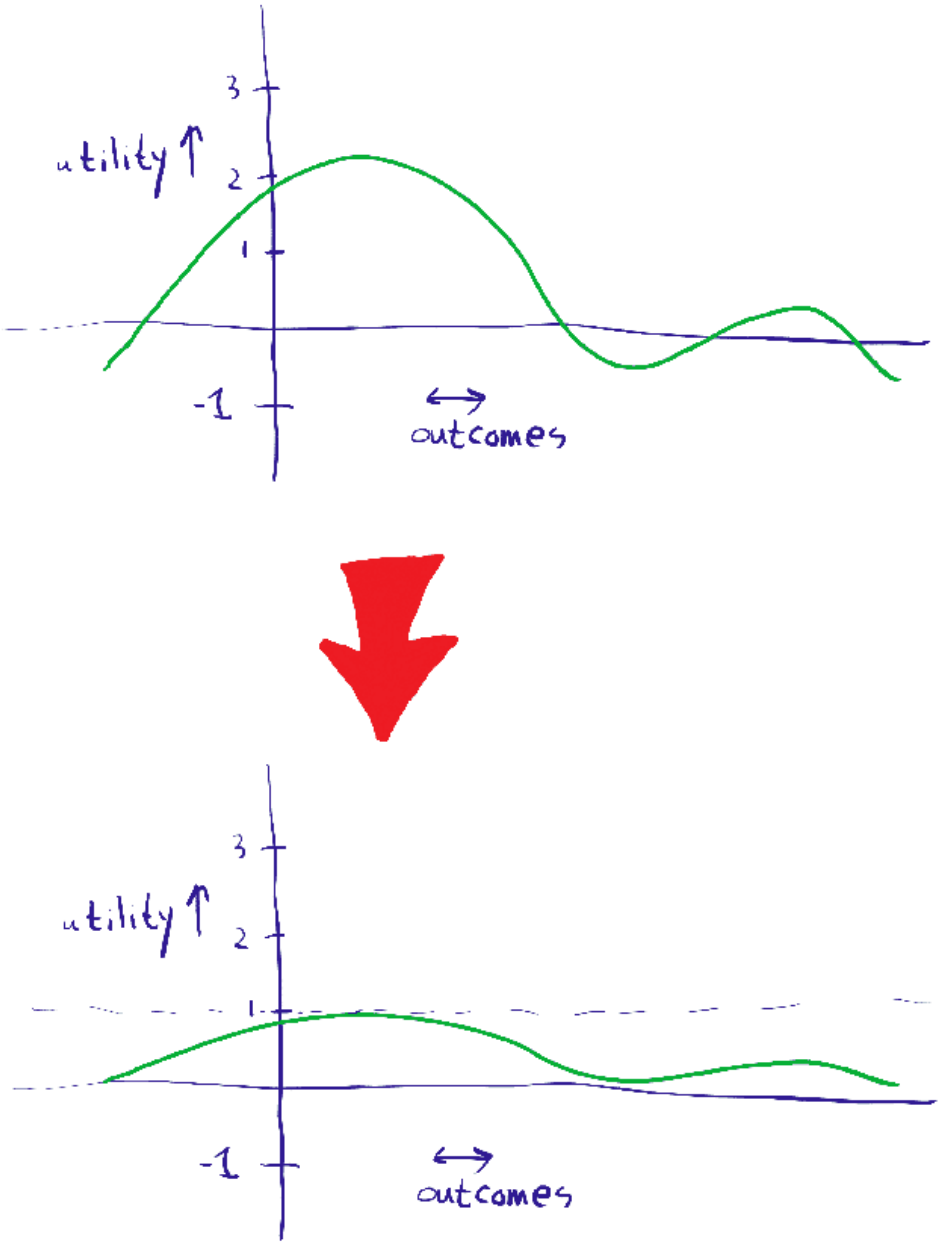
We said in the beginning that the problem was, a utility function $U(x)$ has many equivalent representations $aU(x) + b$. We can address this as a problem of ***normalization:*** we want to take a U and put it into a canonical form, getting rid of the choice between equivalent representations.

One way of thinking about this is ***strategy-proofness***. A utilitarian collective should not be vulnerable to members strategically claiming that their preferences are stronger (larger b), or that they should get more because they're worse off than everyone (smaller a -- although, remember that we haven't talked about any setup which actually cares about that, yet).

Warm-Up: Range Normalization

Unfortunately, some obvious ways to normalize utility functions are not going to be strategy-proof.

One of the simplest normalization techniques is to squish everything into a specified range, such as [0,1]:



This is analogous to range voting: everyone reports their preferences for different outcomes on a fixed scale, and these all get summed together in order to make decisions.

If you're an agent in a collective which uses range normalization, then you may want to strategically mis-report your preferences. In the example shown, the agent has a big hump around outcomes they like, and a small hump on a secondary "just OK" outcome. The agent might want to get rid of the second hump, forcing the group outcome into the more favored region.

I believe that in the extreme, the optimal strategy for range voting is to choose some utility threshold. Anything below that threshold goes to zero, feigning maximal disapproval of the outcome. Anything above the threshold goes to one, feigning maximal approval. In other words, under strategic voting, range voting becomes approval voting (range voting where the only options are zero and one).

If it's not possible to mis-report your preferences, then the incentive becomes to *self-modify to literally have these extreme preferences.* This could perhaps have a real-life analogue in political outrage and black-and-white thinking. If we use this normalization scheme, that's the closest you can get to being a utility monster.

Variance Normalization

We'd *like* to avoid *any* incentive to misrepresent/modify your utility function. Is there a way to achieve that?

Owen Cotton-Barratt discusses different normalization techniques in illuminating detail, and argues for *variance normalization:* divide utility functions by their variance, making the variance one. (*Geometric reasons for normalizing variance to aggregate preferences*, O Cotton-Barratt, 2013.) Variance normalization is strategy-proof under the assumption that everyone participating in an election shares beliefs about how probable the different outcomes are! (Note that *variance of utility* is only well-defined under some assumption about *probability of outcome.*) That's pretty good. It's probably the best we can get, in terms of strategy-proofness of voting. Will MacAskill also argues for variance normalization in the context of normative uncertainty (*Normative Uncertainty,* Will MacAskill, 2014).

Intuitively, variance normalization directly addresses the issue we encountered with range normalization: an individual attempts to make their preferences "loud" by extremizing everything to 0 or 1. This increases variance, so, is directly punished by variance normalization.

However, Jameson Quinn, LessWrong's resident voting theory expert, has warned me rather strongly about variance normalization.

1. The assumption of shared beliefs about election outcomes is far from true in practice. Jameson Quinn tells me that, in fact, the strategic voting incentivized by quadratic voting is *particularly bad* amongst normalization techniques.
2. Strategy-proofness isn't, after all, the final arbiter of the quality of a voting method. The final arbiter should be something like the utilitarian quality of an election's outcome. This question gets a bit weird and recursive in the current context, where I'm using elections as an analogy to ask how we should define utilitarian outcomes. But the point still, to some extent, stands.

I didn't understand the full justification behind his point, but I came away thinking that range normalization was probably better in practice. After all, it reduces to approval voting, which is actually a pretty good form of voting. But if you want to do the best we can with the state of voting theory, Jameson Quinn suggested 3-2-1 voting. (I don't think 3-2-1 voting gives us any nice theory about how to combine utility functions, though, so it isn't so useful for our purposes.)

Open Question: *Is there a variant of variance normalization which takes differing beliefs into account, to achieve strategy-proofness (IE honest reporting of utility)?*

Anyway, so much for normalization techniques. These techniques ignore the broader context. They attempt to be fair and even-handed *in the way we choose the multiplicative and additive constants.* But we could also explicitly try to be fair and even-handed *in the way we choose between Pareto-optimal outcomes*, as with this next technique.

Nash Bargaining Solution

It's important to remember that the Nash bargaining solution is a solution *to the Nash bargaining problem*, which isn't quite our problem here. But I'm going to gloss over that. Just imagine that we're setting the social choice function through a massive negotiation, so that we can apply bargaining theory.

Nash offers a very simple solution, which I'll get to in a minute. But first, a few words on how this solution is derived. Nash provides two seperate justifications for his solution. The first is a game-theoretic derivation of the solution as an especially robust Nash equilibrium. I won't detail that here; I quite recommend his original paper (*The Bargaining Problem,* 1950); but, just keep in mind that there is at least some reason to expect selfishly rational agents to hit upon this particular solution. The second, unrelated justification is an axiomatic one:

1. *Invariance to equivalent utility functions.* This is the same motivation I gave when discussing normalization.
2. *Pareto optimality.* We've already discussed this as well.
3. *Independence of Irrelevant Alternatives (IIA).* This says that we shouldn't change the outcome of bargaining by removing options which won't ultimately get chosen anyway. This isn't even technically one of the VNM axioms, but it *essentially* is -- the VNM axioms are posed for binary preferences (a > b). IIA is the assumption we need to break down multi-choice preferences to binary choices. We can justify IIA with a kind of money pump.
4. *Symmetry.* This says that the outcome doesn't depend on the order of the bargainers; we don't prefer Player 1 in case of a tie, or anything like that.

Nash proved that *the only way to meet these four criteria* is to maximize the **product** of gains from cooperation. More formally, choose the outcome x which maximizes:

$$(U_1(x) - U_1(d))(U_2(x) - U_2(d))$$

The d here is a "status quo" outcome. You can think of this as what happens if the bargaining fails. This is sometimes called a "threat point", since strategic players should carefully set what they do *if negotiation fails* so as to maximize their bargaining position. However, you might also want to rule that out, forcing d to be a Nash equilibrium in the hypothetical game where there is no bargaining opportunity. As such, d is also known as the *best alternative to negotiated agreement (BATNA)*, or sometimes the "disagreement point" (since it's what players get if they can't agree). We can think of subtracting out U(d) as just a way of adjusting the additive constant, in which case we really are just maximizing the product of utilities. (The BATNA point is always (0,0) after we subtract out things that way.)

The Nash solution differs significantly from the other solutions considered so far.

1. Maximize the *product??* Didn't Harsanyi's theorem guarantee we only need to worry about sums?

2. This is the first proposal where the additive constants matter. Indeed, now the *multiplicative* constants are the ones that don't matter!
3. Why wouldn't *any* utility-normalization approach satisfy those four axioms?

Last question first: how do normalization approaches violate the Nash axioms?

Well, both range normalization and variance normalization violate IIA! If you remove one of the possible outcomes, the normalization may change. This makes the social choice function display inconsistent preferences across different scenarios. (But how bad is that, really?)

As for why we can get away with maximizing the product, rather than the sum:

The Pareto-optimality of Nash's approach guarantees that it *can be seen* as maximizing a linear function of the individual utilities. So Harsanyi's theorem is still satisfied. However, Nash's solution points to a very *specific* outcome, which Harsanyi doesn't do for us.

Imagine you and me are trying to split a dollar. If we can't agree on how to split it, then we'll end up destroying it (ripping it during a desperate attempt to wrestle it from each other's hands, obviously). Thankfully, John Nash is standing by, and we each agree to respect his judgement. No matter which of us claims to value the dollar more, Nash will allocate 50 cents to each of us.

Harsanyi happens to see this exchange, and explains that Nash has chosen a social choice function which normalized our utility functions to be equal to each other. That's the only way Harsanyi can explain the choice made by Nash -- the value of the dollar was precisely tied between you and me, so a 50-50 split was as good as any other outcome. Harsanyi's justification is indeed *consistent* with the observation. But why, then, did Nash choose 50-50 *precisely?* 49-51 would have had exactly the same collective utility, as would 40-60, or any other split!

Hence, Nash's principle is far more useful than Harsanyi's, even though Harsanyi can justify any rational outcome retrospectively.

However, Nash does rely somewhat on that pesky IIA assumption, whose importance is perhaps not so clear. Let's try getting rid of that.

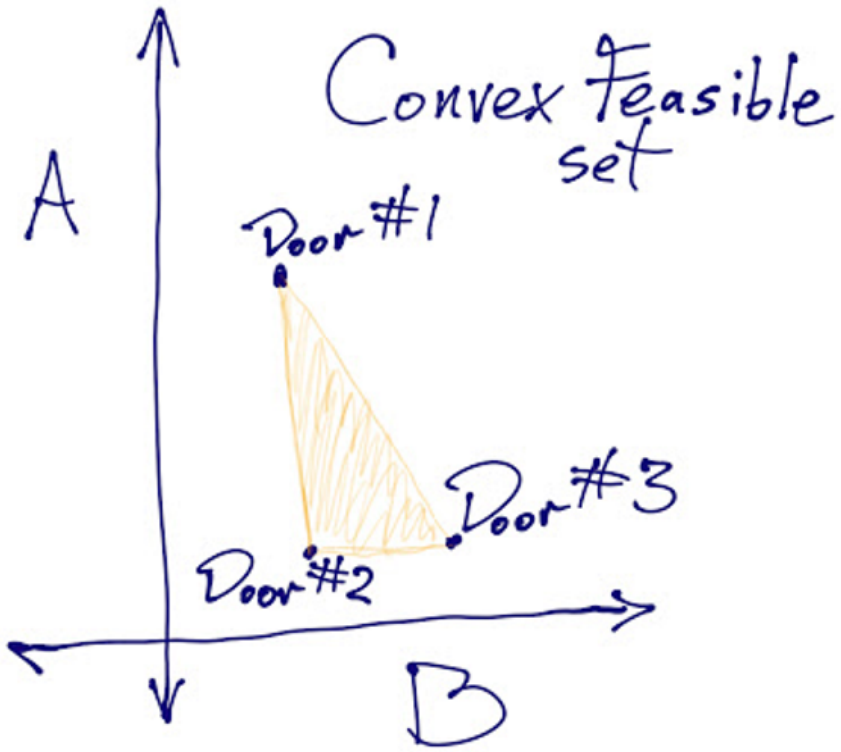
Kalai–Smorodinsky

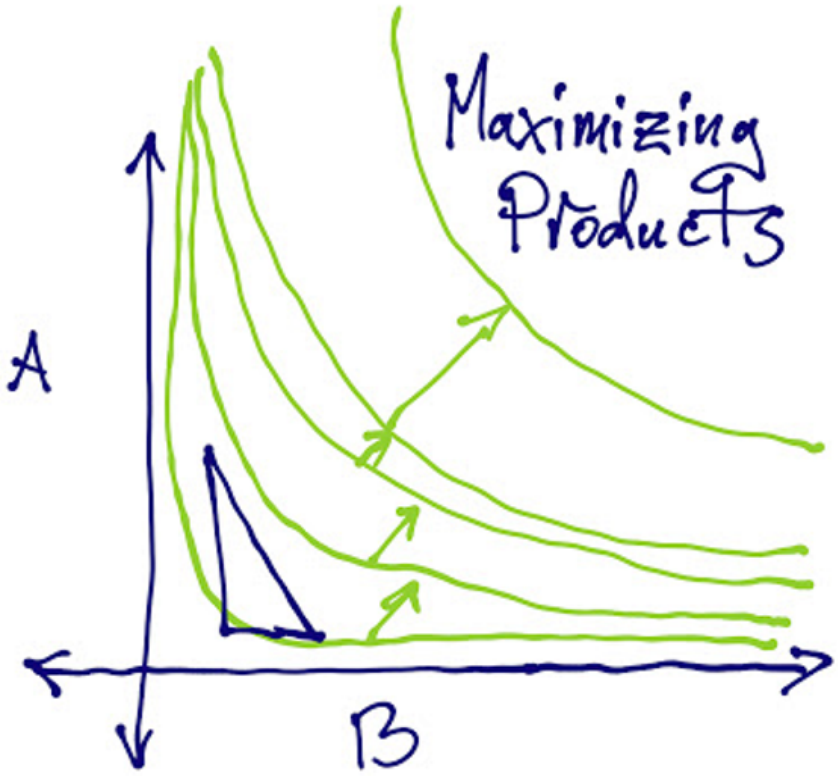
Although the Nash bargaining solution is the most famous, there are other proposed solutions to Nash's bargaining problem. I want to mention just one more, Kalai-Smorodinsky (I'll call it KS).

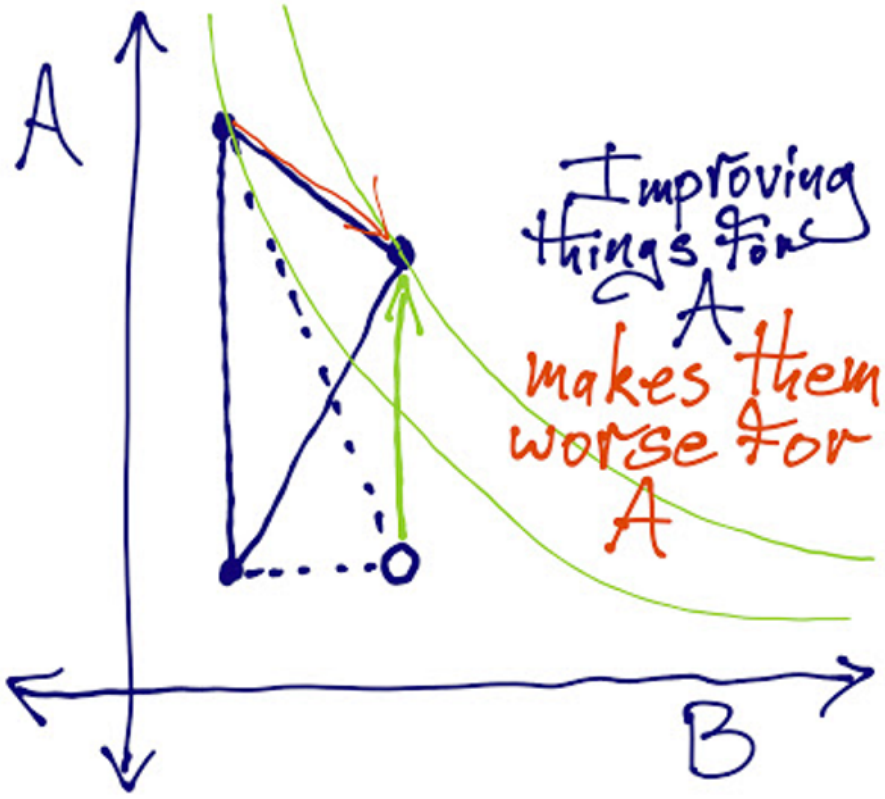
KS throws out IIA as irrelevant. After all, the set of alternatives *will* affect bargaining. Even in the Nash solution, the set of alternatives may have an influence by changing the BATNA! So perhaps this assumption isn't so important.

KS instead adds a *monotonicity* assumption: being in a better position should never make me worse off after bargaining.

Here's an illustration, due to Daniel Demski, of a case where Nash bargaining fails monotonicity:


Convex feasible set
A
Door #1
Door #3
Door #2
B

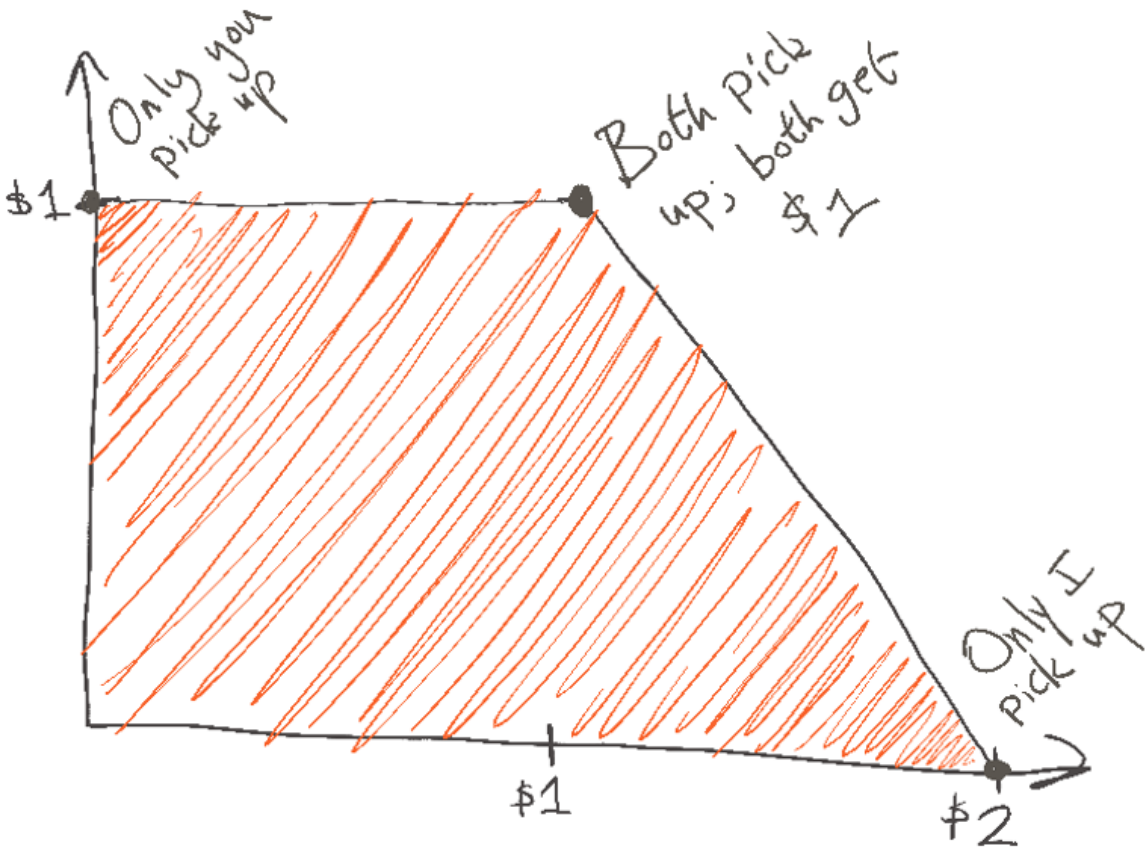

Maximizing
Products
A
B



I'm not that sure monotonicity really should be an axiom, but it does kind of suck to be in an apparently better position and end up worse off for it. Maybe we could relate this to strategy-proofness? A little? Not sure about that.

Let's look at the formula for KS bargaining.

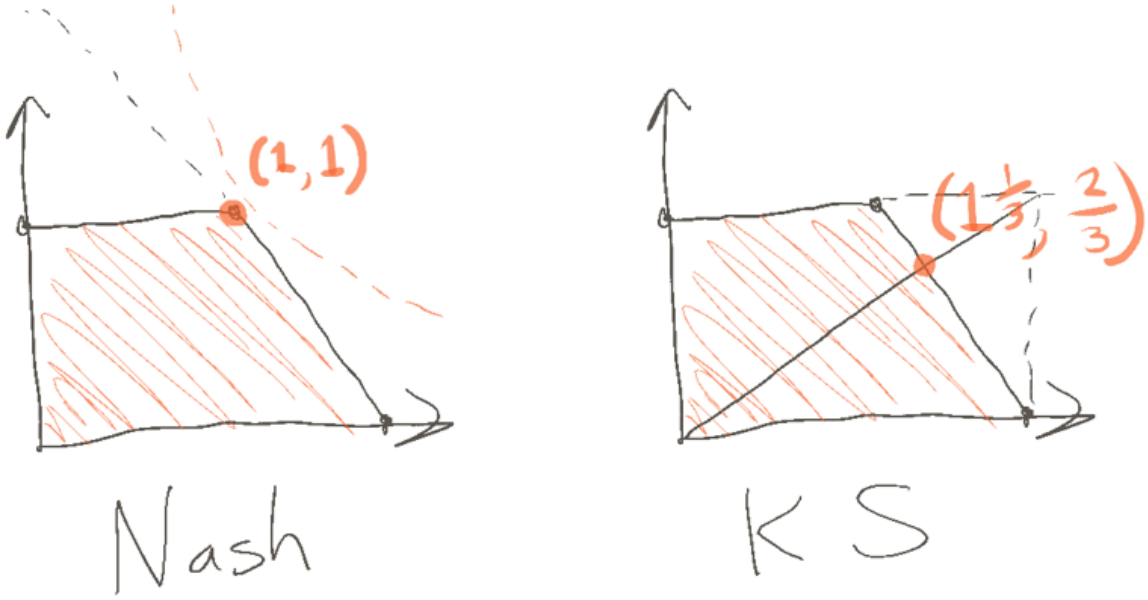
Suppose there are a couple of dollars on the ground: one which you'll walk by first, and one which I'll walk by. If you pick up your dollar, you can keep it. If I pick up my dollar, I can keep mine. But also, if you *don't* pick up yours, then I'll eventually walk by it and can pick it up. So we get the following:



(The box is filled in because we can also use mixed strategies to get values intermediate between any pure strategies.)

Obviously in the real world we just both pick up our dollars. But, let's suppose we bargain about it, just for fun.

The way KS works is, you look at the maximum *one* player can get (you can get $1), and the maximum the *other* player could get (I can get $2). Then, although we can't usually jointly achieve those payoffs (I can't get $2 at the same time as you get $1), KS bargaining insists we achieve the same *ratio* (I should get twice as much as you). In this case, that means I get $1.33, while you get $0.66. We can visualize this as drawing a bounding box around the feasible solutions, and drawing a diagonal line. Here's the Nash and KS solutions side by side:



As in Daniel's illustrations, we can visualize maximizing the product as drawing the largest hyperbola we can that still touches the orange shape. (Orange dotted line.) This suggests that we each get $1; exactly the same solution as Nash would give for splitting $2. (The black dotted line illustrates how we'd continue the feasible region to represent a dollar-splitting game, getting the full triangle rather than a chopped off portion.) Nash doesn't care that one of us can do better than the other; it just looks for the most equal division of funds possible, since that's how we maximize the product.

KS, on the other hand, cares what the max possible is for both of us. It therefore suggests that you give up some of your dollar to me.

I suspect most readers will ***not*** find the KS solution to be more intuitively appealing?

Note that the KS monotonicity property does NOT imply the desirable-sounding property "if there are more opportunities for good outcomes, everyone gets more or is at least not worse off." (I mention this mainly because I initially misinterpreted KS's monotonicity property this way.) In my dollar-collecting example, KS bargaining makes you worse off simply because there's an opportunity for me to take your dollar if you don't.

Like Nash bargaining, KS bargaining ignores multiplicative constants on utility functions, and can be seen as normalizing additive constants by treating d as (0,0). (Note that, in the illustration, I assumed d is chosen as (minimal achievable for one player, minimal achievable for the other). this need not be the case in general.)

A peculiar aspect of KS bargaining is that it doesn't really give us an obvious quantity to maximize, unlike Nash or Harsanyi. It only describes the optimal point. This seems far less practical, for realistic decision-making.

OK, so, should we use bargaining solutions to compare utilities?

My intuition is that, because of the need to choose the BATNA point d, bargaining solutions end up rewarding destructive threats in a disturbing way. For example, suppose that we are playing the dollar-splitting game again, except that I can costlessly destroy $20 of your money, so d now involves both the destruction of the $1, and the destruction of $20. Nash

bargaining now hands the entire dollar to me, because you are "up $20" in that deal, so the fairest possible outcome is to give me the $1. KS bargaining splits things up a little, but I still get most of the dollar.

If utilitarians were to trade off utilities that way in the real world, it would benefit powerful people, especially those willing to exploit their power to make credible threats. If X can take everything away from Y, then Nash bargaining sees everything Y has as already counting toward "gains from trade".

As I mentioned before, sometimes people try to define BATNAs in a way which excludes these kinds of threats. However, I see this as ripe for strategic utility-spoofing (IE, lying about your preferences, or self-modifying to have more advantageous preferences).

So, this might favor normalization approaches.

On the other hand, Nash and KS both do way better in the split-the-dollar game than any normalization technique, because they can optimize for fairness of outcome, rather than just fairness of multiplicative constants chosen to compare utility functions with.

Is there any approach which combines the advantages of bargaining and normalization??

Animals, etc.

An essay on utility comparison would be incomplete without at least mentioning the problem of animals, plants, and so on.

- Option one: some cutoff for "moral patients" is defined, such that a utilitarian only considers preferences of agents who exceed the cutoff.
- Option two: some more continuous notion is selected, such that we care more about some organisms than others.

Option two tends to be more appealing to me, despite the non-egalitarian implications (e.g., if animals differ on this spectrum, than humans could have some variation as well).

As already discussed, bargaining approaches do seem to have this feature: animals would tend to get less consideration, because they've got less "bargaining power" (they can do less harm to humans than humans can do to them). However, this has a distasteful might-makes-right flavor to it.

This also brings to the forefront the question of how we view something as an agent. Something like a plant might have quite deterministic ways of reacting to environmental stimulus. Can we view it as making choices, and thus, as having preferences? Perhaps "to some degree" -- if such a degree could be defined, numerically, it could factor into utility comparisons, giving a formal way of valuing plants and animals *somewhat,* but "not too much".

Altruistic agents.

Another puzzling case, which I think needs to be handled carefully, is accounting for the preferences of altruistic agents.

Let's proceed with a simplistic model where agents have "personal preferences" (preferences which just have to do with themselves, in some sense) and "***cofrences***" (co-preferences; preferences having to do with other agents).

Here's an agent named Sandy:

Sandy			
Personal Preferences		Cofrences	
Candy	+.1	Alice	+.1
Pizza	+.2	Bob	-.2
Rainbows	+10	Cathy	+.3
Kittens	-20	Dennis	+.4

The cofrences represent coefficients on other agent's utility functions. Sandy's preferences are supposed to be understood as a utility function representing Sandy's *personal* preferences, plus a weighted sum of the utility functions of Alice, Bob, Cathy, and Dennis. (Note that the weights can, hypothetically, be negative -- for example, screw Bob.)

The first problem is that utility functions are not comparable, so we have to say more before we can understand what "weighted sum" is supposed to mean. But suppose we've chosen some utility normalization technique. There are still other problems.

Notice that we can't totally define Sandy's utility function until we've defined Alice's, Bob's, Cathy's, and Dennis'. But any of those four might have cofrences which involve Sandy, as well!

Suppose we have Avery and Briar, two lovers who "only care about each other" -- their only preference is a cofrence, which places 1.0 value on the other's utility function. We could ascribe *any values at all* to them, so long as they're both the same!

With some technical assumptions (something along the lines of: your cofrences always sum to less than 1), we can ensure a unique fixed point, eliminating any ambiguity from the interpretation of cofrences. However, I'm skeptical of just taking the fixed point here.

Suppose we have five siblings: Primus, Secundus, Tertius, Quartus, et Quintus. All of them value each other at .1, except Primus, who values all siblings at .2.

If we simply take the fixed point, Primus is going to get the short end of the stick all the time: because Primus cares about everyone else more, everyone else cares about Primus' personal preferences *less* than anyone else's.

Simply put, I don't think more altruistic individuals should be punished! In this setup, the "utility monster" is the perfectly selfish individual. Altruists will be scrambling to help this person while the selfish person does nothing in return.

A different way to do things is to interpret cofrences as *integrating only the personal preferences of the other person.* So Sandy wants to help Alice, Cathy, and Dennis (and harm Bob), but does *not* automatically extend that to wanting to help any of their friends (or harm Bob's friends).

This is a little weird, but gives us a more intuitive outcome in the case of the five siblings: Primus will more often be voluntarily helpful to the other siblings, but the other siblings won't be prejudice *against* the personal preferences of Primus when weighing between their various siblings.

I realize altruism isn't *exactly* supposed to be like a bargain struck between selfish agents. But if I think of utilitarianism like a coalition of all agents, then I don't want it to punish the (selfish component of) the most altruistic members. It seems like utilitarianism should have better incentives than that?

(Try to take this section as more of a problem statement and less of a solution. Note that the concept of *cofrence* can include, more generally, preferences such as "I want to be better off

than other people" or "I don't want my utility to be too different from other people's in either direction".)

Utility monsters.

Returning to some of the points I raised in the "non-obvious consequences" section -- now we can see how "utility monsters" are/aren't a concern.

On my analysis, a utility monster is just an agent who, according to your metric for comparing utility functions, has a very large influence on the social choice function.

This might be a bug, in which case you should reconsider how you are comparing utilities. But, since you've hopefully chosen your approach carefully, it could also not be a bug. In that case, you'd want to bite the bullet fully, defending the claim that such an agent should receive "disproportionate" consideration. Presumably this claim could be backed up, on the strength of your argument for the utility-comparison approach.

Average utilitarianism vs total utilitarianism.

Now that we have given some options for utility comparison, can we use them to make sense of the distinction between average utilitarianism and total utilitarianism?

No. Utility comparison doesn't really help us there.

The average vs total debate is a debate about population ethics. Harsanyi's utilitarianism theorem and related approaches let us think about altruistic policies for a fixed set of agents. They don't tell us how to think about a set which changes over time, as new agents come into existence.

Allowing the set to vary over time like this feels similar to allowing a single agent to change its utility function. There is no rule against this. An agent can prefer to have different preferences than it does. A collective of agents can prefer to extend its altruism to new agents who come into existence.

However, I see no reason why population ethics needs to be *simple*. We can have relatively complex preferences here. So, I don't find paradoxes such as the Repugnant Conclusion to be especially concerning. To me there's just this complicated question about what everyone collectively wants for the future.

One of the basic questions about utilitarianism shouldn't be "average vs total?". To me, this is a type error. It seems to me, more basic questions for a (preference) utilitarian are:

- How do you combine individual preferences into a collective utility function?
    - How do you compare utilities between people (and animals, etc)?
        - Do you care about an "objective" solution to this, or do you see it as a subjective aspect of altruistic preferences, which can be set in an unprincipled way?
        - Do you range-normalize?
        - Do you variance-normalize?
        - Do you care about strategy-proofness?
        - How do you evaluate the bargaining framing? Is it relevant, or irrelevant?
        - Do you care about Nash's axioms?
        - Do you care about monotonicity?

        - What distinguishes humans from animals and plants, and how do you use it in utility comparison? Intelligence? Agenticness? Power? Bargaining position?
    - How do you handle cofrences?

*: Agents need not have a concept of outcome, in which case they don't really have a utility function (because utility functions are functions *of outcomes*). However, this does not significantly impact any of the points made in this post.

[AN #116]: How to make explanations of neurons compositional

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

HIGHLIGHTS

Compositional Explanations of Neurons *(Jesse Mu et al)* (summarized by Robert): Network dissection is an interpretability technique introduced in 2017, which uses a dataset of images with dense (i.e. pixel) labels of concepts, objects and textures. The method measures the areas of high activation of specific channels in a convolutional neural network, then compares these areas with the labelled areas in the dataset. If there's a high similarity for a particular channel (measured by the intersection divided by the union of the two areas), then we can say this channel is recognising or responding to this human-interpretable concept.

This paper introduces an extension of this idea, where instead of just using the basic concepts (and matching areas in the dataset), they search through logical combinations of concepts (respectively areas) to try and find a compositional concept which matches the channel's activations. For example, a channel might respond to (water OR river) AND NOT blue. This is still a concept humans can understand (bodies of water which aren't blue), but enables us to explain the behaviour of a larger number of neurons than in the original network dissection method. Their work also extends the method to natural language inference (NLI), and they interpret neurons in the penultimate layer of a BiLSTM-based network trained to know whether a sentence entails, contradicts, or is neutral with respect to another. Here they create their own features based on words, lexical similarity between the two sentences, and part-of-speech tags.

Using their method, they find that channels in image classifiers do learn compositional concepts that seem useful. Some of these concepts are semantically coherent (i.e. the example above), and some seem to have multiple unrelated concepts entangled together (i.e. operating room OR castle OR bathroom). In the NLI network, they see that many neurons seem to learn shallow heuristics based on bias in the dataset - i.e. the appearance of single words (like nobody) which are highly informative about the classification.

Finally, they use their method to create copy-paste adversarial examples (like in Activation Atlas (AN #49)). In the Places365 dataset (where the goal is to classify places), they can crudely add images which appear in compositional concepts aligned with highly contributing neurons, to make that neuron fire more, and hence change the classification. Some of these examples generalise across classifier architectures, implying a bias present in the dataset.

Robert's opinion: I think work which targets specific neurons and what they're doing is interesting as it can give us a very low-level understanding of the model, which I feel is necessary to achieve the level of understanding required by alignment solutions which use interpretability (i.e. those in **An overview of 11 proposals for building safe advanced AI** (**AN #102**)). The main limitation of this approach is that

it currently requires a large amount of dense human labelling of the datasets, and if a concept isn't in the labels of the dataset, then the method won't be able to explain a neuron using this concept. Also, the fact that their interpretability method is able to give insights (in the form of creating copy-paste examples) is a useful sign it's actually doing something meaningful, which I think some other interpretability methods lack.

TECHNICAL AI ALIGNMENT

LEARNING HUMAN INTENT

Learning to Summarize with Human Feedback *(Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler et al)* (summarized by Rohin): OpenAI has been working on **finetuning language models from human preferences** (**AN #67**). This blog post and paper show the progress they have made on text summarization in particular since their last release.

As a reminder, the basic setup is similar to that of **Deep RL from Human Preferences**: we get candidate summaries by executing the policy, have humans compare which of two summaries is better, and use this feedback to train a reward model that can then be used to improve the policy. The main differences in this paper are:

1. They put in a lot of effort to ensure high data quality. Rather than having MTurk workers compare between summaries, they hire a few contractors who are paid a flat hourly rate, and they put a lot of effort into communicating what they care about to ensure high agreement between labelers and researchers.

2. Rather than collecting preferences in an online training setup, they collect large batches at a time, and run a relatively small number of iterations of alternating between training the reward model and training the policy. My understanding is that this primarily makes it simpler from a practical perspective, e.g. you can look at the large batch of data you collected from humans and analyze it as a unit.

3. They initialize the policy from a model that is first pretrained in an unsupervised manner (as in **GPT-3** (**AN #102**)) and then finetuned on the reference summaries using supervised learning.

On the Reddit task they train on, their summaries are preferred over the reference summaries (though since the reference summaries have varying quality, this does not imply that their model is superhuman). They also transfer the policy to summarize CNN / DailyMail news articles and find that it still outperforms the supervised model, despite not being trained at all for this setting (except inasmuch as the unsupervised pretraining step saw CNN / DailyMail articles).

An important ingredient to this success is that they ensure their policy doesn’t overoptimize the reward, by adding a term to the reward function that penalizes deviation from the supervised learning baseline. They show that if they put a very low weight on this term, the model overfits to the reward model and starts producing bad outputs.

Read more: **Paper: Learning to summarize from human feedback**

Rohin's opinion: This paper is a great look at what reward learning would look like at scale. The most salient takeaways for me were that data quality becomes very important and having very large models does not mean that the reward can now be optimized arbitrarily.

FORECASTING

Does Economic History Point Toward a Singularity? *(Ben Garfinkel)* (summarized by Rohin): One important question for the long-term future is whether we can expect accelerating growth in the near future (see e.g. this **recent report** (**AN #105**)). For AI alignment in particular, the answer to this question could have a significant impact on AI timelines: if some arguments suggested that it would be very unlikely for us to have accelerating growth soon, we should probably be more skeptical that we will develop transformative AI soon.

So far, the case for accelerating growth relies on one main argument that the author calls the *Hyperbolic Growth Hypothesis* (HGH). This hypothesis posits that the growth *rate* rises in tandem with the population size (intuitively, a higher population means more ideas for technological progress which means higher growth rates). This document explores the *empirical* support for this hypothesis.

I'll skip the messy empirical details and jump straight to the conclusion: while the author agrees that growth rates have been increasing in the modern era (roughly, the Industrial Revolution and everything after), he does not see much support for the HGH prior to the modern era. The data seems very noisy and hard to interpret, and even when using this noisy data it seems that models with constant growth rates fit the pre-modern era better than hyperbolic models. Thus, we should be uncertain between the HGH and the hypothesis that the industrial revolution triggered a one-off transition to increasing growth rates that have now stabilized.

Rohin's opinion: I'm glad to know that the empirical support for the HGH seems mostly limited to the modern era, and may be weakly disconfirmed by data from the pre-modern era. I'm not entirely sure how I should update -- it seems that both hypotheses would be consistent with future accelerating growth, though HGH predicts it more strongly. It also seems plausible to me that we should still assign more credence to HGH because of its theoretical support and relative simplicity -- it doesn't seem like there is strong evidence suggesting that HGH is false, just that the empirical evidence for it is weaker than we might have thought. See also **Paul Christiano's response**.

NEAR-TERM CONCERNS

MACHINE ETHICS

Reinforcement Learning Under Moral Uncertainty *(Adrien Ecoffet et al)* (summarized by Rohin): Given that we don't have a perfect ethical theory ready to load into an AI system, and we don't seem poised to get one any time soon, it seems worth looking into approaches that can deal with *moral uncertainty*. Drawing on the literature on moral uncertainty in philosophy, the authors consider several methods by which multiple moral theories can be aggregated, such as averaging over the

theories, making decisions through a voting system, and having the theories compete to control the agent's overall actions. They implement several of these in RL agents, and test them on simple gridworld versions of various trolley problems. They find that all of the methods have advantages and disadvantages.

Rohin's opinion: The central challenge here is that normalizing different moral theories so that they are comparable is **difficult** (**AN #60**) (see Section 2.3). This issue plagues even computationally intractable idealizations like **assistance games** (**AN #69**) that can perform full Bayesian updating on different moral theories. I'd love to see better theoretical solutions for this challenge.

OTHER PROGRESS IN AI

DEEP LEARNING

Deploying Lifelong Open-Domain Dialogue Learning *(Kurt Shuster, Jack Urbanek et al)* (summarized by Rohin): Most research in natural language processing (NLP) follows a paradigm in which we first collect a dataset via crowdsourced workers, and then we train a model on this dataset to solve some task. Could we instead have *lifelong learning*, in which a model could continue learning after being deployed, getting better and better the more it is used? This paper shows one instantiation of such an approach, in a fantasy role-playing game.

The authors take the previously developed LIGHT role-playing setting, and gamify it. The human player talks to a language model while playing some role, and earns stars and badges for saying realistic things (as evaluated by another language model). Rather than paying crowdsourced workers to provide data, the authors instead merely advertise their game, which people then play for fun, reducing the cost of data acquisition. They find that in addition to reducing costs, this results in a more diverse dataset, and also leads to faster improvements in automated metrics.

Rohin's opinion: Ultimately we're going to want AI systems that learn and improve over time, even during deployment. It's exciting to see an example of what that might look like.

UNSUPERVISED LEARNING

Understanding View Selection for Contrastive Learning *(Yonglong Tian et al)* (summarized by Flo): **Contrastive multiview learning** (**AN #92**) is a self-supervised approach to pretraining classifiers in which different views of data points are created and an encoder is trained to minimize the distance between encodings of views corresponding to data points with the same label while maximizing the distance between encodings of views with different labels.

The efficacy of this approach depends on the choice of views as well as the downstream task the neural network is going to be trained for. To find the most promising views, the authors propose the Infomin principle: all views should keep task-relevant information while the mutual information between views is minimized. The principle is supported by various observations: Firstly, earlier approaches to contrastive learning in the image domain that use data augmentation to preserve object identity while creating diverse views can be seen as an implicit application of

the Infomin principle. Secondly, varying the mutual information between views (for example by changing the distance between two cropped views of the same image) creates an inverted U-curve for downstream performance corresponding to poor performance if there is too much or too little mutual information between the views. Lastly, the authors also find an inverted U-curve in performance for different colour spaces when using channels as views and the Lab colour space which was built to mimic human colour perception is close to the optimum, meaning that human colour perception might be near-optimal for self-supervised representation learning.

The authors then use the Infomin principle to select image augmentations for contrastive pretraining and improve the state of the art in linear readout on ImageNet from 69.3% to 73% for Top-1 accuracy and from 89% to 91.1% for Top-5 accuracy.

Read more: **What makes for good views for contrastive learning**

Flo's opinion: While the Infomin principle seems powerful and their results look impressive, I am not really convinced that the principle actually played an important role in finding the image augmentations they ended up using, as there is little description of how that happened and the augmentations rather look like the result of combining previously used approaches and doing some hyperparameter optimization.

HIERARCHICAL RL

Decentralized Reinforcement Learning: Global Decision-Making via Local Economic Transactions *(Michael Chang et al)* (summarized by Zach): Increasing the scalability of learning systems is a central challenge to machine learning. One framework is to organize RL agents as 'super' agents, large collections of simpler agents that each make decisions according to their own incentives. If it were possible to get the incentives correct, the dominant equilibria would be identical to the optimal solution for the original RL problem.

In this paper, the authors introduce a framework for decentralizing decision-making by appealing to auction theory. There is a separate simple agent for each action. At every a timestep, a Vickrey auction is run in which each agent can bid for the superagent executing their particular action. The trick is that when an agent successfully wins a bid and acts on a state, it then 'owns' the produced next state, and 'earns' the result of the auction in the next round. (At the end of an episode, the owner of the state earns the reward of the trajectory.) Intuitively, the agent wants to bid on states in which it can make progress towards earning the final reward, as those will be states that other agents want to buy. The authors show that this scheme incentivizes each agent to bid the Q-value of their action in the given state, which would then lead to an optimal policy.

The authors test out this approach with some simple MDPs. They also investigate a task where they try to get the agents to rotate MNIST images so that a classifier will recognize them. Finally, they investigate task transfer by training agents on simple sub-tasks and then reusing those agents to learn a related task making use of both sub-tasks.

Read more: **Paper: Decentralized Reinforcement Learning: Global Decision-Making via Local Economic Transactions**

Zach's opinion: Imagine **Twitch plays**, but you use a reputation to buy and sell your actions. The actual idea in the paper is slightly more mundane than this because the primitives are bidders. **Reinforcement Learning with Competitive Ensembles of Information-Constrained Primitives** (**AN #66**) is a similar piece of work that also uses primitives as the basic level of selection. However, their incentive mechanism is different: agents pay according to how much information from the environment they use and then get a reward back for their actions. However, there's good reason to think options could work as well since in both of these papers there's evidence that primitives that learn sub-tasks are useful in new tasks.

NEWS

Cooperative AI Workshop (summarized by Rohin): This NeurIPS workshop has the goal of improving the *cooperation* skills of AI systems (whether with humans or other machines), which encompasses a *very* wide range of research topics. The deadline to submit is September 18.

Senior Systems Safety Engineer *(OpenAI)* (summarized by Rohin): OpenAI is hiring for a senior systems safety engineer. From my read of the job description, it seems like the goal is to apply the principles from **Engineering a Safer World** (**AN #112**) to AI development.

Early-career funding for individuals interested in improving the long-term future (summarized by Rohin): This Open Philanthropy program aims to provide support for people who want to focus on improving the long-term future. The primary form of support would be funding for graduate school, though other one-off activities that build career capital also count. They explicitly say that people interested in working on AI policy or risks from transformative AI should apply to this program (possibly in addition to their **AI fellowship** (**AN #66**)). The stage 1 deadline is January 1, but if you submit earlier they aim to respond within 10 working days.

Updates Thread

If you've updated your belief about something you think is worth noting, post it here.

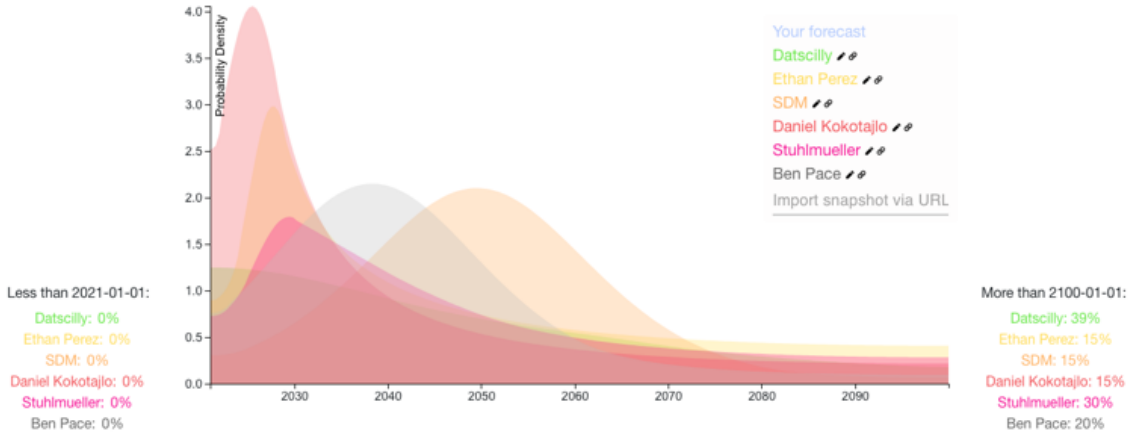
- It doesn't have to be a full blown "mind change", just an incremental update to your beliefs.
- I'm thinking it'd be good to have a low bar for what is "worth noting". Even if it's something trivial, I figure that the act of discussing updates itself is beneficial. For rationality practice, and for fun!
- That said, I also expect that browsing through updates that other people on LessWrong make will lead to readers making similar updates themselves a decent amount of the time.
- I've been developing a strong opinion that journaling and self-reflection in general is incredibly useful. Significantly underrated even among those that preach it. This thread is a way to perform such journaling and self-reflection.

Reflections on AI Timelines Forecasting Thread

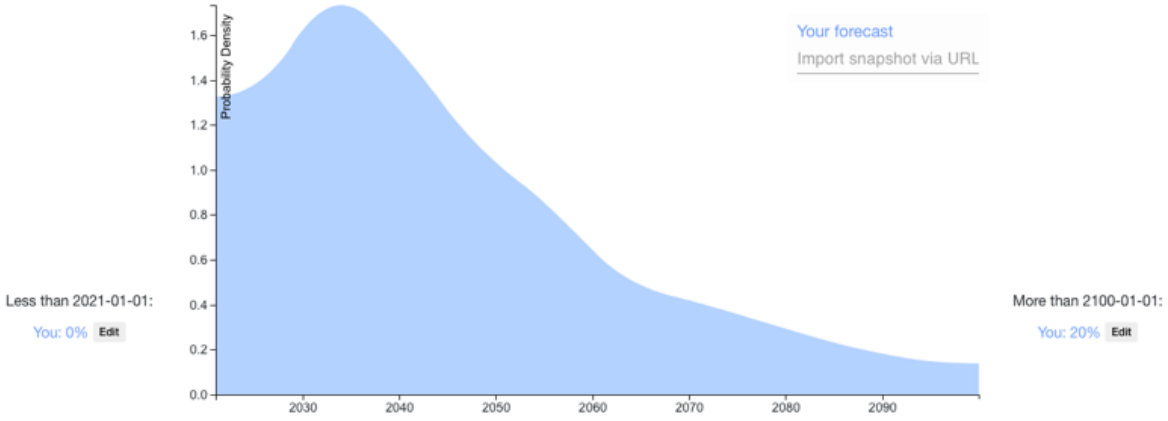
It's been exciting to see people engage with the AI forecasting thread that Ben, Daniel, and I set up! The thread was inspired by Alex Irpan's AGI timeline update, and our hypothesis that visualizing and comparing AGI timelines could generate better predictions. Ought has been working on the probability distribution tool, Elicit, and it was awesome to see it in action.

14 users shared their AGI timelines. Below are a number of their forecasts overlaid, and an aggregation of their forecasts.

Comparison of 6 top-voted forecasts



Aggregation, weighted by votes



The thread generated some interesting learnings about AGI timelines and forecasting. Here I'll discuss my thoughts on the following:

- The object level discussion of AGI timelines
- How much people changed their minds and why
- Learnings about forecasting
- Open questions and next steps

AGI timelines

Summary of beliefs

We calculated an aggregation of the 14 forecasts weighted by the number of votes each comment with a forecast received. The question wasn't precisely specified (people forecasted based on slightly different interpretations) so I'm sharing these numbers mostly for curiosity's sake, rather than to make a specific claim about AGI timelines.

- Aggregated median date: **June 20, 2047**
- Aggregated most likely date: **November 2, 2033**
- Earliest median date of any forecast: **June 25, 2030**
- Latest median date of any forecast: **After 2100**

Emergence of categories

I was pleasantly surprised by the emergence of categorizations of assumptions. Here are some themes in the way people structured their reasoning:

- **AGI from current paradigm (2023 – 2033)**
  - GPT-N gets us to AGI
  - GPT-N + improvements within existing paradigm gets us to AGI
- **AGI from paradigm shift (2035 – 2060)**
  - We need fundamental technical breakthroughs
    - Quantum computing
    - Other new paradigms
- **AGI after 2100, or never (2100 +)**
  - We decide not to build AGI
    - We decide to build tool AI / CAIS instead
    - We move into a stable state
  - It's harder than we expect
    - It's hard to get the right insights
    - We won't have enough compute by 2100
  - We can't built AGI
    - There's a catastrophe that stops us from being able to build AGI
- **Outside view reasoning**
  - With 50% probability, things will last twice as long as they already have
  - We can extrapolate from rate of reaching past AI milestones

When sharing their forecasts, people associated these assumptions with a corresponding date interval for when we would see AGI. I took the median lower bound and median upper bound for each assumption to give a sense of what people are expecting if each assumption is true. Here's a spreadsheet with all of the assumptions. Feel free to make a copy of the spreadsheet if you want to play around and make edits.

Did this thread change people's minds?

One of the goals of making public forecasts is to help people identify disagreements and resolve cruxes. The number of people who updated is one measure of how well this format achieves this goal.

There were two updates in comments on the thread (Ben Pace and Ethan Perez), and several others not explicitly on the thread. Here are some characteristics of the thread that caused people to update (based on conversations and inference from comments):

- **It was easy to notice surprising probabilities.** In most forecasts, Elicit's bin interface meant probabilities were linked to specific assumptions. For example, it was easy to disagree with Ben Pace's specific belief that with 30% probability, we'd reach a stable state and therefore wouldn't get AGI before 2100. Seeing a visual image of people's distributions also made surprising beliefs (like sharp peaks) easy to spot.
- **Visual comparison provided a sense check.** It was easy to verify whether you had too little or too much uncertainty compared to others.
- **Seeing many people's beliefs provides new information.** Separate from the information provided by people's reasoning, there's information in how many people support certain viewpoints. For example, multiple people placed a non-trivial probability mass on the possibility that we could get AGI from scaling GPT-3.
- **The thread catalyzed conversations outside of LessWrong**

Learnings about forecasting

Vaguely defining the question worked surprisingly well

The question in this thread ("Timeline until human-level AGI") was defined much less precisely than similar Metaculus questions. This meant people were able to forecast using their preferred interpretation, which provided more information about the range of possible interpretations and sources of disagreements at the interpretation level. For example:

- tim_dettmers' forecast defined AGI as not making 'any "silly" mistakes,' which generated a substantially different distribution
- datscilly's forecast used the criteria from this Metaculus question and this Metaculus question, including, for example: "Able to reliably pass a Turing test of the type that would win the Loebner Silver Prize."
- Rohin Shah predicted timelines for transformative AI

A good next step would be to create more consensus on the most productive interpretation for AGI timeline predictions.

Value of a template for predictions

When people make informal predictions on AGI, they often define their own intervals and ways of specifying probabilities (e.g. '30% probability by 2035', or 'highly likely by 2100'). For example, this list of predictions shows how vague a lot of timeline predictions are.

Having a standard template for predictions forces people to have numerical beliefs across an entire range. This makes it easier to compare predictions and compute disagreements across any range (e.g. this bet suggestion based on finding the earliest range with substantial disagreement). I'm curious how much more information we can capture over time by encouraging standardized predictions.

Creating AGI forecasting frameworks

Ought's mission is to apply ML to complex reasoning. A key first step is making reasoning about the future explicit (for example, by decomposing the components of a forecast,

isolating assumptions, and putting numbers to beliefs) so that we can then automate parts of the process. We'll share more about this in a blog post that's coming soon!

In this thread, it seemed like a lot of people built their own forecasting structure from scratch. I'm excited about leveraging this work to create structured frameworks that people can start with when making AGI forecasts. This has the benefits of:

- Avoiding replication of cognitive work
- Clearly isolating the assumptions that people disagree with
- Generating more rigorous reasoning by encouraging people to examine the links between different components of a forecast and make them explicit
- Providing data that helps us automate the reasoning process

Here are some ideas for what this might look like:

- **Decomposing the question more comprehensively based on the categories outlined above**
  - For example, creating your overall distribution by calculating: P(Scaling hypothesis is true) * Distribution for when we will get AGI | Scaling hypothesis is true + P(Need paradigm shift) * Distribution for when we will get AGI | Need paradigm shift + P(Something stops us) * Distribution for when we will get AGI | Something stops us
- **Decomposing AGI timelines into the factors that will influence it**
  - For example, compute or investment
- **Inferring distributions from easy questions**
  - For example, asking questions like: "If the scaling hypothesis is true, what's the mean year we get AGI?" and use the answers to infer people's distributions

What's next? Some open questions

I'd be really interested in hearing other people's reflections on this thread.

Questions I'm curious about

- How was the experience for other people who participated?
- What do people who didn't participate but read the thread think?
- What updates did people make?
- What other questions would be good to make forecasting threads on?
- What else can we learn from information in this thread, to capture the work people did?
- How can Elicit be more helpful for these kinds of predictions?
- How else do you want to build on the conversation started in the forecasting thread?

Ideas we have for next steps

- **Running more forecasting threads on other x-risk / catastrophic risks.** For example:
  - When will humanity go extinct from global catastrophic biological risks?
  - How many people will die from nuclear war before 2200?
  - When will humanity go extinct from asteroids?

    - By 2100, how many people will die for reasons that would not have occurred if we solved climate change by 2030?
- **More decomposition and framework creation for AGI timeline predictions**
    - We’re working on making Elicit as useful as we can for this!

Petrov Event Roundup 2020

Petrov Day is this weekend. Various LessWrong folk have taken the opportunity to reflect on the Day that Stanislav Petrov Didn't Destroy the World. Some people have employed ritual, and others parties. Some employ Big Red Buttons of some sort.

This year, many of the traditional manners people have commemorated the event are a bit different.

East Coast Petrov Megameetup

The New York Rationality community will be holding an Online Petrov Megameetup, which includes:

- 3:00pm - 3:15am Welcome
- 3:15pm - 4:15pm Ice Breakers
- 4:15pm - 5:30pm Lightning Talks & Activities
- 5:30pm - 6:00pm Snack/Water Break
- 6:00pm - 7:00pm Ritual

It's recommended that you get 8 candles/candleholders for the ritual if you can, but if not you can listen along, and maybe download a Candle App for your phone that you can use for a few key moments.

Austin Petrov Day

The folks in Austin are holding an in person, masked, outdoor Petrov Day.

> We have modified the ceremonial manual to accommodate the outdoor setting. If possible, please bring a printed copy of the double-sided booklet version (or a mobile device on which you can read the mobile-friendly version), and a pen/pencil.

Littleton, CO Petrov Day

Littleton is having an outdoor potluck:

> Come join us as we celebrate the world not having been destroyed, and raise a glass of Vodka to Stanislav Petrov.
>
> We have chosen an outdoor venue for pandemic concerns, and will have a potluck for anyone willing to break bread during these troubled times. Masks are encouraged if you are not eating or drinking, and social distancing is recommended for anyone not already spending time together (some attendees have created quarantine circles, so don't be surprised if you see people ignoring typical precautions).

If you're hosting a Petrov Day celebration of some kind that people are welcome to join, please list them in the comments below!

Needed: AI infohazard policy

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

The premise of AI risk is that AI is a danger, and therefore research into AI might be dangerous. In the AI alignment community, we're trying to do research which makes AI safer, but occasionally we might come up with results that have significant implications for AI capability as well. Therefore, it seems prudent to come up with a set of guidelines that address:

- Which results should be published?
- What to do with results that shouldn't be published?

These are thorny questions that it seems unreasonable to expect every researcher to solve for themselves. The inputs to these questions involve not only technical knowledge about AI, but also knowledge about the behavior of progress, to the extent we can produce such using historical record or other methods. AI risk organizations might already have internal policies on these issues, but they don't share them and don't discuss or coordinate them with each other (that I know of: maybe some do it in private channels). Moreover, coordination might be important even if each actor is doing something reasonable when regarded in isolation (avoiding bad Nash equilibria). We need to have a public debate on the topic inside the community, so that we arrive at some consensus (that might be updated over time). If not consensus, then at least a reasonable spectrum of possible policies.

Some considerations that such a policy should take into account:

- Some results might have implications that shorten the AI timelines, but are still good to publish since the distribution of outcomes is improved.
- Usually we shouldn't even start working on something which is in the should-not-be-published category, but sometimes the implications only become clear later, and sometimes dangerous knowledge might still be net positive as long as it's contained.
- In the midgame, it is unlikely for any given group to make it all the way to safe AGI by itself. Therefore, safe AGI is a broad collective effort and we should expect most results to be published. In the endgame, it might become likely for a given group to make it all the way to safe AGI. In this case, incentives for secrecy become stronger.
- The policy should not fail to address extreme situations that we only expect to arise rarely, because those situations might have especially major consequences.

Some questions that such a policy should answer:

- What are the criteria that determine whether a certain result should be published?
- What are good channels to ask for advise on such a decision?
- How to decide what to do with a potentially dangerous result? Circulate in a narrow circle? If so, which? Conduct experiments in secret? What kind of experiments?

The last point is also related to a topic with independent significance, namely, what are reasonable precautions for testing new AI algorithms? This has both technical aspects (e.g. testing on particular types of datasets or particular types of environments, throttling computing power) and procedural aspects (who should be called to advice/decide on the manner). I expect to have several tiers of precautions, s.t. a tier can be selected according to our estimate of the new algorithm's potential, and guidelines for producing such an estimate.

I emphasize that I don't presume to have good answers to these questions. My goal here was not to supply answers, but to foster debate.

Rationality and playfulness

Can rationality help us be playful? Can we be playful when we're solitary?

Play is usually interactive. It's about connecting with other people, or even with a pet animal. When people do "playful" things by themselves, it's usually for relaxation or practice.

The presence of a second person changes everything. You can react to each other, surprise or influence each other, create structure together. Group decisions are easier to commit to after a choice is made.

Many activities can be fun, engaging, and interesting, without being obviously "playful." A chess game can involve more mental concentration and stillness than is required of most people at their jobs, yet be a delightful hobby activity for the participants. We even say we "played" a game of chess. So why doesn't chess feel playful?

Partly, it's because play is usually physical. Even if we're just having a playful conversation, our body language and voices can bring a physical element to the exchange. Chess, like writing, reading, and many other fun-but-not-playful activities doesn't typically use our big muscles or our social muscles.

What about exercise? That's not conventionally playful either, even though it uses our muscles. Even athletics, like a game of tennis, can feel fun-but-not-playful, unless the participants are joking around and being social while they play.

It really does seem to be the social element that's key for a sense of play. If we watch a talk show, the participants often have very playful interactions, even though they're mostly just sitting in chairs talking.

Even in a social setting where both participants desire a playful conversation, though, it's often very difficult to achieve. It's so easy for even good friends to feel awkward, formal, and serious in each others' company. Coming up with a playful text message takes *work* for many people. Especially at first. If a conversation chain gets going that has taken on a playful tone, it might stay that way. Positive energy, a combination of kindness and rudeness, and not taking things literally all can be fertile ground for a playful conversation.

If you're leading a solitary life, though, is it possible to be playful? What about in these lonely times?

Can you think playful thoughts? After all, our inner world can often feel like we have multiple perspectives, multiple voices within us. Is it possible for them to have a playful interaction?

Can you find a sense of play in observing the world around you? Can you flirt with a building, joke with the sky, let the trees in on a little secret, tell a story to the sidewalk? I'm not just being poetic. I literally mean that it seems at least possible that there's a way to have a felt sense of playful interaction with the world of objects.

Certainly it's possible to have brief, playful interactions with strangers, especially if they're in a service role. There are ways to be friendly with the cashier at the grocery

store.

What about in being creative, meditative, or just in the activities of daily living? Is there a playful way to clean the bathroom? To meditate? To write a song?

When I imagine trying to do any of these things, my first thought is that I would feel foolish, self-conscious, and pathetic. A person who's so needy that he resorts to seeking connection with the inanimate objects around him. I heard a story once about a man who was so lonely that he took to hugging a support beam in his house.

It occurs to me, though, that those reactions are coming from inside me. It's my self-talk and my imagination that anticipate that sense of bleak foolishness. Observing that, it seems to me that my self-talk and my imagination are responsible, at least in part, for depriving me of playfulness. Of even trying for it.

After all, I do many things just to see if they can be done. Some of those challenges are incredibly difficult. Sometimes I have little idea of how I'll approach it. By throwing myself into it, setting the goal, my intuition starts to devise a way forward. Maybe this could work.

Experiment 1

I try just standing up and seeing what might happen. My perception changes, almost immediately, to a quite different frame of mind than I'm used to. Suddenly I feel like I'm an actor on a stage, even though nobody is home. I feel the urge to take my shirt off. Why not? As I stand there, I notice that the white blanket on the couch looks like a cape. I imagine wearing it that way.

I walk around the house aimlessly. Sometimes I stand looking out the window, or at myself in the mirror in the shadows of the hall. In the kitchen, I find myself gazing at the reflection my kitchen table makes in the mirror, with my silhouette behind it.

I notice how my mind wants to give itself tasks and find distractions. To clean messes here and there. To walk around, set destinations for myself. Sometimes I tap on the walls. Sometimes I just look at objects: the box fan, the thermostat, the pots and pans. Most of the time, it's just a passive noticing. Sometimes, my brain imagines something silly I could do with them, like banging the pots and pans together.

There's a sense of achievement in the moments when I notice something beautiful, like the reflection in the window pane, or how my body looks in the shadowy full-length mirror in the hall.

Experiment 2

After writing all that down, I stand up again. Another experiment. This time, the mindset grows on me more easily. At first, I regard things around me: the drapes, the brick wall on the building outside my window, and my brain wants to find something in them, but I know that there's nothing there. This isn't something you strive for. It's something that should just appear.

Then I walk into the kitchen and look at the hanging fruit basket. I remember how it was given to me several years ago by a friend who was living with me. I observe that I don't usually go back through old memories, especially not when I'm alone. Then I remember a more recent memory associated with it. A week ago, I came home from a

trip, and a potato in it had gone bad - liquified - dripping the most foul-smelling brown liquid. Even after cleaning it up, it took half a day for the smell to disappear.

I look around at the messes that need to be cleaned up after a full day of activity. The boxes of cleaning supplies that just arrived because I'm trying to keep a cleaner house. I think of the reasons why I'm doing that. And so many of the other forces that define my life: school, work, my efforts to maintain my social life. It all feels very big. And very small.

I look at the pots stacked on top of the refrigerator. It looks sloppy. But what can I do? It's the practical way to store them. I regard the cabinet drawer that opens with an awful, nails-on-chalkboard squeaking sound. Will I get around to sanding it down at some point? Then I look at the print I made of the elephant hawk moth, *Deilephila elpenor,* the moth that can see full-color vision in dim starlight. I think about how I learned to make block prints. Notice how I like the rough texture of the print, and the childish simplicity of the lines of its body. How if I don't pick it apart, it looks beautiful and unusual. I think that perhaps *Deilephila elpenor* is a metaphor for this project, of learning how to be playful in solitude.

Experiment 3

I stand up briefly again. For less than a minute. Surveying the kitchen again, I get this conception of how it would be to be a relentless, fast, machine-like worker in my own life. One who cleaned every mess as fast as possible, then immediately transitioned to sanding down that cabinet drawer, to organizing the fridge. That threw on music as I worked. Now as I sit here writing this, perhaps one who spontaneously breaks out dancing all in the middle of that frantic activity. A sense of being magnetized to the world, controlled by it almost like a puppet, drilling down deeper and deeper into what needs to be done until, perhaps, hitting impenetrable rock. Or oil. Or fossils.

Then again, I think about this sense of playfulness. How right now at least, it seems to demand slowness, and stillness. There is the mode of compressing as much accomplishment and activity into the shortest time interval possible. Losing the meta-level and burning yourself up in sheer obvious activity. But don't you lose something like that? Would it be good to practice both, to switch? Is there something important for me to learn in this playful stillness? Am I being playful? Is there also something to drill into in the stillness? Not measured in checking off tasks from a list, but in some other way?

Experiment 4

I won't recount everything I think and experience this time. Suffice to say that my thoughts begin with deep melancholy, dwelling on many sad aspects of my life, the world we live in, the dysfunctions, the ways people fall through the cracks, and the ways we try to escape.

Then it hits me. If I'm not playful, it's because I relentlessly dwell on sadness and dysfunction and a sense of lack.

What if I choose to think about experiences from the day that were pleasant? Or found a playful way to think about the experiences I had?

I reflect on the COVID-19 test I had today, and imagine that it was like having my brains twisted like spaghetti around a fork. The chipper nurse who registered me for the test. How I'm waiting for an iPhone with a functioning camera to arrive in the mail

so I can take pictures for Tinder, how I'm going to have to figure out how to pose, to be a show-off. I think to myself, "this is going to be fun!" I begin to feel as though I'm having a conversation with myself. That I'm playing with myself.

Experiment 5

There's a few stalks of lavender in the vase on my kitchen table. I stole them from the church.

Normally, I would just stare vacantly at them. Or I'd say something like "I took them from the church," stated as a dull fact. But now, it's *I stole them from the church*. As if I'm letting *myself* in on a little secret. That I've been up to something a bit mischievous.

Experiment 6

It occurs to me that I've never felt playful while cooking a meal. It's been work. An attempt to impress. A learning effort. Never play, though. Except once when I was little, and my mom let me throw everything in the spice cabinet into a "cake." I thought it was poisonous and fed it to the birds. Not out of malice. I was just a bit of a stupid child, really. Wasn't thinking overly hard about the birds' wellbeing. I'm sure that if I'd thought twice about it, I'd have found something else to do with it, but instead I took it out and sprinkled the crumbs in the grass.

The links of all the activities I've done for serious motivations, with a serious attitude, spread out before me. What unites them? There's something missing from all of them. It's a story. It's caring. A sense of heart, of play, of connection to myself. It's something I think has been available this whole time. The story I've been telling has been largely bitter, paranoid, anxious, jealous, self-deprecating, sarcastic, arrogant, dull, serious, and wounded, for a very long time. I put a smile on my face. I'd really like to change that.

Experiment 7

My thoughts putter around. A birthday party from two years ago. A woman I met there who I flirted with and haven't spoken with, a friend of a friend. I realize that I still have a bit of a crush on her. The feeling actually registers in my heart. It's not a mental realization, not a plan, not a "what if I got in touch with her?" or a "I should ask my friend if she's single." It's just an emotion, a pleasant twinge, nice to have all on its own.

I double check the name of the woman I matched with on Tinder, but whom I haven't heard back from yet. Her name is the same as the one from a song I liked when I was a kid and haven't listened to in many years.

I check myself out in the mirror. I realize that after changing my grooming habits dramatically, I feel attracted to myself in a way that I haven't ever experienced before. It's a nice feeling.

All I'm doing is gently encouraging my mind to land on pleasant memories, objects with good associations. No need to control or actively seek them out. It's like my mind is a butterfly that has finally learned to seek out flowers to land on. Sometimes it's "in between" thoughts, just traveling, or blank.

I think I should give my house a name.

Experiment 8

My brothers' jade plant is half-hidden behind a wall, peeking out around it with two of its branches. It's in a big, beautiful clay pot with Chinese dragon designs all around it. Right next to the shoe rack. Needs better Feng Shui!

There's a way of paying attention to objects so that they reveal themselves to you. If you stare hard at a point on the wall, it feels neurotic. There's nothing there. But allow your gaze to trace over the whole house, and suddenly you're in a place that's full of memories, potential, and meaning. *This is my house. I live here. It's a place where I can invite guests in, where they feel privileged to feel welcome. It's a place that I rent, but that is mine for as long as I am doing so. I remember when I first moved in. I remember when I didn't have a house, when I lived out of my car for a summer. I think about the other people living nearby: the intriguing apartments filled with plants, Christmas lights, and comfortable-looking furniture across the alley. Who lives there?*

I should pick out my favorite houses, and imagine the lives that people lead there.

Experiment 9

Think about linoleum tiles. Somebody designed the color scheme on these ones. They're sort of flecked with different shades of blue and white. It's kind of pretty, actually. Did the linoleum tile designer hope that somebody would appreciate the way they make the kitchen floor look a bit like an abstract archipelago of sandy cream tiles and blueish watery tiles?

Experiment 10

I'm looking at the stove. At first, it seems tiny, cramped: this is all I can afford. Then it changes. It's cozy. It's all I need. I imagine hanging up a little earthy bundle of plants behind it. A rose or a bundle of grass. Just to mark it. To give it some love. Maybe it would catch fire. But in any case, I don't need to. It's enough to practice that mental shift. To see the thing, to honor it, to appreciate it for what it is, to find beauty in it.

Experiment 11

Other things I think about. A stone I brought back from Iceland transports me back there. Looking at the fingerling potatoes I bought makes me think of my breakfast tomorrow morning. Black coffee, potatoes chopped thin with eggs, green onions, and hot sauce. I so rarely think about meals the next day, or even later the same day.

There's a garden spider on a web outside of my window. I draw up close to peer at it. Hairy legs, a pattern of white crosses on its abdomen. The web blows in the wind, rippling the spider just a little closer to me as it hangs in the darkness. I wonder if spiders can feel cold.

Experiment 12

Among many other observations, one thing I notice developing is an awareness of how my mind can look at things in two very different ways. One is gentle, detached, and moves like light over the surfaces of things. The other is piercing, aggressive, and tunnels like a deep borer digging a tunnel. The latter is all to easy for me. It's the default. I like the developing ability to have gentle thoughts.

It occurs to me that in all our investigation here on Less Wrong about the problems of rational thought, the difficulties of synthesis, and how bias and emotion affect our judgment, it's never seemed quite possible to bring it all together. The problem of good thinking feels impossibly large, for even one single issue. The arguments endless, the proofs too large for the mind of humanity.

I have an inkling. I'm standing in the hall, and becoming aware of all the machines and electronics that are running in the house. My computer. The lights. The refrigerator. And I can hear an airplane flying overhead, a car outside. Smoke is thick in the air from wildfires. A physical awareness of the constant energy usage dawns on me. How little I think of these things most of the time. Every appliance in my house is sucking in energy through a straw from some central power source. So is every other apartment, in every building, throughout this city, and in every city.

The activity is relentless. Manipulation of words on the computer. Of bits, of atoms. The construction company that builds houses, that built my house. The factories refining raw materials into useful ones, and turning those into products that people put to use. The way that sometimes, it comes together in ways that feel meaningful, useful. The side effects, of waste, of CO2 entering the atmosphere, and how it heats up the woods and leads to the forest fire, and how the smoke in the air is keeping me from running, and how this virus and these smokey days are destroying what could have been beautiful and social times in my and in our lives. The argument stops being words on a page. It's a connected series of images and objects that simply *are related*. Science has allowed my mind to move, to wander the globe, in ways that make sense.

This is what it feels like to understand something. Rationality isn't fundamentally argumentative. It is fundamentally experiential. It is observational. It is imaginative and visual. The reason why winning an argument never works is because you have *completely* missed the mark when you argue. Convincing somebody is about helping them to see as you see, to help their mind learn how to wander in the directions you know it's capable of. To help it see the turn it consistently misses and convince it to open certain doors and have a look inside.

So there is a reason why we are stuck right now on this earth.

We are missing the art of opening doors in each others' minds, guiding each other along new paths, and allowing ourselves to be guided. Instead, we are erecting arguments, slogans, screeds, that separate people into camps: those who disagree and reject the thing wholesale, and those who agree and add more links in the chain. Some places relationships and online spaces are nothing but collections of these steely monuments, landmines, booby traps, flags, orders, coded messages, propaganda.

Or maybe that is just my mindset at its most paranoid. Perhaps the fault is not in the words. Maybe this is an era of an extraordinary flowering of the human mind. It may be that we are only just beginning to learn how to *open* ourselves to it. When we stand outside these word-gardens, these strange sculptures with messages we won't understand until we've meditated among them, they seem frightening to us. Who build them, and why? What am I doing here? This experience feels like an intrusion in my life.

Perhaps there is a way of finding playfulness with the world-sculptures, too. Connecting with them, just like tonight I've been able to connect with a reflection in

the window, with a stone, with a fruit basket, with the sight of the community center next door, with my own body, with a spider on its web in the darkness, with a white blanket folded on the couch that looks like a cape I might wear.

What are good rationality exercises?

I want to know what are good rationality exercises.

I was just on a call with Liron and PhilH, hanging out after the weekly LessWrong weekend event, and we discussed exercises that could happen on LessWrong.

Here is the list we generated:

- Thinking Physics
- Fermi Estimates
- Project Euler
- Calibration Training
- Basic probabilistic reasoning
- Basic have-you-read-the-sequences knowledge test (e.g. "Which of the following is an example of 'belief as attire'?")

Another user on the call (whose name I forget) suggested it could be fun to have a daily Fermi Estimate on LessWrong, where everyone submits their number and the model they used to reach the number. I think this would be quite exciting.

Please write answers with other exercises that you think are or might be great for rationality training, some explanation of why you think it could be good, and a suggestion of how it could be incorporated into LessWrong. I'll probably add some of the above myself.

How Much Computational Power Does It Take to Match the Human Brain?

This is a linkpost for https://www.openphilanthropy.org/brain-computation-report

Joe Carlsmith with a really detailed report on computational upper bounds and lower bounds on simulating a human brain:

> Open Philanthropy is interested in when AI systems will be able to perform various tasks that humans can perform ("AI timelines"). To inform our thinking, I investigated what evidence the human brain provides about the computational power sufficient to match its capabilities. This is the full report on what I learned. A medium-depth summary is available here. The executive summary below gives a shorter overview.
>
> [...]
>
> Let's grant that in principle, sufficiently powerful computers can perform any cognitive task that the human brain can. How powerful is sufficiently powerful? I investigated what we can learn from the brain about this. I consulted with more than 30 experts, and considered four methods of generating estimates, focusing on floating point operations per second (FLOP/s) as a metric of computational power.
>
> These methods were:
>
> 1. Estimate the FLOP/s required to model the brain's mechanisms at a level of detail adequate to replicate task-performance (the "mechanistic method").**1**
> 2. Identify a portion of the brain whose function we can already approximate with artificial systems, and then scale up to a FLOP/s estimate for the whole brain (the "functional method").
> 3. Use the brain's energy budget, together with physical limits set by Landauer's principle, to upper-bound required FLOP/s (the "limit method").
> 4. Use the communication bandwidth in the brain as evidence about its computational capacity (the "communication method"). I discuss this method only briefly.
>
> None of these methods are direct guides to the *minimum possible* FLOP/s budget, as the most efficient ways of performing tasks need not resemble the brain's ways, or those of current artificial systems. But if sound, these methods would provide evidence that certain budgets are, at least, big enough (*if* you had the right software, which may be very hard to create – see discussion in section 1.3).**2**
>
> Here are some of the numbers these methods produce, plotted alongside the FLOP/s capacity of some current computers.

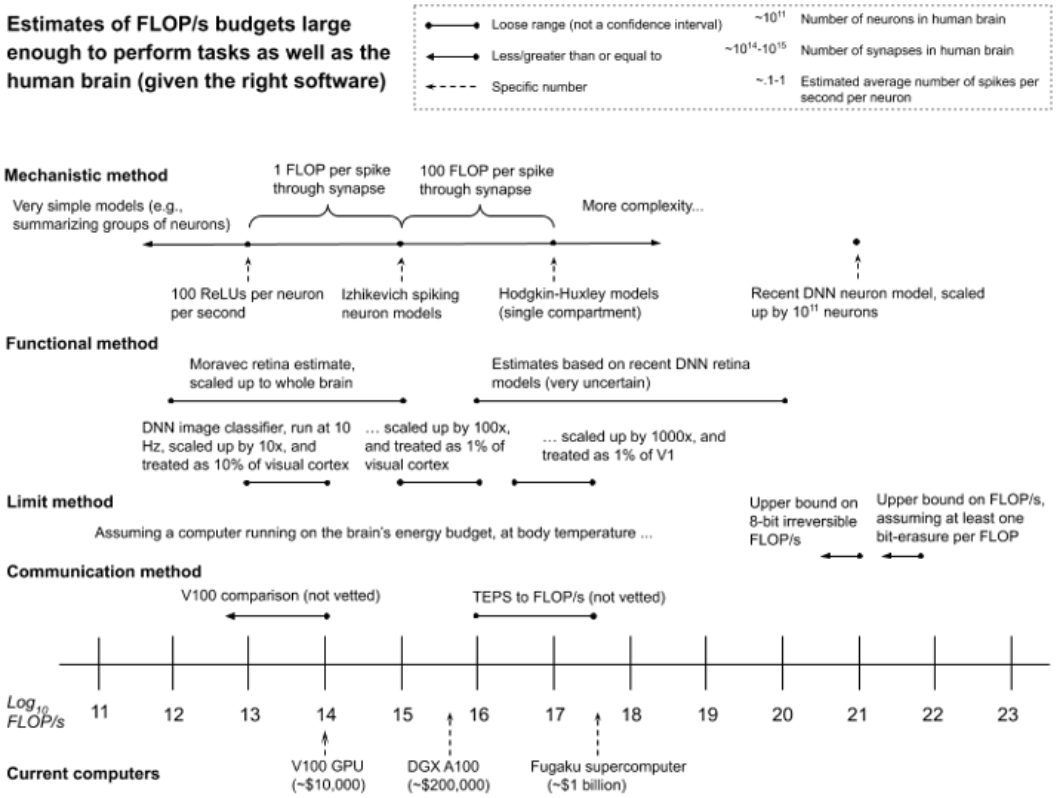


Figure 1: The report's main estimates. *See the conclusion for a list that describes them in more detail, and summarizes my evaluation of each.*

These numbers should be held lightly. They are back-of-the-envelope calculations, offered alongside initial discussion of complications and objections. The science here is very far from settled.

Let the AI teach you how to flirt

"It's Not You, it's Me: Detecting Flirting and its Misperception in Speed-Dates" is a fascinating approach to the study of flirtation. It uses a machine learning model to parse speed-dating data and detect whether the participants were flirting. Here's a sci-hub link. I found three key insights in the paper.

First of all, people basically assume that others share their own intentions. If they were flirting, they assume their partner was too. They're quite bad at guessing whether their partner was flirting, but they do a bit better than chance.

Secondly, the machine learning model was about 70% accurate in detecting flirtation. It's much better than the speed date participants themselves, despite having far less information to draw upon and the fact that the authors used a more forgiving standard of success for people's detection rates than for the detection rates of the machine learning model.

Thirdly, storytelling and conversations about friends seem to be the strongest signals of flirtation. Talking about the mundane details of student life (this was on a college campus) were the strongest signals of non-flirtation.

Finally, men and women have quite different approaches to flirtation:

> Men who say they are flirting ask more questions, and use more you and we. They laugh more, and use more sexual, anger [*hate/hated, hell, ridiculous, stupid, kill, screwed, blame, sucks, mad, bother, shit*], and negative [*bad, weird, hate, crazy, problem*, difficult, tough, awkward, boring, wrong, sad, worry*] emotional words. Prosodically they speak faster, with higher pitch, but quieter (lower intensity min). Features of the alter (the woman) that helped our system detect men who say they are flirting include the woman's laughing, sexual words [*love, passion, virgin, sex, screw*] or swear words, talking more, and having a higher f0 (max).
>
> Women who say they are flirting have a much expanded pitch range (lower pitch min, higher pitch max), laugh more, use more I and well, use repair questions [*Wait, Excuse me*] but not other kinds of questions, use more sexual terms, use far less appreciations [*Wow, That's true, Oh, great*] and backchannels [*Uh-huh., Yeah., Right., Oh, okay.*], and use fewer, longer turns, with more words in general. Features of the alter (the man) that helped our system detect women who say they are flirting include the male use of you, questions, and faster and quieter speech.

This paper has changed the way I think about skillful heterosexual flirtation. I used to think that flirting was a unisex behavior, and that men and women were decently skilled at detecting it. In much the same way that it's harder to write a novel than to read one, I thought that the hard part was signalling your own intentions, not interpreting theirs.

Now, I think that a strategy for skillful flirtation is to get the other person to broadcast *their* intentions, and learn to interpret their signals correctly. Men and women have different flirting styles. Each person knows when they themselves are trying to flirt. But they're bad at guessing when their partner is trying to flirt. This suggests that if you can get your partner to engage in their own natural flirting style, and get good at

detecting it, then you can guess their intentions with much more confidence than the average person is capable of.

Both men and women should try to make each other laugh, let their voices be more musical, and provoke each other to talk about love and sex. They should tell stories about their lives and friendships and try to avoid mundane details.

A man who wants to signal flirtation to a woman should ask lots of questions that provoke the woman to talk about herself at length. Note that the "appreciations" and "backchannels" that are negatively correlated with women's flirtation are responses that women tend to give to men who keep going on about themselves. This is the old standard advice.

A woman who wants to signal flirtation to a man should maybe find topics they can complain about together - hopefully in a lighthearted way. She could also talk about her life in such a way that it provokes him to be curious and ask questions about her or observe connections between himself and her.

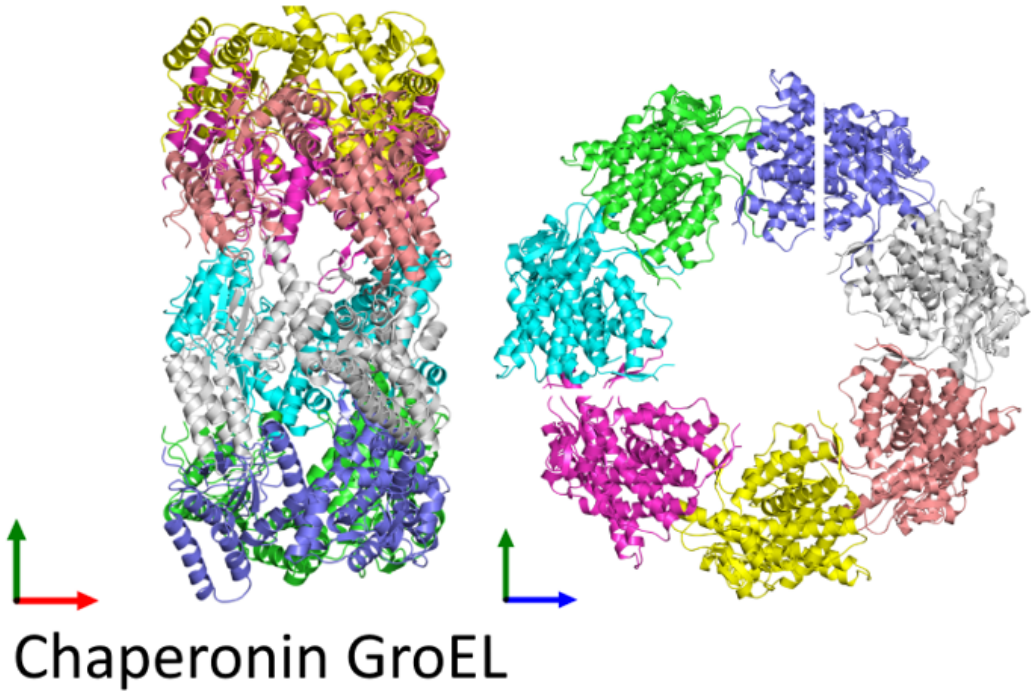
Not all communication is manipulation: Chaperones don't manipulate proteins

Epistemic status: Original work, explanation of a mental model that I developed for a few years that brings together knowledge from existing fields.

Is all communication manipulation? I hear this sentiment frequently expressed and want to explain in this article that there's nonmanipulative communication by using protein folding as an intuition pump.

It is common knowledge within molecular biology that proteins fold into their native state. That native state is the folded shape that possesses a minimum of free energy. Finding global minima is however a hard problem. For bigger proteins, it's at the time of writing - still impossible to calculate the shape.

Even in vivo protein folding is a hard problem. Cells are densely packed with many different molecules that push against each other. Frequently, resources are wasted when a protein misfolds into a shape that's not its native state.



Nature is clever and developed a way to help proteins fold into their native state. Cells produce chaperones. A chaperone surrounds an unfolded protein to protect it from outside influences to help the protein to fold into its native state. A chaperone doesn't need to know the native state of a protein to help the protein fold into that state. Instead of manipulating the protein like a sculpture, it holds space for a protein to be safe from outside influences, while it folds into its native form.

This allows a chaperone that works in an uncomplicated way to achieve a result that very complex machine learning algorithms currently don't achieve. The machine learning algorithm tries to figure out the best way for the protein to fold while the chaperone just lets the protein find this way by itself.

The psychologist Carl Rogers advocated that good psychologists act in the same way *nonmanipulative* with their patients. In his view, it's not the job of the therapist to solve the problem of their patient by manipulating the patient into a healthy form. A good therapist isn’t like ta sculptor sculpts a sculpture. The job of the therapist is rather to hold a space for the patient in which the patient is safe from certain forces that prevent the patient from finding their healthy authentic *native state*.

I don't intend to argue for *nonmanipulative communication* from a moral perspective. In cases where you know how to fix the problem of the person you are talking with and are confident that the other person will follow your advice, go ahead. If you don't know what will help a person, taking a nonmanipulative approach is often more effective than giving the person advice that they have already heard a dozen times.

If you tell an obese person that they should lose weight *again*, you add additional stress which can make it harder for them to think about the issue. In the Rogerian model effective change isn't about creating enough pressure by telling the obese to lose weight till they finally get it. For an obese person who feels shame for being obese, it can be hard to clearly think about the issue when they are alone. Providing the person a space where they can speak about their challenges in a way where they aren't feeling judged can help them to make progress for themselves.

There's a mystic quality to being *nonmanipulative*. Even Carl Rogers, who proposed the ideal, that all interactions should be nonmanipulative, sometimes fell short of it. For practical purposes it's often more useful to do what makes sense in the moment and what helps the other than to live up to an ideal of being perfectly nonmanipulative.

On the other hand, having a mental model of what it means to be *nonmanipulative* can be very helpful to understand communication practices like Rogerian psychotherapy, Gestalt Therapy and Circling.

I invite you to explore communicating in a way that holds the space for others to find themselves.

A Toy Model of Hingeyness

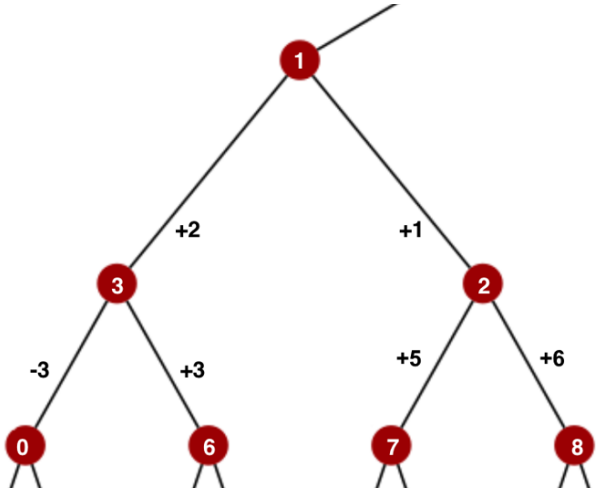
[This is a crosspost from the Effective Altruism forum]

Epistemic status: Attempt to clarify a vague concept. This should be seen as a jumping of point and not as a definitive model.

Definition of Hingeyness

> The **Hinge of History** refers to a time when we have an unusually high amount of influence over the future of civilization, compared to people who lived in the eras before and after ours.

I will use the model I made for my [previous question post] to explain why I don't think this definition is very useful. As before, in this model are only two possible choices per year. The number inside the circle refers to the amount of utility that year experiences and the two lines are the two options that this year has to decide on. The amount of utility which each option will add to the next year is written next to the lines. (*[link to image]*)



Older decisions are hingier

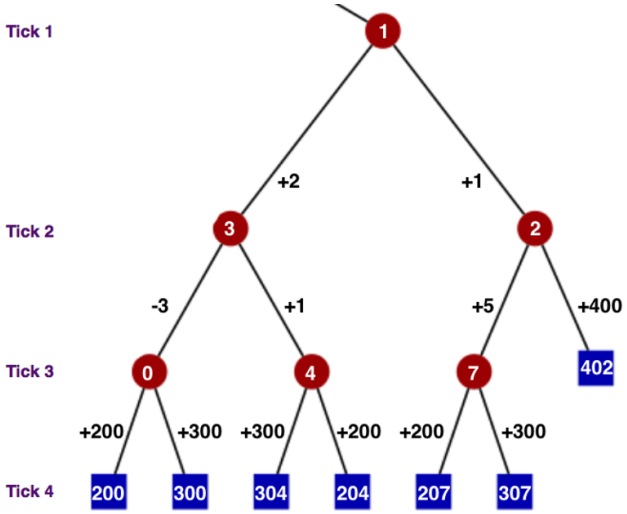
I think we all agree that we should try to avoid the option that will lead to better results in the next year, but will create less utility in the long run. In this model the year with 1 utility could choose the +2 option, but it should choose the +1 option because it leads to better options next year. Let's assume that all life dies after the last batch of years. The 1 utility then 3 utility then 0 utility option is the worst because you've generated 4 utility in total. 1-3-6 is just as good as 1-2-7, but 1-2- 8 is clearly the best path.

The implication is that later decisions are never hingier than earlier ones. 1 gets a range of options that ranges from 4 utility to 11 utility, no other option get's that kind of range. In fact, it's mathematically impossible that future decisions have a range of

options that's larger than the previous decisions had (assuming the universe will end and isn't some kind of loop). It's also mathematically impossible that future decisions have ranges where the best and worst case scenarios give you more utility than the range of the previous years. This is, unless negative utility is possible, which might arguably exist when you have a universe of beings being kept alive and tortured against their will (but it's rare in any case).

Decrease in range

Does that mean that hingeyness is now a useless concept? Not necessarily. The range will never grow, but the amount by which it narrows from year to year varies widely. Let's look at an extreme example. *(link to image)*



So the decisions made in 1 will always have the broadest range [204-405], but if you look at the difference in range between 3 [203-311] and 4 [208-311] it's not that much. So hingeyness may still be useful to think about how quickly our range is decreasing. It's even possible that the range doesn't shrink at all.

Going extinct quickly isn't necessarily bad

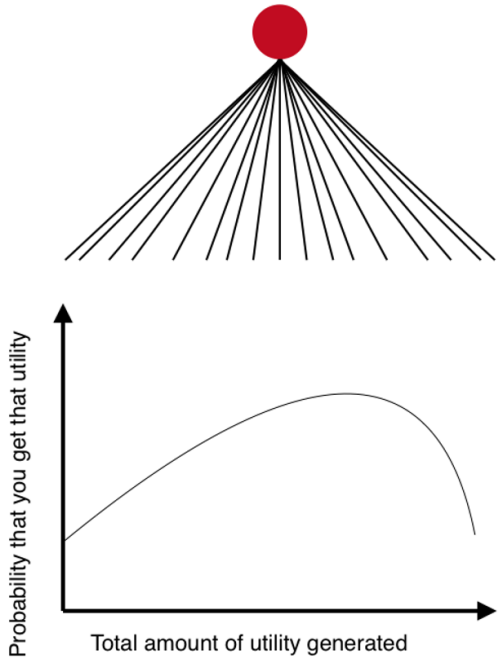
In the previous post I said that choosing for times where we survive for longer is almost always better (assuming you're a positive utilitarian and negative utility is impossible), this is an example of when this is not the case. The 1-2-402 chain gives the world the most utility even though it goes extinct one tick quicker. We (naturally) focus on reducing x-risk, but I wanted to visualize here why it might be possible that dying quickly in a blaze of utility is better than fizzling on for longer with low amounts of utility (especially if negative utility is possible). Although it should be noted that this model gives you clear ticks which might not exist in real life. Maybe planck time? Or maybe the time it takes to go from one state of pleasure to another a.k.a the time it takes to fire a neuron? Depending on how you answer that question this argument might fall flat.

Is hingeyness related to slack?

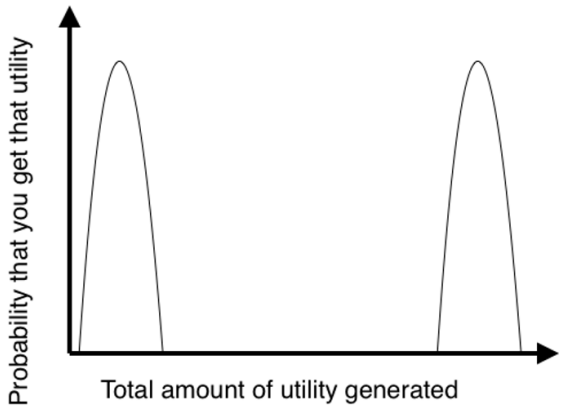
I'm starting to see similarities between the range of possible choices you keep and the amount of slack. I previously expressed that I see the slack/moloch trade-off as similar to the exploration/exploitation trade-off. Since we can't accurately predict which branches will give us the most utility it might be useful to keep a broad range of options open a.k.a to give yourself a lot of slack. In fact if we look at the first image you can see that someone who is pursuing linear utility exploitation will go from 1 to 3 (giving himself a +2 instead of a +1). Since this gives you worse results later this is basically the same thing as moloch pushing you into an inadequate equilibria. Having the slack/exploration to choose a sub-optimal route in the short-run but a better route in the long-run can only work if you have a lot of hingeyness.

How probability fits in

In reality of course you get more than two options, but the principle stays the same. Instead of a range you get a probability distribution. *(link to image)*



The probability that you get a certain amount of utility is equal to the amount of chains that generate that specific amount of utility (If you think certain chains have inherently less chance of existing you can just multiply the two factors). The range we are talking about is the difference between the lowest amount you could possibly generate and the highest. This will always either stay the same or shrink. This is not necessarily a bad thing as a we would rather face a narrow range of options between several good outcomes than a broad range of options between a lot of bad outcomes. But what about a distribution that looks like this *(link to image)*:



This is what I think a lot of people think about when we talk about the hinge of history; a time in history where the decisions we make can either turn out to have very good outcomes or very bad outcomes with very little in between. Our range may be smaller than the previous eras, but the probability that we either gain or lose lots of utility have never been higher. I won't decide what the "true definition of hingeyness" is since language belongs to it's users. I'm just pointing out that "the range of total amount of utility generated", "how quickly that range is decreasing" and "how polarized the probability distribution is" are very different concepts and we should probably have different labels for them. I will suggest three in the conclusion.

How much risk should we take?

I previously asked:

> When you are looking at the potential branches in the future, should you make the choice that will lead you to the cluster of outcomes with the highest average utility or to the cluster with the highest possible utility?

EmmaAbele answered:

> I'd say the one with the highest average utility if they are all equally likely. Basically, go with the one with the highest expected value.

But what about the cluster of branches with the median amount of utility, or mode or whatever? I don't think these questions have one definitively correct answer. Instead I would argue that we should use meta-preference utilitarianism to choose the options that most people want to choose.

Conclusion

There are three concepts that could be described as Hingeyness:

1) The range of the amount of utility you can potentially generate with your decision (maybe call it 'hinge broadness'?)

2) How much that range will narrow when you make a decision (maybe call it 'hinge reduction'?)

3) How polarized the probability is that you get either a lot or very little utility in the future (maybe call it 'hinge precipiceness'?)

Having lot's of "hinge broadness" is crucial for having slack. This toy model can be used to visualize all of these concepts.

Anthropomorphisation vs value learning: type 1 vs type 2 errors

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

The Occam's razor paper showed that one cannot deduce an agent H's reward function (R_H - using the notation from that paper) or their level of rationality (p_H) by observing their behaviour or even by knowing their policy (π_H). Subsequently, in a LessWrong post, it was demonstrated that even knowing the agent's full algorithm (call this a_H) would not be enough to deduce either R_H or p_H individually.

In an online video, I argued that the reason humans can do this when assessing other humans, is because we have an empathy module/theory of mind E_H, that allows us to model the rationality and motives of other humans. These E_H are, crucially, quite similar from human to human, and when we turn them on ourselves, the results are similar to what happens when others assess us. So, roughly speaking, there is an approximate 'what humans want', at least in typical environments[1], that most humans can agree on.

I struggled to convince people that, without this module, we would fail to deduce the motives of other humans. It is hard to imagine what we would be like if we were fundamentally different.

But there is an opposite error that people know very well: anthropomorphisation. In this situation, humans attribute motives to the behaviour of the wind, the weather, the stars, the stock market, cute animals, uncute animals...

So the same module that allows us to, somewhat correctly, deduce the motivations of other humans, also sets us up to fail for many other potential agents. If we started 'weakening' E_H, then we would reduce the number of anthropomorphisation errors we made, but we'd start making more errors about actual humans.

So our E_H can radically fail at assessing the motivations of non-humans, and also sometimes fails at assessing the motivations of humans. Therefore I'm relatively confident in arguing that E_H is not some "a priori" object, coming from pure logic, but is

contingent and dependent on human evolution. If we met an alien race, they we would likely assess their motives in ways they would find incorrect - and they'd assess our motives in ways we would find incorrect, no matter how much information either of us had.

---

1. See these posts for how we can and do extend this beyond typical environments. ↩

Human Biases that Obscure AI Progress

There are some common biases that are often used to discount AI progress. We should keep these in mind, as they can prevent us from having an objective understanding of progress in the field.

I'm going to use AI here instead of ML because usually these biases are relevant to any AI technique, not just ML. But in practice, ML is usually what I'm referring to.

1. AI uses the easiest solution it can find. Often people argue that a system isn't intelligent because it did some task in a simpler way than humans would have done it. This is especially applicable if there is some simple heuristic the AI found that did well enough. If finding the word "death" in a sentence is sufficient to do perfect classification, the AI will probably only learn to do that, no matter how intelligent it is. If your test set includes a case where that rule isn't sufficient, but the training set does not, the AI will probably fail on the test set case, because it has no way of knowing that "looking for the word death" wasn't the rule you wanted. AI will only do more complicated things once that simple thing is no longer sufficient, and it'll keep around the dumb heuristics whenever they work.

2. AI is different than us. Sometimes people argue that a system isn't intelligent because it does something in a more complicated way than is necessary (or just in a different way than what humans would have done, but not necessarily more or less complicated). Just because a specific rule is intuitive to us humans doesn't mean it's the easiest for an AI to find. A more complicated rule that also works and is easier for the AI to find will be found and used, no matter how intelligent an AI is. Regularization could in theory help with this, but only if regularization pushes towards the natural "human" approach, and only if the researchers continue training it after it's performance is very good. It's also not guaranteed that the rules that seem sensible to us are as good as what the AI does, our approach could be worse.

3. Many models will be better than humans in some ways and worse in others, and which aspects those are often won't be correlated with what we view as "difficult".

There are certain things that are easy for us to do (continuing to use the same name to refer to someone over a long story, preserving world details like the size of a horse) and things that are difficult for us to do (advanced reasoning, logic involving detailed steps, stories that involve 30 or more characters, etc.). We have a bias that the things that are easy for us to do should be easy for an AI to do, and the things that are hard for us to do should be hard for an AI to do. This leads to two important assumptions, both of which are *false*:

If an AI cannot do those easy things, it cannot do those difficult things.

If an AI can do those difficult things, it can easily do the easy things.

These days this bias is more apparent: visual recognition is much more difficult than playing chess, or solving complicated arithmetic expressions. But this point has deeper implications that I feel like some people miss. A model may become superhuman in all ways but one, and in that way it is still subhuman. If that subhuman

aspect is something that is easy for humans to do, we will write off the model as "not intelligent". **The key bias here is that our internal intuition for the difficulty of tasks should not be used to judge how intelligent something is, it can only be used to determine how "humanlike" something is in its way of thinking**. We can still use the intuition as a rough guiding pole for being surprised when AI does well at some task, but the point is that intelligence is multifaceted and the ordering of tasks solved by AI is not a given. A model could be better at theorem proving than any mathematician on earth, and simultaneously struggle to be consistent about the name of a character for more than a few paragraphs.

This is a really really important point. Superintelligent agents will probably still be dumb about *something* for a long time, even after they are dangerous. It's unreasonable to expect them to converge to our way of thinking: they function differently, so different things are easy and hard for us than they are for them.

4. AI models can be capable of being dumb or smart depending on the prompt. Generative models need to model *all* of human behaviour, including the stupid mistakes we all sometimes make. I've seen this one most concisely stated as "**sampling cannot be used to prove the absence of knowledge or ability, only the presence of it**". If your prompt doesn't give you what you want, it's possible that PEBCAK, and consider trying a different prompt, or trying different sampling settings.

5. World knowledge isn't intelligence. A superintelligent alien that landed on earth would not know whether a horse is larger than a toaster. Testing world knowledge is an interesting task, and being able to pick up world knowledge is an important sign of intelligence. But lack of world knowledge should not be used to discount intelligence or reasoning ability.

6. Exponential growth is counterintuitive. Somehow AI progress seems to be exponential in the same way that Moore's law is. This means that everything always seems too early to do anything about, right up until it is too late. In practice, most things are probably S curves that level off eventually, but where they level off may be far past the relevant danger points, so we should still try and keep this in mind.

7. Advertising/Celebrities. Significant progress made by small labs or independent researchers may not get nearly as much attention as progress made by big organizations or well known researchers. This is a difficult problem caused in part by the large amount of papers. In theory it could be helped by recommendation and exploration software to improve paper discovery, but either way the bias is important to keep in mind.

Let me know if you think of other biases and I'll add them to the list.

PS: When people claim that an AI is "memorizing and pattern matching" and not "truly understanding", in practice I find it usually comes down to either them referring to point 1 or 2, or when they see a model make a mistake a human wouldn't normally make (3).

Gems from the Wiki: Do The Math, Then Burn The Math and Go With Your Gut

This is a linkpost for https://www.lesswrong.com/tag/do-the-math-then-burn-the-math-and-go-with-your-gut

During the LessWrong 1.0 Wiki Import we (the LessWrong team) discovered a number of great articles that most of the LessWrong team hadn't read before. Since we expect many others to also not have have read these, we are creating a series of the best posts from the Wiki to help give those hidden gems some more time to shine.

The original wiki article was fully written by riceissa , who I've added as a coauthor to this post. Thank you for your work on the wiki!

"**Do the math, then burn the math and go with your gut**"[1] is a procedure for decision-making that has been described by Eliezer Yudkowsky. The basic procedure is to go through the process of assigning numbers and probabilities that are relevant to some decision ("do the math") and then to throw away this calculation and instead make the final decision with one's gut feelings ("burn the math and go with your gut"). The purpose of the first step is to force oneself to think through all the details of the decision and to spot inconsistencies.

History

In July 2008, Eliezer Yudkowsky wrote the blog post "When (Not) To Use Probabilities", which discusses the situations under which it is a bad idea to verbally assign probabilities. Specifically, the post claims that while theoretical arguments in favor of using probabilities (such as Dutch book and coherence arguments) always apply, humans have evolved algorithms for reasoning under uncertainty that don't involve verbally assigning probabilities (such as using "gut feelings"), which in practice often perform better than actually assigning probabilities. In other words, the post argues in favor of using humans' non-verbal/built-in forms of reasoning under uncertainty even if this makes humans incoherent/subject to Dutch books, because forcing humans to articulate probabilities would actually lead to worse outcomes. The post also contains the quote "there *are* benefits from trying to translate your gut feelings of uncertainty into verbal probabilities. It may help you spot problems like the conjunction fallacy. It may help you spot internal inconsistencies – though it may not show you any way to remedy them."[2]

In October 2011, LessWrong user bentarm gave an outline of the procedure in a comment in the context of the Amanda Knox case. The steps were: "(1) write down a list of all of the relevant facts on either side of the argument. (2) assign numerical weights to each of the facts, according to how much they point you in one direction or another. (3) burn the piece of paper on which you wrote down the facts, and go with your gut." This description was endorsed by Yudkowsky in a follow-up comment.

bentarm's comment claims that Yudkowsky described the procedure during summer of 2011.[3]

In December 2016, Anna Salamon described the procedure parenthetically at the end of a blog post. Salamon described the procedure as follows: "Eliezer once described what I take to be the a similar ritual for avoiding bucket errors, as follows: When deciding which apartment to rent (he said), one should first do out the math, and estimate the number of dollars each would cost, the number of minutes of commute time times the rate at which one values one's time, and so on. But at the end of the day, if the math says the wrong thing, one should do the right thing anyway."[4]

See also

- CFAR Exercise Prize – Andrew Critch's Bayes game, described on this page, gives another technique for dealing with uncertainty in real-life situations

External links

- A Facebook post by Julia Galef from May 2018 inquiring about this procedure
- "Why we can't take expected value estimates literally (even when they're unbiased)" (August 2011) by GiveWell co-founder Holden Karnofsky makes a similar point: "It's my view that my brain instinctively processes huge amounts of information, coming from many different reference classes, and arrives at a prior; if I attempt to formalize my prior, counting only what I can name and justify, I can worsen the accuracy a lot relative to going with my gut. Of course there is a problem here: going with one's gut can be an excuse for going with what one wants to believe, and a lot of what enters into my gut belief could be irrelevant to proper Bayesian analysis. There is an appeal to formulas, which is that they seem to be susceptible to outsiders' checking them for fairness and consistency."
- "The Optimizer's Curse & Wrong-Way Reductions" by Christian Smith discusses similar issues
- Verbal overshadowing page on Wikipedia

---

1. Qiaochu Yuan. "Qiaochu_Yuan comments on A Sketch of Good Communication". March 31, 2018. *LessWrong*.↩
2. Eliezer Yudkowsky. "When (Not) To Use Probabilities". July 23, 2008. *LessWrong*.↩
3. bentarm. "bentarm comments on Amanda Knox: post mortem". October 21, 2011. *LessWrong*.↩
4. Anna Salamon. "'Flinching away from truth' is often about *protecting* the epistemology". December 20, 2016. *LessWrong*.↩

What Does "Signalling" Mean?

[Epistemic status: shortly after writing this post, I thought I'd regret it. It was not rated very highly by LW karma, and on outside view, it's the sort of rant I often don't endorse later on. But re-reading it after a few weeks, I think it holds up. I still endorse everything I've written here, with the exception of my suggestion that "virtue signalling" is an appropriate replacement for what most people mean.]

I still feel a strong empathy for the post You Can't Signal to Rubes, which called out LessWrong for using the word "signalling" incorrectly. That post got heavily, and rightly, downvoted because ***it also got the definition wrong.*** *:(* But it had a point!

At the time of writing, the current definition of signalling on the LessWrong tag is:

> **Signaling** is behavior whose main purpose is to demonstrate to others that you possess some desirable trait. For example, a bird performing an impressive mating display signals that it is healthy and has good genes.

I'm not even sure I should correct it, because this *does* seem to summarize the LessWrong consensus on what signalling means. But we *already have* a term for signalling desirable properties about yourself: virtue signalling! Maybe you'll object that "virtue signalling" isn't quite right. Ok. But, could you find another word? I would prefer for "signalling" to point to the subject of signalling theory, which I understand to be the game theory of communication (often focusing on evolutionary game theory).

Scott Alexander's What Is Signaling, Really? seems to get most things right:

> In conclusion, a signal is a method of conveying information among not-necessarily-trustworthy parties by performing an action which is more likely or less costly if the information is true than if it is not true. Because signals are often costly, they can sometimes lead to a depressing waste of resources, but in other cases they may be the only way to believably convey important information.

Although all of his examples are about signalling self-properties, he never *stipulates* that, instead always using the more general conveying-information definition. He also avoids the *signalling is automatically bad* pitfall. Instead, he explains that signalling is often unfortunately costly, but is nonetheless a very useful tool.

However, reading it, I'm not sure whether he means to *contrast* signalling with "mere assertion", or whether he considers assertion to be a kind of signalling:

> Life frequently throws us into situations where we want to convince other people of something. If we are employees, we want to convince bosses we are skillful, honest, and hard-working. If we run the company, we want to convince customers we have superior products. If we are on the dating scene, we want to show potential mates that we are charming, funny, wealthy, interesting, you name it.
>
> In some of these cases, mere assertion goes a long way.
>
> [...]
>
> In other cases, mere assertion doesn't work.

> [...]

I'll charitably assume that he meant both cases to be types of signalling. But for anyone who was mislead by the wording: ***signalling is the theory of conveying information! Mere assertions, if they carry information, count as signalling!***

So, to summarize the points I've raised so far:

1. Sometimes people talk like signalling is just the bad thing (the dishonest or not-maximally-honest practice of making yourself look good).
2. Relatedly, people tend to exclude "mere assertion" from signalling, making signaling and literal use of language mutually exclusive.
3. Often people restrict signalling to signalling facts *about yourself*. (In fact, often restricted to *status* signalling.)

To be honest, I'm not even sure *academic* uses of the term "signalling" avoid the "mistakes" I'm pointing at! The Wikipedia article Signalling (economics) currently begins with the following:

> In contract theory, **signalling** (or **signaling**; see spelling differences) is the idea that one party (termed the agent) credibly conveys some information about itself to another party (the principal).

[Note that I've defaulted to the Wikipedia spelling of signalling; spelling on LessWrong seems mixed.]

On the other hand, the page on Signalling Theory (a page which is very biology-focused, despite the broader applicability of the theory) includes examples such as alarm calls (eg, birds warning each other that there is a snake in the grass). These signals cannot be interpreted as facts about the signaller.

Perhaps it is a quirk of *economics* which restricts the term "signalling" to hidden information *about the agent*, and LessWrong inherited this restricted sense via Robin Hanson?