Best of LessWrong: April 2018

1. A voting theory primer for rationalists
2. Noticing the Taste of Lotus
3. Local Validity as a Key to Sanity and Civilization
4. On exact mathematical formulae
5. I'm going to help you quit Facebook with some science
6. Announcement: AI alignment prize round 2 winners and next round
7. Adult Neurogenesis – A Pointed Review
8. The Chromatic Number of the Plane is at Least 5 - Aubrey de Grey
9. Is Rhetoric Worth Learning?
10. Raven Paradox Revisited
11. Idea: OpenAI Gym environments where the AI is a part of the environment
12. Schelling Shifts During AI Self-Modification
13. Reframing misaligned AGI's: well-intentioned non-neurotypical assistants
14. Critique my Model: The EV of AGI to Selfish Individuals
15. Announcing Rational Newsletter
16. [Draft for commenting] Near-Term AI risks predictions
17. [Preprint for commenting] Fighting Aging as an Effective Altruism Cause
18. On Equivalence of Supergoals
19. Double Cruxing the AI Foom debate
20. The many ways AIs behave badly
21. An Argument For Prioritizing: "Positively Shaping the Development of Crypto-assets"
22. Recommendations vs. Guidelines
23. 5 general voting pathologies: lesser names of Moloch
24. Can corrigibility be learned safely?
25. Hold On To The Curiosity
26. Some Simple Observations Five Years After Starting Mindfulness Meditation
27. Reward function learning: the learning process
28. My confusions with Paul's Agenda
29. The Alignment Newsletter #3: 04/23/18
30. Inefficient Doesn’t Mean Indifferent
31. Form Your Own Opinions
32. Community Page Mini-Guide
33. Reward function learning: the value function
34. Weird question: could we see distant aliens?
35. Announcing the Alignment Newsletter
36. Multi-winner Voting: a question of Alignment
37. Death in Groups
38. Believable Promises
39. Review of CZEA "Intense EA Weekend" retreat
40. Bounds of Attention
41. Corrigible but misaligned: a superintelligent messiah
42. Internal Diet Crux
43. Specification gaming examples in AI
44. Understanding Iterated Distillation and Amplification: Claims and Oversight
45. Ten Commandments for Aspiring Superforecasters
46. Implicit extortion
47. The First Rung: Insights from 'Linear Algebra Done Right'
48. Metaphilosophical competence can't be disentangled from alignment
49. My take on agent foundations: formalizing metaphilosophical competence
50. Global insect declines: Why aren't we all dead yet?

Best of LessWrong: April 2018

1. A voting theory primer for rationalists
2. Noticing the Taste of Lotus
3. Local Validity as a Key to Sanity and Civilization
4. On exact mathematical formulae
5. I'm going to help you quit Facebook with some science
6. Announcement: AI alignment prize round 2 winners and next round
7. Adult Neurogenesis – A Pointed Review
8. The Chromatic Number of the Plane is at Least 5 - Aubrey de Grey
9. Is Rhetoric Worth Learning?
10. Raven Paradox Revisited
11. Idea: OpenAI Gym environments where the AI is a part of the environment
12. Schelling Shifts During AI Self-Modification
13. Reframing misaligned AGI's: well-intentioned non-neurotypical assistants
14. Critique my Model: The EV of AGI to Selfish Individuals
15. Announcing Rational Newsletter
16. [Draft for commenting] Near-Term AI risks predictions
17. [Preprint for commenting] Fighting Aging as an Effective Altruism Cause
18. On Equivalence of Supergoals
19. Double Cruxing the AI Foom debate
20. The many ways AIs behave badly
21. An Argument For Prioritizing: "Positively Shaping the Development of Crypto-assets"
22. Recommendations vs. Guidelines
23. 5 general voting pathologies: lesser names of Moloch
24. Can corrigibility be learned safely?
25. Hold On To The Curiosity
26. Some Simple Observations Five Years After Starting Mindfulness Meditation
27. Reward function learning: the learning process
28. My confusions with Paul's Agenda
29. The Alignment Newsletter #3: 04/23/18
30. Inefficient Doesn’t Mean Indifferent
31. Form Your Own Opinions
32. Community Page Mini-Guide
33. Reward function learning: the value function
34. Weird question: could we see distant aliens?
35. Announcing the Alignment Newsletter
36. Multi-winner Voting: a question of Alignment
37. Death in Groups
38. Believable Promises
39. Review of CZEA "Intense EA Weekend" retreat
40. Bounds of Attention
41. Corrigible but misaligned: a superintelligent messiah
42. Internal Diet Crux
43. Specification gaming examples in AI
44. Understanding Iterated Distillation and Amplification: Claims and Oversight
45. Ten Commandments for Aspiring Superforecasters
46. Implicit extortion
47. The First Rung: Insights from 'Linear Algebra Done Right'
48. Metaphilosophical competence can't be disentangled from alignment

49. My take on agent foundations: formalizing metaphilosophical competence
50. Global insect declines: Why aren't we all dead yet?

A voting theory primer for rationalists

What is voting theory?

Voting theory, also called social choice theory, is the study of the design and evaulation of democratic voting methods (that's the activists' word; game theorists call them "voting mechanisms", engineers call them "electoral algorithms", and political scientists say "electoral formulas"). In other words, for a given list of candidates and voters, a voting method specifies a set of valid ways to fill out a ballot, and, given a valid ballot from each voter, produces an outcome.

(An "electoral system" includes a voting method, but also other implementation details, such as how the candidates and voters are validated, how often elections happen and for what offices, etc. "Voting system" is an ambiguous term that can refer to a full electoral system, just to the voting method, or even to the machinery for counting votes.)

Most voting theory limits itself to studying "democratic" voting methods. That typically has both empirical and normative implications. Empirically, "democratic" means:

- There are many voters
- There can be more than two candidates

In order to be considered "democratic", voting methods generally should meet various normative criteria as well. There are many possible such criteria, and on many of them theorists do not agree; but in general they do agree on this minimal set:

- Anonymity; permuting the ballots does not change the probability of any election outcome.
- Neutrality; permuting the candidates on all ballots does not change the probability of any election outcome.
- Unanimity: If voters universally vote a preference for a given outcome over all others, that outcome is selected. (This is a weak criterion, and is implied by many other stronger ones; but those stronger ones are often disputed, while this one rarely is.)
- Methods typically do not directly involve money changing hands or other enduring state-changes for individual voters. (There can be exceptions to this, but there are good reasons to want to understand "moneyless" elections.)

Why is voting theory important for rationalists?

First off, because democratic processes in the real world are important loci of power. That means that it's useful to understand the dynamics of the voting methods used in such real-world elections.

Second, because these real-world democratic processes have all been created and/or evolved in the past, and so there are likely to be opportunities to replace, reform, or

add to them in the future. If you want to make political change of any kind over a medium-to-long time horizon, these systemic reforms should probably be part of your agenda. The fact is that **FPTP, the voting method we use in most of the English-speaking world, is absolutely horrible**, and there is reason to believe that reforming it would substantially (though not of course completely) alleviate much political dysfunction and suffering.

Third, because understanding social choice theory helps clarify ideas about how it's possible and/or desirable to resolve value disputes between multiple agents. For instance, if you believe that superintelligences should perform a "values handshake" when meeting, replacing each of their individual value functions by some common one so as to avoid the dead weight loss of a conflict, then social choice theory suggests both questions and answers about what that might look like. (Note that the ethical and practical importance of such considerations is not at all limited to "post-singularity" examples like that one.)

In fact, on that third point: my own ideas of ethics and of fun theory are deeply informed by my decades of interest in voting theory. To simplify into a few words my complex thoughts on this, I believe that voting theory elucidates "ethical incompleteness" (that is, that it's possible to put world-states into ethical preference order partially but not fully) and that this incompleteness is a good thing because it leaves room for fun even in an ethically unsurpassed world.

What are the branches of voting theory?

Generally, you can divide voting methods up into "single-winner" and "multi-winner". Single-winner methods are useful for electing offices like president, governor, and mayor. Multi-winner methods are useful for dividing up some finite, but to some extent divisible, resource, such as voting power in a legislature, between various options. Multi-winner methods can be further subdivided into seat-based (where a set of similar "seats" are assigned one winner each) or weighted (where each candidate can be given a different fraction of the voting power).

What are the basics of single-winner voting theory?

(Note: Some readers may wish to skip to the summary below, or to read the later section on multi-winner theory and proportional representation first. Either is valid.)

Some of the earliest known work in voting theory was by Ramon Llull before his death in 1315, but most of that was lost until recently. Perhaps a better place to start would be in the French Academy in the late 1700s; this allows us to couch it as a debate (American Chopper meme?) between Jean-Charles de Borda and Nicolas de Condorcet.

> **Condorcet**: "Plurality (or 'FPTP', for First Past the Post) elections, where each voter votes for just one candidate and the candidate with the most votes wins, are often spoiled by vote-splitting."

Borda: "Better to have voters rank candidates, give candidates points for favorable rankings, and choose a winner based on points." (Borda Count)

Condorcet: "Ranking candidates, rather than voting for just one, is good. But your point system is subject to strategy. Everyone will rate some candidate they believe can't win in second place, to avoid giving points to a serious rival to their favorite. So somebody could win precisely because nobody takes them seriously!"

Borda: "My method is made for honest men!"

Condorcet: "Instead, you should use the rankings to see who would have a majority in every possible pairwise contest. If somebody wins all such contests, obviously they should be the overall winner."

In my view, Borda was the clear loser there. And most voting theorists today agree with me. The one exception is the mathematician Donald Saari, enamored with the mathematical symmetry of the Borda count. This is totally worth mentioning because his last name is a great source of puns.

But Condorcet soon realized there was a problem with his proposal too: it's possible for A to beat B pairwise, and B to beat C, while C still beats A. That is, pairwise victories can be cyclical, not transitive. Naturally speaking, this is rare; but if there's a decision between A and B, the voters who favor B might have the power to artificially create a "poison pill" amendment C which can beat A and then lose to B.

How would a Condorcet cycle occur? Imagine the following election:

1: A>B>C

1: B>C>A

1: C>A>B

(This notation means that there's 1 voter of each of three types, and that the first voter prefers A over B over C.) In this election, A beats B by 2 to 1, and similarly B beats C and C beats A.

Fast-forward to 1950, when theorists at the RAND corporation were inventing game theory in order to reason about the possibility of nuclear war. One such scientist, Kenneth Arrow, proved that the problem that Condorcet (and Llull) had seen was in fact a fundamental issue with any ranked voting method. He posed 3 basic "fairness criteria" and showed that no ranked method can meet all of them:

- Ranked unanimity: if every voter prefers X to Y, then the outcome has X above Y.
- Independence of irrelevant alternatives: If every voter's preferences between some subset of candidates remain the same, the order of those candidates in the outcome will remain the same, even if other candidates outside the set are added, dropped, or changed.
- Non-dictatorial: the outcome depends on more than one ballot.

Arrow's result was important in and of itself; intuitively, most people might have guessed that a ranked voting method could be fair in all those ways. But even more important than the specific result was the idea of an impossibility proof for voting.

Using this idea, it wasn't long until Gibbard and Satterthwaite independently came up with a follow-up theorem, showing that no voting system (ranked or otherwise) could possibly avoid creating strategic incentives for some voters in some situations. That is to say, there is no non-dictatorial voting system for more than two possible outcomes and more than two voters in which every voter has a single "honest" ballot that depends only on their own feelings about the candidates, such that they can't sometimes get a better result by casting a ballot that isn't their "honest" one.

There's another way that Arrow's theorem was an important foundation, particularly for rationalists. He was explicitly thinking about voting methods not just as real-world ways of electing politicians, but as theoretical possibilities for reconciling values. In this more philosophical sense, Arrow's theorem says something depressing about morality: if morality is to be based on (potentially revealed) preferences rather than interpersonal comparison of (subjective) utilities, it cannot simply be a democratic matter; "the greatest good for the greatest number" doesn't work without inherently-subjective comparisons of goodness. Amartya Sen continued exploring the philosophical implications of voting theory, showing that the idea of "private autonomy" is incompatible with Pareto efficiency.

Now, in discussing Arrow's theorem, I've said several times that it only applies to "ranked" voting systems. What does that mean? "Ranked" (also sometimes termed "ordinal" or "preferential") systems are those where valid ballots consist of nothing besides a transitive preferential ordering of the candidates (partial or complete). That is, you can say that you prefer A over B or B over A (or in some cases, that you like both of them equally), but you cannot say how strong each preference is, or provide other information that's used to choose a winner. In Arrow's view, the voting method is then responsible for ordering the candidates, picking not just a winner but a second place etc. Since neutrality wasn't one of Arrow's criteria, ties can be broken arbitrarily.

This excludes an important class of voting methods from consideration: those I'd call rated (or graded or evaluational), where you as a voter can give information about strength of preference. Arrow consciously excluded those methods because he believed (as Gibbard and Satterthwaite later confirmed) that they'd inevitably be subject to strategic voting. But since ranked voting systems are also inevitably subject to strategy, that isn't necessarily a good reason. In any case, Arrow's choice to ignore such systems set a trend; it wasn't until approval voting was reinvented around 1980 and score voting around 2000 that rated methods came into their own. Personally, for reasons I'll explain further below, I tend to prefer rated systems over purely ranked ones, so I think that Arrow's initial neglect of ranked methods got the field off on a bit of a wrong foot.

And there's another way I feel that Arrow set us off in the wrong direction. His idea of reasoning axiomatically about voting methods was brilliant, but ultimately, I think the field has been too focused on this axiomatic "Arrovian" paradigm, where the entire goal is to prove certain criteria can be met by some specific voting method, or cannot be met by any method. Since it's impossible to meet all desirable criteria in all cases, I'd rather look at things in a more probabilistic and quantitative way: how often and how badly does a given system fail desirable criteria.

The person I consider to be the founder of this latter, "statistical" paradigm for evaluating voting methods is Warren Smith. Now, where Kenneth Arrow won the Nobel Prize, Warren Smith has to my knowledge never managed to publish a paper in a peer-reviewed journal. He's a smart and creative mathematician, but... let's just say, not exemplary for his social graces. In particular, he's not reluctant to opine in varied

fields of politics where he lacks obvious credentials. So there's plenty in the academic world who'd just dismiss him as a crackpot, if they are even aware of his existence. This is unfortunate, because his work on voting theory is groundbreaking.

In his 2000 paper on "Range Voting" (what we'd now call Score Voting), he performed systematic utilitarian Monte-Carlo evaluation of a wide range of voting systems under a wide range of assumptions about how voters vote. In other words, in each of his simulations, he assumed certain numbers of candidates and of voters, as well as a statistical model for voter utilities and a strategy model for voters. Using the statistical model, he assigned each virtual voter a utility for each candidate; using the strategy model, he turned those utilities into a ballot in each voting method; and then he measured the total utility of the winning candidate, as compared to that of the highest-total-utility candidate in the race. Nowadays the name for the difference between these numbers, scaled so that the latter would be 100% and the average randomly-selected candidate would be 0%, is "Voter Satisfaction Efficiency" (VSE).

Smith wasn't the first to do something like this. But he was certainly the first to do it so systematically, across various voting methods, utility models, and strategic models. Because he did such a sensitivity analysis across utility and strategic models, he was able to see which voting methods consistently outperformed others, almost regardless of the specifics of the models he used. In particular, score voting, in which each voter gives each candidate a numerical score from a certain range (say, 0 to 100) and the highest total score wins, was almost always on top, while FPTP was almost always near the bottom.

More recently, I've done further work on VSE, using more-realistic voter and strategy models than what Smith had, and adding a variety of "media" models to allow varying the information on which the virtual voters base their strategizing. While this work confirmed many of Smith's results — for instance, I still consistently find that FPTP is lower than IRV is lower than approval is lower than score — it has unseated score voting as the undisputed highest-VSE method. Other methods with better strategy resistance can end up doing better than score.

Of course, something else happened in the year 2000 that was important to the field of single-winner voting theory: the Bush-Gore election, in which Bush won the state of Florida and thus the presidency of the USA by a microscopic margin of about 500 votes. Along with the many "electoral system" irregularities in the Florida election (a mass purge of the voter rolls of those with the same name as known felons, a confusing ballot design in Palm Beach, antiquated punch-card ballots with difficult-to-interpret "hanging chads", etc.) was one important "voting method" irregularity: the fact that Ralph Nader, a candidate whom most considered to be ideologically closer to Gore than to Bush, got far more votes than the margin between the two, leading many to argue that under almost any alternative voting method, Gore would have won. This, understandably, increased many people's interest in voting theory and voting reform. Like Smith, many other amateurs began to make worthwhile progress in various ways, progress which was often not well covered in the academic literature.

In the years since, substantial progress has been made. But we activists for voting reform still haven't managed to use our common hatred for FPTP to unite behind a common proposal. (The irony that our expertise in methods for reconciling different priorities into a common purpose hasn't let us do so in our own field is not lost on us.)

In my opinion, aside from the utilitarian perspective offered by VSE, the key to evaluating voting methods is an understanding of strategic voting; this is what I'd call

the "mechanism design" perspective. I'd say that there are 5 common "anti-patterns" that voting methods can fall into; either where voting strategy can lead to pathological results, or vice versa. I'd pose them as a series of 5 increasingly-difficult hurdles for a voting method to pass. Because the earlier hurdles deal with situations that are more common or more serious, I'd say that if a method trips on an earlier hurdle, it doesn't much matter that it could have passed a later hurdle. Here they are:

(0. Dark Horse. As in Condorcet's takedown of Borda above, this is where a candidate wins precisely because nobody expects them to. Very bad, but not a serious problem in most voting methods, except for the Borda Count.)

1. Vote-splitting / "spoiled" elections. Adding a minor candidate causes a similar major candidate to lose. Very bad because it leads to rampant strategic dishonesty and in extreme cases 2-party dominance, as in Duverger's Law. Problematic in FPTP, resolved by most other voting methods.

2. Center squeeze. A centrist candidate is eliminated because they have lost first-choice support to rivals on both sides, so that one of the rivals wins, even though the centrist could have beaten either one of them in a one-on-one (pairwise) election. Though the direct consequences of this pathology are much less severe than those of vote-splitting, the indirect consequences of voters strategizing to avoid the problem would be exactly the same: self-perpetuating two-party dominance. This problem is related to failures of the "favorite betrayal criterion" (FBC). Problematic in IRV, resolved by most other methods.

3. Chicken dilemma (aka Burr dilemma, for Hamilton fans). Two similar candidates must combine strength in order to beat a third rival. But whichever of the two cooperates less will be the winner, leading to a game of "chicken" where both can end up losing to the rival. This problem is related to failures of the "later-no-harm" (LNH) criterion. Because LNH is incompatible with FBC, it is impossible to completely avoid the chicken dilemma without creating a center squeeze vulnerability, but systems like STAR voting or 3-2-1 minimize it.

4. Condorcet cycle. As above, a situation where, with honest votes, A beats B beats C beats A. There is no "correct" winner in this case, and so no voting method can really do anything to avoid getting a "wrong" winner. Luckily, in natural elections (that is, where bad actors are not able to create artificial Condorcet cycles by strategically engineering "poison pills"), this probably happens less than 5% of the time.

Note that there's a general pattern in the pathologies above: the outcome of honest voting and that of strategic voting are in some sense polar opposites. For instance, under honest voting, vote-splitting destabilizes major parties; but under strategic voting, it makes their status unassailable. This is a common occurrence in voting theory. And it's a reason that naive attempts to "fix" a problem in a voting system by adding rules can actually make the original problem worse.

(I wrote a separate article with further discussion of these pathologies)

Here are a few of the various single-winner voting systems people favor, and a few (biased) words about the groups that favor them:

FPTP (aka plurality voting, or choose-one single-winner): Universally reviled by voting theorists, this is still favored by various groups who like the status quo in countries like

the US, Canada, and the UK. In particular, incumbent politicians and lobbyists tend to be at best skeptical and at worst outright reactionary in response to reformers.

IRV (Instant runoff voting), aka Alternative Vote or RCV (Ranked Choice Voting... I hate that name, which deliberately appropriates the entire "ranked" category for this one specific method): This is a ranked system where to start out with, only first-choice votes are tallied. To find the winner, you successively eliminate the last-place candidate, transferring those votes to their next surviving preference (if any), until some candidate has a majority of the votes remaining. It's supported by FairVote, the largest electoral reform nonprofit in the US, which grew out of the movement for STV proportional representation (see the multi-winner section below for more details). IRV supporters tend to think that discussing its theoretical characteristics is a waste of time, since it's so obvious that FPTP is bad and since IRV is the reform proposal with by far the longest track record and most well-developed movement behind it. Insofar as they do consider theory, they favor the "later-no-harm" criterion, and prefer to ignore things like the favorite betrayal criterion, summability, or spoiled ballots. They also don't talk about the failed Alternative Vote referendum in the UK.

Approval voting: This is the system where voters can approve (or not) each candidate, and the candidate approved by the most voters wins. Because of its simplicity, it's something of a "Schelling point" for reformers of various stripes; that is, a natural point of agreement as an initial reform for those who don't agree on which method would be an ideal end state. This method was used in Greek elections from about 1860-1920, but was not "invented" as a subject of voting theory until the late 70s by Brams and Fishburn. It can be seen as a simplistic special case of many other voting methods, in particular score voting, so it does well on Warren Smith's utilitarian measures, and fans of his work tend to support it. This is the system promoted by the Center for Election Science (electology.org), a voting reform nonprofit that was founded in 2012 by people frustrated with FairVote's anti-voting-theory tendencies. (Full disclosure: I'm on the board of the CES, which is growing substantially this year due to a significant grant by the Open Philanthropy Project. Thanks!)

Condorcet methods: These are methods that are guaranteed to elect a pairwise beats-all winner (Condorcet winner) if one exists. Supported by people like Erik Maskin (a Nobel prize winner in economics here at Harvard; brilliant, but seemingly out of touch with the non-academic work on voting methods), and Markus Schulze (a capable self-promoter who invented a specific Condorcet method and has gotten groups like Debian to use it in their internal voting). In my view, these methods give good outcomes, but the complications of resolving spoil their theoretical cleanness, while the difficulty of reading a matrix makes presenting results in an easy-to-grasp form basically impossible. So I personally wouldn't recommend these methods for real-world adoption in most cases. Recent work in "improved" Condorcet methods has showed that these methods can be made good at avoiding the chicken dilemma, but I would hate to try to explain that work to a layperson.

Bucklin methods (aka median-based methods; especially, Majority Judgment): Based on choosing a winner with the highest median rating, just as score voting is based on choosing one with the highest average rating. Because medians are more robust to outliers than averages, median methods are more robust to strategy than score. Supported by French researchers Balinski and Laraki, these methods have an interesting history in the progressive-era USA. Their VSE is not outstanding though; better than IRV, plurality, and Borda, but not as good as most other methods.

Delegation-based methods, especially SODA (simple optionally-delegated approval): It turns out that this kind of method can actually do the impossible and "avoid the Gibbard-Satterthwaite theorem in practice". The key words there are "in practice" — the proof relies on a domain restriction, in which voters honest preferences all agree with their favorite candidate, and these preference orders are non-cyclical, and voters mutually know each other to be rational. Still, this is the only voting system I know of that's 100% strategy free (including chicken dilemma) in even such a limited domain. (The proof of this is based on complicated arguments about convexity in high-dimensional space, so Saari, it doesn't fit here.) Due to its complexity, this is probably not a practical proposal, though.

Rated runoff methods (in particular STAR and 3-2-1): These are methods where rated ballots are used to reduce the field to two candidates, who are then compared pairwise using those same ballots. They combine the VSE advantages of score or approval with extra resistance to the chicken dilemma. These are currently my own favorites as ultimate goals for practical reform, though I still support approval as the first step.

Quadratic voting: Unlike all the methods above, this is based on the universal solvent of mechanism design: money (or other finite transferrable resources). Voters can buy votes, and the cost for n votes is proportional to n^2. This has some excellent characteristics with honest voters, and so I've seen that various rationalists think it's a good idea; but in my opinion, it's got irresolvable problems with coordinated strategies. I realize that there are responses to these objections, but as far as I can tell every problem you fix with this idea leads to two more.

TL; DR?

- Plurality voting is really bad. (Borda count is too.)
- Arrow's theorem shows no ranked voting method is perfect.
- Gibbard-Satterthwaite theorem shows that no voting method, ranked or not, is strategy-free in all cases.
- Rated voting methods such as approval or score can get around Arrow, but not Gibbard-Satterthwaite.
- Utilitarian measures, known as VSE, are one useful way to evaluate voting methods.
- Another way is mechanism design. There are (1+)4 voting pathologies to worry about. Starting from the most important and going down: (Dark horse rules out Borda;) vote-splitting rules out plurality; center squeeze would rule out IRV; chicken dilemma argues against approval or score and in favor of rated runoff methods; and Condorcet cycles mean that even the best voting methods will "fail" in a few percent of cases.

What are the basics of multi-winner voting theory?

Multi-winner voting theory originated under parliamentary systems, where theorists wanted a system to guarantee that seats in a legislature would be awarded in proportion to votes. This is known as proportional representation (PR, prop-rep, or #PropRep). Early theorists include Henry Droop and Charles Dodgson (Lewis Carroll).

We should also recognize Thomas Jefferson and Daniel Webster's work on the related problem of apportioning congressional seats across states.

Because there are a number of seats to allocate, it's generally easier to get a good answer to this problem than in the case of single-winner voting. It's especially easy in the case where we're allowed to give winners different voting weights; in that case, a simple chain of delegated voting weight guarantees perfect proportionality. (This idea has been known by many names: Dodgson's method, asset voting, delegated proxy, liquid democracy, etc. There are still some details to work out if there is to be a lower bound on final voting weight, but generally it's not hard to find ways to resolve those.)

When seats are constrained to be equally-weighted, there is inevitably an element of rounding error in proportionality. Generally, for each kind of method, there are two main versions: those that tend to round towards smaller parties (Sainte-Laguë, Webster, Hare, etc.) and those that tend to round towards larger ones (D'Hondt, Jefferson, Droop, etc.).

Most abstract proportional voting methods can be considered as greedy methods to optimize some outcome measure. Non-greedy methods exist, but algorithms for finding non-greedy optima are often considered too complex for use in public elections. (I believe that these problems are NP-complete in many cases, but fast algorithms to find provably-optimal outcomes in all practical cases usually exist. But most people don't want to trust voting to algorithms that nobody they know actually understands.)

Basically, the outcome measures being implicitly optimized are either "least remainder" (as in STV, single transferable vote), or "least squares" (not used by any real-world system, but proposed in Sweden in the 1890s by Thiele and Phragmen). STV's greedy algorithm is based on elimination, which can lead to problems, as with IRV's center-squeeze. A better solution, akin to Bucklin/median methods in the single-winner case, is BTV (Bucklin transferable vote). But the difference is probably not a big enough deal to overcome STV's advantage in terms of real-world track record.

Both STV and BTV are methods that rely on reweighting ballots when they help elect a winner. There are various reweighting formulas that each lead to proportionality in the case of pure partisan voting. This leads to an explosion of possible voting methods, all theoretically reasonable.

Because the theoretical pros and cons of various multi-winner methods are much smaller than those of single-winner ones, the debate tends to focus on practical aspects that are important politically but that a mathematician would consider trivial or ad hoc. Among these are:

- The role of parties. For instance, STV makes partisan labels formally irrelevant, while list proportional methods (widely used, but the best example system is Bavaria's MMP/mixed member proportional method) put parties at the center of the decision. STV's non-partisan nature helped it get some traction in the US in the 1920s-1960s, but the only remnant of that is Cambridge, MA (which happens to be where I'm sitting). (The other remnant is that former STV advocates were key in founding FairVote in the 1990s and pushing for IRV after the 2000 election.) Political scientist @jacksantucci is the expert on this history.
- Ballot simplicity and precinct summability. STV requires voters to rank candidates, and then requires keeping track of how many ballots of each type there are, with the number of possible types exceeding the factorial of the

number of candidates. In practice, that means that vote-counting must be centralized, rather than being performed at the precinct level and then summed. That creates logistical hurdles and fraud vulnerabilities. Traditionally, the way to resolve this has been list methods, including mixed methods with lists in one part. Recent proposals for delegated methods such as my PLACE voting (proportional, locally-accountable, candidate endorsement; here's an example) provide another way out of the bind.
- Locality. Voters who are used to FPTP (plurality in single-member districts) are used to having "their local representative", while pure proportional methods ignore geography. If you want both locality and proportionality, you can either use hybrid methods like MMP, or biproportional methods like LPR, DMP, or PLACE.
- Breadth of choice. Ideally, voters should be able to choose from as many viable options as possible, without overwhelming them with ballot complexity. My proposal of PLACE is designed to meet that ideal.

Prop-rep methods would solve the problem of gerrymandering in the US. I believe that PLACE is the most viable proposal in that regard: maintains the locality and ballot simplicity of the current system, is relatively non-disruptive to incumbents, and maximizes breadth of voter choice to help increase turnout.

Oh, I should also probably mention that I was the main designer, in collaboration with dozens of commenters on the website Making Light, of the proportional voting method E Pluribus Hugo, which is now used by the Hugo Awards to minimize the impact and incentives of bloc voting in the nominations phase.

Anticlimactic sign-off

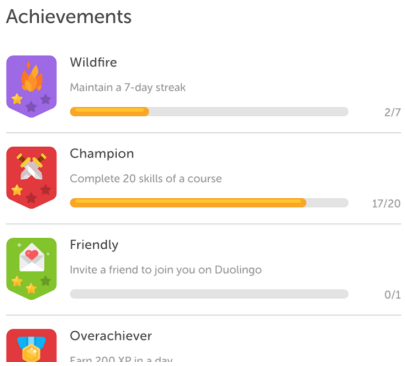
OK, that's a long article, but it does a better job of brain-dumping my >20 years of interest in this topic than anything I've ever written. On the subject of single-winner methods, I'll be putting out a playable exploration version of all of this sometime this summer, based off the work of the invaluable nicky case (as well as other collaborators).

I've now added a third article on this topic, in which I included a paragraph at the end asking people to contact me if they're interested in activism on this. I believe this is a viable target for effective altruism.

Noticing the Taste of Lotus

Recently I started picking up French again. I remembered getting something out of Duolingo a few years ago, so I logged in.

Since the last time I was there, they added an “achievements” mechanic:



I noticed this by earning one. I think it was “Sharpshooter”. They gave me the first of three stars for something like doing five lessons without mistakes. In the “achievements” section, it showed me that I could earn the *second* star by doing *twenty* lessons in a row flawlessly.

And my brain cared.

I watched myself hungering to get the achievements. These arbitrary things that someone had just stuck on there… in order to get me to want them. I noticed that I could get the second and maybe third star of “Sharpshooter” by doing earlier lessons and googling words and phrases I wasn’t quite sure about…

…which really doesn’t help me learn French.

Yes, we could quibble about that. Maybe perfect practice makes perfect, yada yada. But the point is: I disagree, I think my disagreement comes from knowing what I’m talking about when it comes to my learning, and someone’s arbitrary gold stars immediately overrode all that insight by grabbing my motivations directly.

I don’t have a problem with gamification per se. What bugs me here is that this specific gamification didn’t fit my goals, and that fact didn’t at all affect how well the system grabbed my wanting. I just… wanted those achievements. Because they were there.

If I hadn’t noticed this, and if I’m right about what I need to learn French, then I would have wasted a bunch of time pursuing a useless proxy goal. *And I would have felt pleasure in achieving it*. I might have even thought that was a meaningful sign that I was learning French — never mind that my goal of *holding my own in conversations* isn’t really helped by carefully avoiding typos.

Duncan Sabien sometimes talks about “lotus-eating”. He’s referring to a part of the Odyssey where they land on an island of “lotus-eaters”. It turns out that once you eat some of this kind of lotus, all you want to do is eat more. You stop caring about your other goals. The lotus just grabs your wants directly.

I claim you can notice when something grabs your wanting. Just… *look*. Just pay attention. Here are some lotuses I've noticed:

- Most computer games are full of these. I sometimes play one called Alto's Adventure. You flip a little character over and land some tricks, and then get a speed boost. If you collect enough coins, you can get special items or level them up to a maximum. If I start playing it, I notice I care about these arbitrary coins and flips and so on. And if I've been playing it recently, I notice myself wanting to pull the game out and play it some more. But what is gained by doing so? Maybe something, but if so then that's a happy accident. My life isn't any better after unlocking all the made-up achievements on this little made-up game. But each time I land a trick: BAM! A tiny burst of satisfaction, and a wanting to keep going.
- Scrolling down on Facebook. There's something about *wanting to scroll a little farther*. I get a "Yes!" and a "Just a little more" each time I scroll down and see a new post. Just another couple more minutes on Facebook, right? Oops.
- Email. Where does the impulse to check email several times a day come from? Or to "catch up" on email? What are you trying to *do*? What does it feel like when you've just clicked "Send"?
- Inbox zero in particular does this a lot for me. If I have just *two* emails, I want to reply to them right away, so I can get back to that oh so sacred inbox zero state. But then people *reply*, and I reply back, and my time gets eaten up… but at least I'm maintaining inbox zero, right?
- Porn is loudly lotuses. The website Your Brain On Porn goes into this a ton.
- YouTube has lotus nature. It's actually designed for it, just like Facebook. When you watch a video, it tries to guess what video you might want to watch next, and adjusts depending on what you click or sit through.
- The card game Dominion has a bunch of expansions. I found myself wanting to buy each expansion as it came out, because then my set would be *complete*, you see. Notice how the completeness is defined by *someone else*.

I think this kind of thing isn't very hard to notice if you try. What suddenly has you caring? What drives you into a kind of action? Just notice.

Also notice when someone else built the want-grabber. Their incentives are probably different from yours. If you don't pay attention, you'll get hijacked.

And then you're prone to rationalizing your addiction — like thinking that Facebook keeps you connected to your friends, but not really caring that maybe that's false.

I claim you can come to notice what lotuses taste like. Then you can choose to break useless addictions. And it'll feel good to do so: you're breaking free of distractions *and can tell*.

I find this gets easier if you give yourself permission to eat lotuses if you want to. Then I don't have to lie to myself about whether I am or not. I can just play Alto's Adventure, or clear out my email, or whatever, and it's fine. I just pay attention to the actual consequences — including the impact on what I later find myself *wanting to do* .

I ended up finding the taste of Duolingo's lotus disgusting. I could tell I wanted more, and that wanting was distracting me from my goal. I *could* do more, but now I just don't want to. It feels satisfying and empowering to resist the impulse to go back there and get one more star. I'm listening to French radio instead.

I invite you to learn what lotuses taste like, and reclaim your wanting for yourself.

Local Validity as a Key to Sanity and Civilization

(Cross-posted from Facebook.)

0.

Tl;dr: There's a similarity between these three concepts:

- A locally valid proof step in mathematics is one that, in general, produces only true statements from true statements. This is a property of a single step, irrespective of whether the final conclusion is true or false.
- There's such a thing as a bad argument even for a good conclusion. In order to arrive at sane answers to questions of fact and policy, we need to be curious about whether arguments are good or bad, independently of their conclusions. The rules against fallacies must be enforced even against arguments for conclusions we like.
- For civilization to hold together, we need to make coordinated steps away from Nash equilibria in lockstep. This requires general rules that are allowed to impose penalties on people we like or reward people we don't like. When people stop believing the general rules are being evaluated sufficiently fairly, they go back to the Nash equilibrium and civilization falls.

i.

The notion of a locally evaluated argument step is simplest in mathematics, where it is a formalizable idea in model theory. In math, a general type of step is 'valid' if it only produces semantically true statements from other semantically true statements, relative to a given model. If $x = y$ in some set of variable assignments, then $2x = 2y$ in the same model. Maybe x doesn't equal y, in some model, but even if it doesn't, the local step from "$x = y$" to "$2x = 2y$" is a locally valid step of argument. It won't introduce any *new* problems.

Conversely, $xy = xz$ does not imply $y = z$. It happens to work when $x = 2$, $y = 3$, and $z = 3$, in which case the two statements say "6 = 6" and "3 = 3" respectively. But if $x = 0$, $y = 4$, $z = 17$, then we have "0 = 0" on one side and "4 = 17" on the other. We can feed in a true statement and get a false statement out the other end. This argument is not *locally* okay.

You can't get the concept of a "mathematical proof" unless on some level—though often an intuitive level rather than an explicit one—you understand the notion of a single step of argument that is locally okay or locally not okay, independent of whether you globally agreed with the final conclusion. There's a kind of approval you give to the pieces of the argument, rather than looking the whole thing over and deciding whether you like what came out the other end.

Once you've grasped that, it may even be possible to convince you of mathematical results that sound counterintuitive. When your understanding of the rules governing allowable argument steps has become stronger than your faith in your ability to judge

whole intuitive conclusions, you may be convinced of truths you would not otherwise have grasped.

ii.

More generally in life, even outside of mathematics, there are such things as bad arguments for good conclusions.

There are even such things as genuinely good arguments for false conclusions, though of course those are much rarer. By the Bayesian definition of evidence, "strong evidence" is exactly that kind of evidence which we very rarely expect to find supporting a false conclusion. Lord Kelvin's careful and multiply-supported lines of reasoning arguing that the Earth could not possibly be so much as a hundred million years old, all failed simultaneously in a surprising way because that era didn't know about nuclear reactions. But most of the time this does not happen.

On the other hand, bad arguments for true conclusions are extremely easy to come by, because there are tiny elves that whisper them to people. There isn't anything the least bit more difficult in making an argument terrible when it leads to a good conclusion, since the tiny elves own lawnmowers.

One of the marks of an intellectually strong mind is that they are able to take a curious interest in whether a particular argument is a good argument or a bad argument, independently of whether they agree with the conclusion of that argument.

Even if they happen to start out believing that, say, the intelligence explosion thesis for Artificial General Intelligence is false, they are capable of frowning at the argument that the intelligence explosion is impossible because hypercomputation is impossible, or that there's really no such thing as intelligence because of the no-free-lunch theorem, and saying, "Even if I agree with your conclusion, I think that's a terrible argument for it." Even if they agree with the mainstream scientific consensus on anthropogenic global warming, they still wince and perhaps even offer a correction when somebody offers as evidence favoring global warming that there was a really scorching day last summer.

There are weaker and stronger versions of this attribute. Some people will think to themselves, "Well, it's important to use only valid arguments... but there was a sustained pattern of record highs worldwide over multiple years which does count as evidence, and that particular very hot day was a part of that pattern, so it's valid evidence for global warming." Other people will think to themselves, "I'd roll my eyes at someone who offers a single very cold day as an argument that global warming is false. So it can't be okay to use a single very hot day to argue that global warming is true."

I'd much rather buy a used car from the second person than the first person. I think I'd pay at least a 5% price premium.

Metaphorically speaking, the first person will court-martial an allied argument if they must, but they will favor allied soldiers when they can. They still have a sense of motion toward the Right Final Answer as being progress, and motion away from the right final answer as anti-progress, and they dislike not making progress.

The second person has something more like the strict mindset of a mathematician when it comes to local validity. They are able to praise some proof steps as obeying the rules, irrespective of which side those steps are on, without a sense that they are thereby betraying their side.

iii.

This essay has been bubbling in the back of my mind for a while, since I read that potential juror #70 for the Martin Shkreli trial was rejected during selection when, asked if they thought they could render impartial judgment, they replied, "I can be fair to one side but not the other." And I thought maybe I should write something about why that was possibly a harbinger of the collapse of civilization. I've been musing recently about how a lot of the standard Code of the Light isn't really written down anywhere anyone can find.

The thought recurred during the recent #MeToo saga when some Democrats were debating whether it made sense to kick Al Franken out of the Senate. I don't want to derail into debating Franken's behavior and whether that degree of censure was warranted per se, and I'll delete any such comments. What brought on this essay was that I read some unusually frank concerns from people who did think that Franken's behavior was per se cause to not represent the Democratic Party in the Senate; but who worried that the Democrats would police themselves, the Republicans wouldn't, and so the Republicans would end up controlling the Senate.

I've heard less of that since some upstanding Republican voters in Alabama stayed home on election night and put Doug Jones in the Senate.

But at the time, some people were replying, "That seems horrifyingly cynical and realpolitik. Is the idea here that sexual line-crossing is only bad and worthy of punishment when Republicans do it? Are we deciding that explicitly now?" And others were saying, "Look, the end result of your way of doing things is to just hand over the Senate to the Republican Party."

This is a conceptual knot that, I'm guessing, results from not explicitly distinguishing game theory from goodness.

There is, I think, a certain intuitive idea that *ideally* the Law is supposed to embody a subset of morality insofar as it is ever wise to enforce certain kinds of goodness. Murder is bad, and so there's a law against this bad behavior of murder. There's a lot of places where the law is in fact evil, like the laws criminalizing marijuana; that means the law is departing from its purpose, falling short of what it should be. Those who are not real-life straw authoritarians (who are sadly common) will cheerfully agree that there are some forms of goodness, even most forms of goodness, that it is not wise to try to legislate. But insofar as it *is* ever wise to make law, there's an intuitive sense that law should reflect some particular subset of morally good behavior that we have decided it is wise to enforce with guns, such as "Don't kill people."

It's from this perspective that "As a matter of pragmatic realpolitik we are going to not enforce sexual line-crossing rules against Democratic senators" seems like giving up, and maybe a harbinger of the fall of civilization if things have really gotten that bad.

But there's more than one function of legal codes, the way that money is both a store of value and a medium of exchange but these are different functions of money.

You can also look at laws as a kind of game theory played with people who might not share your morality at all. Some people take this perspective almost exclusively, at least in their verbal reports. They'll say, "Well, yes, I'd like it if I could walk into your house and take all your stuff, but I would dislike it even more if you could walk into my house and take *my* stuff, and that's why we have laws." I'm never quite sure how seriously to take the claim that they'd be happy walking into my house and taking my stuff. It seems to me that law enforcement and even social enforcement are simply not effective enough to count for the vast majority of human cooperation, and I have a sense that civilization is free-riding a whole lot on innate altruism... but game theory is certainly *a* function served by law.

The same way that money is both medium of exchange and store of value, the law is both collective utility function fragment and game theory.

In its function as game theory, the law (ideally) enables people with different utility functions to move from bad Nash equilibria to better Nash equilibria, closer to the Pareto frontier. Instead of mutual defection getting a payoff of (2, 2), both sides pay 0.1 for law enforcement and move to enforced mutual cooperation at (2.9, 2.9).

From this perspective, everything rests on notions like "fairness", "impartiality", "equality before the law", "it doesn't matter whose ox is being gored". If the so-called law punishes your defection but lets the other's defection pass, and this happens systematically enough and often enough, it is in your interest to blow up the current equilibrium if you have a chance.

It is coherent to say, "Crossing this behavioral line is universally bad when anyone does it, and also we're not going to punish Democratic senators unless you also punish Republican senators." Though as the saga of Senator Doug Jones of Alabama also shows, you should be careful about preemptively assuming the other side won't cooperate; there are sad lost opportunities there.

iv.

The way humans do law, it depends on the existence of what *feel like* simple general rules that apply to all cases.

This is not a universal truth of decision theory, it's a consequence of our cognitive limitations. Two superintelligences could negotiate a compromise with complicated detailed boundaries going right up to the Pareto frontier. They could agree on mutually verified pieces of cognitive code designed to intelligently decide future events according to known principles.

Humans use simpler laws than that.

To be clear, the kind of "law" I'm talking about here is not to be confused with the enormous modern morass of unreadable regulations. Think of, say, the written laws that actually got enforced in a small town in California in 1820. Or Democrats debating whether to enforce a sanction against Democratic senators if it's not being enforced against Republican senators. Or a small community's elders' star-chamber meeting to debate an accusation of sexual assault. Or the laws that cops will enforce even against other cops. These are the kinds of laws that must be simple in order to exist.

The reason that hunter-gatherer tribes don't have 100,000 pages of written legalism... is *not* that they've wisely realized that lengthy rules are easier to fill with loopholes, and that complicated regulations favor large corporations with legal departments, and that laws often have unintended consequences which don't resemble their stated justifications, and that deadweight losses increase quadratically. It's *very clear* that a supermajority of human beings are not that wise. Rather, hunter-gatherers just don't have enough time, energy, and paper to screw up that badly.

When humans try to verbalize The Law that isn't to be confused with written law, the law that cops will enforce against other cops, it comes out in universally quantified short sentences like "Anyone who defects in the Prisoner's Dilemma will be penalized TEN points even if that costs us fifteen" or "If you kill somebody who wasn't attacking you first, we'll exile you."

At one point somebody had the bright idea of trying to write down The Law. That way everyone could have common knowledge of what The Law was; and if you didn't break what was written, you could know you were safe from at least the official sanctions. Robert Heinlein called it the most important moment in political history, declaring that the law was above the politicians.

I for one rather doubt the Code of Hammurabi was universally enforced. I expect that hunter-gatherer tribes long before writing had a sense of there being Laws that were above the decisions of individual elders. I suspect that even in the best of times most of the The Law was never written down, and that more than half of what was written down was never really The Law.

But unfortunately, once somebody had the bright idea of writing down The Law, somebody else had the bright idea of writing down more words on the same clay tablet.

Today we live in a post-legalist era, when almost all of that which serves the true function of Law can no longer be written down. The government legalist system is too expensive in time and money and energy, too unreliable, and too slow, for any sane victim of sexual assault to appeal to the criminal justice system instead of the media justice system or the whispernet justice system. The civil legalist system outside of small claims court is a bludgeoning contest between entities that can afford lawyers, and the real law between corporations is enforced by merchant reputation and the threat of starting a bludgeoning contest. If you're in a lower-class neighborhood in the US, you can't get together and create order using your own town guards, because the police won't allow it. From your perspective, the function of the police is to prevent open gunfights and to not allow any more effective order than that to form.

But so it goes. We can't always keep the nice things we used to have, like written laws. The privilege was abused, and has been revoked.

When remains of The Law must indeed be simple, because our written-law privileges have been revoked, and so The Law relies on everyone knowing The Law without it being written down. It isn't even recited in memorable verse, as once it was. The Law relies on the community agreeing on the application of The Law without there being professional judges or a precedent-based judiciary. If not universal agreement, it must at least seem that the choices of the elders are trying to appeal to The Law instead of just naked self-interest. To the extent a voluntary association can't agree on The Law in this sense, it will soon cease to be a voluntary association.

The Law also breaks down if people start believing that, when the simple rules say one thing, the deciders will instead look at whose ox got gored, evaluate their personal interest, and enforce a different conclusion instead.

Which is to say: human law ends up with what people at least *believe* to be a set of simple rules that can be locally checked to test okay behavior. It's not actually algorithmically simple any more than walking is cheaply computable, but it feels simple the way that walking feels easy. Whatever doesn't feel like part of that small simple set won't be systematically enforced by the community, regardless of whether your civilization has reached the stage where police are seizing the cars of black people but not white people who use marijuana.

v.

The game-theoretic function of law can make following those simple rules feel like losing something, taking a step backward. You don't get to defect in the Prisoner's Dilemma, you don't get that delicious (5, 0) payoff instead of (3, 3). The law may punish one of your allies. You may be losing something according to your actual value function, which feels like the law having an objectively bad immoral result. You may coherently hold that the universe is a worse place for an instance of the enforcement of a good law, relative to its counterfactual state if that law could be lifted in just that instance without affecting any other instances. Though this does require seeing that law as having a game-theoretic function as well as a moral function.

So long as the rules are seen as moving from a bad global equilibrium to a global equilibrium seen as better, and so long as the rules are mostly-equally enforced on everyone, people are sometimes able to take a step backward and see that larger picture. Or, in a less abstract way, trade off the reified interest of The Law against their own desires and wishes.

This mental motion goes by names like "justice", "fairness", and "impartiality". It has ancient exemplars like a story I couldn't seem to Google, about a Chinese general who prohibited his troops from looting, and then his son appropriated a straw hat from a peasant; so the general sentenced his own son to death with tears running down his eyes.

Here's a fragment of thought as it was before the Great Stagnation, as depicted in passing in H. Beam Piper's *Little Fuzzy*, one of the earliest books I read as a child. It's from 1962, when the memetic collapse had started but not spread very far into science fiction. It stuck in my mind long ago and became one more tiny little piece of who I am now.

> “Pendarvis is going to try the case himself,” Emmert said. “I always thought he was a reasonable man, but what’s he trying to do now? Cut the Company’s throat?”
>
> “He isn’t anti-Company. He isn’t pro-Company either. He’s just pro-law. The law says that a planet with native sapient inhabitants is a Class-IV planet, and has to have a Class-IV colonial government. If Zarathustra is a Class-IV planet, he wants it established, and the proper laws applied. If it’s a Class-IV planet, the Zarathustra Company is illegally chartered. It’s his job to put a stop to illegality. Frederic Pendarvis’ religion is the law, and he is its priest. You never get anywhere by arguing religion with a priest.”

There is no suggestion in 1962 that the speakers are gullible, or that Pendarvis is a naif, or that Pendarvis is weird for thinking like this. Pendarvis isn't the defiant hero or even much of a side character. It's just a kind of judge you sometimes run into, part of a normal environment as projected from the author's mind that wrote the story.

If you don't have some people like Pendarvis, and you don't appreciate what they're trying to do even when they rule against you, sooner or later your tribe ends.

I mean, I doubt the United States will literally fall into anarchy this way before the AGI timeline runs out. But the concept applies on a smaller scale than countries. It applies on a smaller scale than communities, to bargains between three people or two.

The notion that you can "be fair to one side but not the other", that what's called "fairness" is a kind of favor you do for people you like, says that even the *instinctive* sense people had of law-as-game-theory is being lost in the modern memetic collapse. People are being exposed to so many social-media-viral depictions of the Other Side defecting, and viewpoints exclusively from Our Side without any leavening of any other viewpoint that might ask for a game-theoretic compromise, that they're losing the ability to appreciate the kind of anecdotes they used to tell in ancient China.

(Or maybe it's hormonelike chemicals leached from plastic food containers. Let's not forget all the psychological explanations offered for a wave of violence that turned out to be lead poisoning.)

vi.

And to take the point full circle:

The mental motion to evenhandedly apply The Rules irrespective of their conclusion is a kind of thinking that human beings appreciate intuitively, or at least they appreciated it in ancient China and mid-20th-century science fiction. In fact, we appreciate The Law more natively than we appreciate the notion of local syntactic rules capturing semantically valid steps in mathematical proofs, go figure.

So the legal metaphor is where a lot of people get started on epistemology: by seeing the local rules of valid argument as The Law, fallacies as crimes. The unusually healthy of mind will reject bad allied arguments with an emotional sense of practicing the way of an impartial judge.

It's ironic, in a way, because there is no game theory and no morality to the true way of the map that reflects the territory. A paperclip maximizer would also strive to debias its cognitive processes, alone in its sterile universe.

But I would venture a guess and hypothesis that you are better off buying a used car from a random mathematician than a random non-mathematician, even after controlling for IQ. The reasoning being that mathematicians are people whose sense of Law was strong enough to be appropriated for proofs, and that this will correlate, if imperfectly, with mathematicians abiding by what they see as The Law in other places as well. I could be wrong, and would be interested in seeing the results of any study like this if it were ever done. (But no studies on self-reports of criminal behavior, please. Unless there's some reason to believe that the self-report metric isn't measuring "honesty times criminality" rather than "criminality".)

I have no grand agenda in having said all this. I've just sometimes thought of late that it would be nice if more of the extremely basic rules of thinking were written down.

On exact mathematical formulae

This is inspired by the review on "Linear Algebra done right". I decided to do a top-level post, because it hits a misconception that is pretty common.

The starting point of this post is this quote from "Linear Algebra done right":

> Remarkably, mathematicians have proved that no formula exists for the zeros of polynomials of degree 5 or higher. But computers and calculators can use clever numerical methods to find good approximations to the zeros of any polynomial, even when exact zeros cannot be found.
>
> For example, no one will ever be able to give an exact formula for a zero of the polynomial p defined by $p(x) = x^5 - 5x^4 - 6x^3 + 17x^2 + 4^x - 7$.

The authors misrepresent an important point that is understood by most mathematicians, but not properly understood by many laypeople.

What does it mean to solve a problem? What does it mean to have an exact formula for the solution of a problem?

The answers to both are a social convention that has historically changed and is expected to continue to evolve in the future.

Back in the days, people only considered rational numbers, ie fractions. Oh, but what about the positive solution to $x^2 = 2$? Ok, we can't express this as a rational number (important theorem). Because these kinds of problems occured quite often, the mathematical community arrived at the consensus that $\sqrt{2}$, or more generally $\sqrt{r}$ for nonnegative r should be considered an explicit solution. Amazingly, this allows us to express the solution to any quadratic equation $0 = ax^2 + bx + c$ explicitly, with our expanded notion of "explicit". From an algebraic viewpoint it was natural to bless the positive solution to $x^n = a$ as an "explicit formula" next; historically it was a more contentious thing, because greek geometry wanted numbers to be constructible using a ruler and compass only. "Doubling the cube", ie expressing the positive solution to $x^3 = 2$ as a geometric construction was a famous old problem (proven impossible in 1837, after having been a very prominent mathematical research problems for more than 2000 years).

Now, this obviously says not a lot about the cube root of 2, but says a lot about "constructible with ruler and compass".

In other words: "Explicit solutions" are a messy historical map to mathematical territory, nothing more.

The same holds if you ask for explicit formulas for zeros of polynomials after having grudgingly admitted nth roots as "explicit". The same holds if you ask about explicit

integrals of explicit functions (also after having grudgingly admitted eg elliptic integrals as "explicit"). The same holds for solutions of differential equations.

In mathematics, asking about an "explicit formula" for solutions to problems means just: Assuming a general background in mathematics, is the solution something I already have spent years of my life developing an intuition for?

And if the answer happens to be "yes, unconditionally", then it is worthwhile.

If the "explicit" formula uses things that are not commonly taught anymore (crazy "special functions" that 100 years ago constituted a perfectly fine explicit solution), or is too lenghty/complicated to inform intuitions, then it is functionally equivalent to "we don't know", which is functionally equivalent to "we can prove that no formula using terms of type xyz exists".

So there is nothing surprising or scary about problems not having an "explicit" solution.

The true value of Galois theory is that it properly elucidates the hidden structure of polynomial equations, not that it tells us that no "explicit solution formula" exists for degree 5 polynomials for this very historical notion of "explicit". The "explicit" degree 4 formula is nothing more than a curiosity with interesting history, but absolutely worthless from both an intuitive and numerical standpoint.

I most often encountered the unjustified bias towards "explicit solutions" for implicit functions (the function $y = g(x)$ is defined by $f(x, g(x)) = 0$ for some fixed $f = f(x, y)$, implicit function theorem + newton solver) and solutions to differential equations. Integrals are mostly considered "explicit" today.

I'm going to help you quit Facebook with some science

Cross posted from http://bearlamp.com.au/im-going-to-help-you-quit-facebook-with-some-science/

I was a serial Facebook addict. I used to spend 2+ hours a day on Facebook, most days. Until I worked out how to change my mind.

Let's talk about the news feed. We all have this feeling that the news feed is drivel. Even curated, mine was still full of crud. Even super curated it was dull at best. Eventually I realised, something had to give.

As with many conflicts, indecision feels uncomfortable. Personally, I was super uncomfortable sitting in the cognitive dissonance of two conflicting beliefs:

Belief 1: Facebook is drivel and I want to spend less time on Facebook.

Belief 2: Facebook has good content from my friends that I want to keep up with.

There are three possible ways this can go. Either Facebook is in fact drivel and I will be happy to avoid it at any cost. Or Facebook has good in it and I'm staying around for the good stuff because I know it's worth it. Or Facebook is sometimes bad and sometimes good in some other complicated fashion, and I should check Facebook in some complicated intermittent fashion because of that...

This is how I worked out which belief was right.

You will need:

- your news feed
- pen and paper
- 5-10 minutes

Basic premise: Facebook has some good content and some bad content. But how much of each is ideal, acceptable or tolerable?

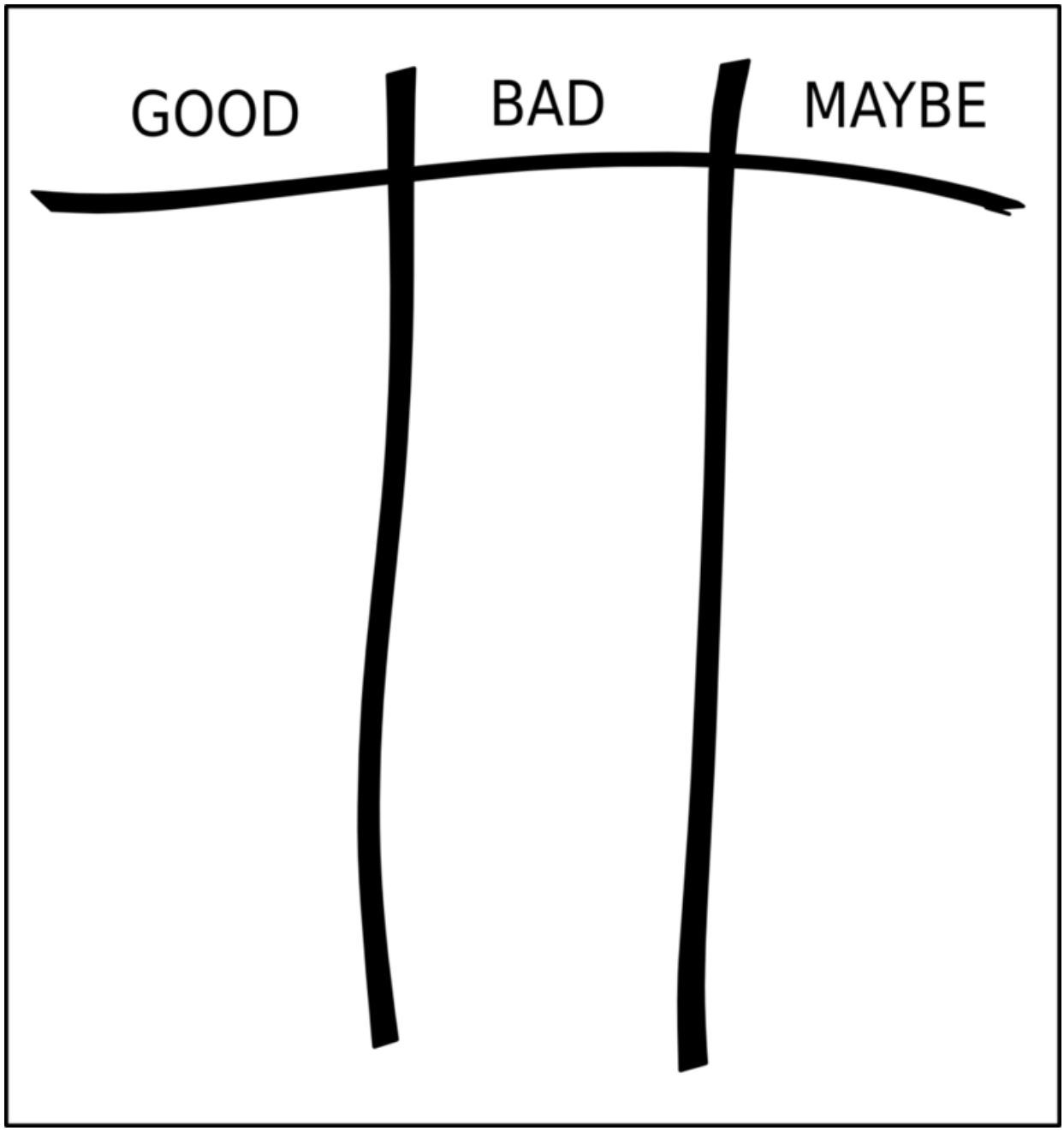
- If there were 10 good posts for every 3 bad posts, I might be willing to accept that. Maybe I can take some rubbish with the good! I should visit more often. 10:3
- If there was 1 good post to every bad post, I could still accept that. 1:1
- If there was 1 good post for every 5 bad posts, maybe I could suffer that. After all, not everything is perfect. 1:5
- But what if there was only 1 good post for every 10, 20, 30 bad posts? I don't think I'd be okay with that. 1:10
- And if it was worse - 1 good for every 50 bad - that would be enough to leave the platform. There's no point digging for diamonds in a dung heap. I'd rather just read a book. 1:50

Think about the possible ratios and write down your pre-commitment. What ratios do you consider ideal, acceptable, and unacceptable? Don't worry about getting it perfect;

you're an adult who can change their mind at any time. The purpose of writing down a ratio is to establish a baseline expectation before testing. I'm not coming to sneak up behind you and see your piece of paper and judge you if you change your mind. It’s fun to try set a belief and then test it. This is doing science.

Once you've decided your ideal, acceptable and unacceptable ratios, write down another number. What do you think the actual ratio is? Try to be honest. No one is looking.

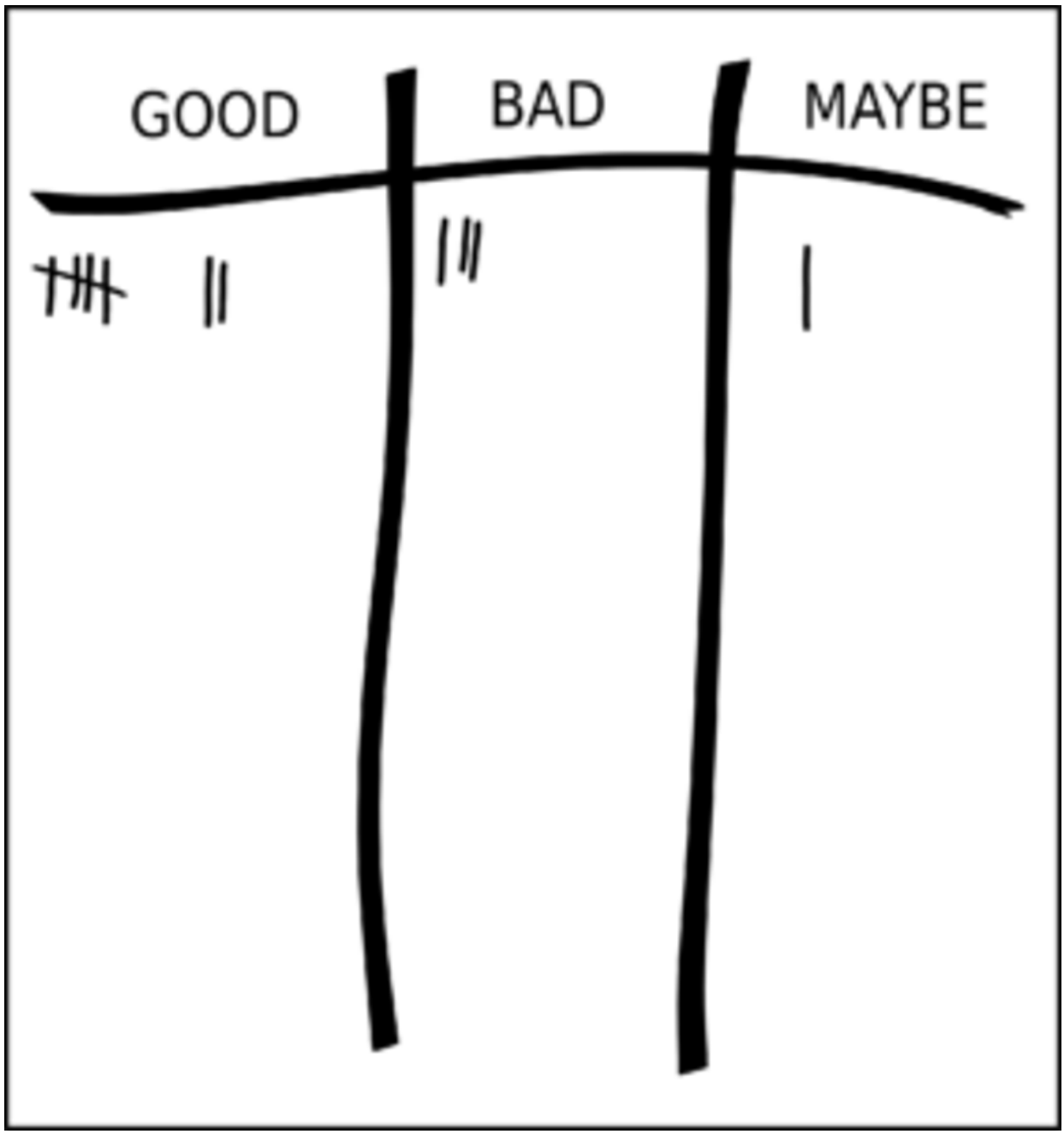
Next, draw this table on your page:



Now comes the part that takes the time. I want you to go down your news feed and I want you to count if you think posts are good or bad. There's also the "maybe" list for if you can't decide. Be honest with yourself. Try not to count posts a particular way. Try not to push the result somewhere. As a scientist, be curious about what it's going to be. No need to bias the results. It won't work as an exercise if you lie to yourself. It's also unnecessary to lie.

What's good and what's bad? I don't know. It’s subjective anyway. I can't tell you how to do that. You might want to think about:

- Do you want to hear this?
- Do you like the person it came from, is that enough to make the post good?
- Is it user generated content or is it shared content from elsewhere?
- is it "funny"?
- is it "news"?
- is it "happy"?
- is it "political"?
- Think of your own version of what factors might matter.
- Is this an ad?



And keep going. You can go to 100 posts, you can go for 5 minutes, go till you get bored. Up to you how you decide when to stop. With a warning: If you don't pre-determine the stopping rule you can bias the numbers a little. What if post 101 is a good one, you might stop at 101 not 100. That means you swayed everything a little more good on the ratios than you would otherwise have measured.

Then what? Count the columns. Then? Do nothing. It's just a ratio. We measured, but we don't have to do anything different.

How does this work? What happened? What did you do to me?

I say do nothing. But I am confident that you are going to allocate less time to Facebook than previously. Just naturally end up on there less often.

The trouble with feelings is that they are based in System 1. I have a hunch that Facebook is boring. But I don't always naturally know what to do with that hunch. The task above takes the feeling and brings that into System 2. We can count and measure the exact quantity of the feeling. Then maybe we can be better informed to act. And act you can.

Depending on what you uncover, you can choose what to do next. You now know exactly how good or bad your news feed was today. Take this information and choose to look at Facebook less, or choose to look at Facebook more. Maybe it's where you find all the good ideas! And any time you like - retest. See how your intuition for the site, mismatches or matches to reality.

Another interesting thing I noticed on Facebook:

If I hit the "like" button, Facebook tries it's valiant and hardest to show me more of that same thing. That means if it's pictures, Facebook will deliver slightly more pictures. If it's from a group, posts from that group. If it's from a friend, I'd be hearing more from that friend over the next week. I kind of stopped hitting the like buttons. Facebook doesn't need to know my likes. I also hate like notifications so much that I got a browser add-on that hides them from view and notification.

Now I wasn't hitting the like button so much. But I would still comment occasionally. I would watch flame wars as they happened. And I started getting interested in arguments. I tried to work out where they started. It seemed that they always started earlier than I expected. Well before a flame war, people are getting aggravated. But while that was interesting to learn, I wasn't commenting - I didn't want to interrupt the arguments that I was seeing, I was trying to be an objective observer. When I would make the occasional comment, same as likes, Facebook would deliver content that was similar to my future browsing.

if you are curious how the news feed will change, you can follow me in stopping likes and commenting. Try it. Also try other experiments. Science is fun!

Meta: I did this 6+ months ago and the results stuck. I spend a lot less time on Facebook than I used to. But the amazing thing is that the effect was almost overnight. It was obvious. My ratio was worse than 1:20. That was unacceptable to me.

When you try this, make sure you put down some predictions. Part of the scientific method is to make predictions and then test those predictions. In the process of modelling the world, predicting the future and generally being awesome, it’s okay to be wrong, it’s okay to be right - that’s why we test. It’s not okay to fake the results and lie to yourself. Wouldn’t it be something if you were surprised. Or confused. Or you changed your mind in response to the evidence you found.

Facebook and social media are becoming an entrenched part of our lives. Hopefully you ask the question about who's in control. Facebook is out to get you after all.

Announcement: AI alignment prize round 2 winners and next round

We (Zvi Mowshowitz and Vladimir Slepnev) are happy to announce the results of the second round of the AI Alignment Prize, funded by Paul Christiano. From January 15 to April 1 we received 37 entries. Once again, we received an abundance of worthy entries. In this post we name five winners who receive $15,000 in total, an increase from the planned $10,000.

We are also announcing the next round of the prize, which will run until June 30th, largely under the same rules as before.

The winners

First prize of $5,000 goes to Tom Everitt (Australian National University and DeepMind) and Marcus Hutter (Australian National University) for the paper The Alignment Problem for History-Based Bayesian Reinforcement Learners. We're happy to see such a detailed and rigorous write up of possible sources of misalignment, tying together a lot of previous work on the subject.

Second prize of $4,000 goes to Scott Garrabrant (MIRI) for these LW posts:

- Robustness to Scale
- Sources of Intuitions and Data on AGI
- Don't Condition on no Catastrophes
- Knowledge is Freedom

Each of these represents a small but noticeable step forward, adding up to a sizeable overall contribution. Scott also won first prize in the previous round.

Third prize of $3,000 goes to Stuart Armstrong (FHI) for his post Resolving human values, completely and adequately and other LW posts during this round. Human values can be under-defined in many possible ways, and Stuart has been very productive at teasing them out and suggesting workarounds.

Fourth prize of $2,000 goes to Vanessa Kosoy (MIRI) for the post Quantilal control for finite MDPs. The idea of quantilization might help mitigate the drawbacks of extreme optimization, and it's good to see a rigorous treatment of it. Vanessa is also a second time winner.

Fifth prize of $1,000 goes to Alex Zhu (unaffiliated) for these LW posts:

- Reframing misaligned AGI's: well-intentioned non-neurotypical assistants
- Metaphilosophical competence can't be disentangled from alignment
- Corrigible but misaligned: a superintelligent messiah
- My take on agent foundations: formalizing metaphilosophical competence

Alex's posts have good framings of several problems related to AI alignment, and led to a surprising amount of good discussion.

We will contact each winner by email to arrange transfer of money.

We would also like to thank everyone else who sent in their work! The only way to make progress on AI alignment is by working on it, so your participation is *the whole point*.

The next round

We are now also announcing the third round of the AI alignment prize.

We're looking for technical, philosophical and strategic ideas for AI alignment, posted publicly between January 1, 2018 and June 30, 2018 and not submitted for previous iterations of the AI alignment prize. You can submit your entries in the comments here or by email to apply@ai-alignment.com. We may give feedback on early entries to allow improvement, though our ability to do this may become limited by the volume of entries.

The minimum prize pool will again be $10,000, with a minimum first prize of $5,000.

Thank you!

Adult Neurogenesis – A Pointed Review

[I am not a neuroscientist and apologize in advance for any errors in this article.]

Hey, let's review the literature on adult neurogenesis! This'll be really fun, promise.

Gage's Neurogenesis In The Adult Brain, published in the *Journal Of Neuroscience* and cited 834 times, begins:

> A milestone is marked in our understanding of the brain with the recent acceptance, contrary to early dogma, that the adult nervous system can generate new neurons. One could wonder how this dogma originally came about, particularly because all organisms have some cells that continue to divide, adding to the size of the organism and repairing damage. All mammals have replicating cells in many organs and in some cases, notably the blood, skin, and gut, stem cells have been shown to exist throughout life, contributing to rapid cell replacement. Furthermore, insects, fish, and amphibia can replicate neural cells throughout life. An exception to this rule of self-repair and continued growth was thought to be the mammalian brain and spinal cord. In fact, because we knew that microglia, astrocytes, and oligodendrocytes all normally divide in the adult and respond to injury by dividing, it was only neurons that were considered to be refractory to replication. Now we know that this long accepted limitation is not completely true

Subsequent investigation has found adult neurogenesis in all sorts of brain regions. Wikipedia notes that "In humans, new neurons are continually born throughout adulthood in two regions of the brain: the subgranular zone and the striatum", but adds that "some authors (particularly Elizabeth Gould) have suggested that adult neurogenesis may also occur in regions within the brain not generally associated with neurogenesis including the neocortex", and there's also some research pointing to the cerebellum.

Some research has looked at the exact mechanism by which neurogenesis takes place; for example, in a paper in *Nature* cited 1581 times, Song et al determine that astroglia have an important role in promoting neurogenesis from FGF-2-dependent stem cells. Other research has tried to determine the rate; for example, Cameron et al (1609 citations) find that there is "a substantial pool of immature granule neurons" that may generate as many as 250,000 new cells per month. Still other research looks at the chemical regulators – a study by Lie et al, cited 1312 times, finds that Wnt3 signaling is involved.

(which is making you more nervous – the fact that I keep emphasizing how many citations these studies have, or the fact that one of the principal investigators is named "Lie"?)

But the most exciting research has been the work identifying the many important roles that neurogenesis plays in the adult brain – roles vital in understanding learning, memory, and disease.

Snyder et al (775 citations) finds "a new role for adult neurogenesis in the formation and/or consolidation of long-term, hippocampus-dependent, spatial memories."

Dupret et al go further and find that "spatial relational memory *requires* hippcampal adult neurogenesis". Aimone et al (633 citations) find "a possible role" for adult neurogenesis in explaining the "temporal clusters of long-term episodic memories seen in some human psychological studies". And Jessberger et al (506 citations) finds a role in object recognition memory as well.

In terms of learning, one of the major studies was Gould et al in Nature Neuroscience (2207 citations) finding that Learning Enhances Adult Neurogenesis In The Hippocampal Formation. Lledo et al (1288 citations) find that neurogenesis plays a part in explaining the brain's amazing plasticity, and is "highly modulated, revealing a plastic mechanism by which the brain's performance can be optimized for a given environment". Clemenson et al (17 citations) find that "from mice to humans", enviromental enrichment improves neurogenesis, and this "may one day lead us to a way to enrich our own lives and enhance performance on hippocampal behaviors".

But I've always been most interested in the link with depression. In 2000, Jacobs et al published Adult Brain Neurogenesis And Psychiatry: A Novel Theory Of Depression (961 citations). It's important enough that I want to quote the whole abstract:

> Neurogenesis (the birth of new neurons) continues postnatally and into adulthood in the brains of many animal species, including humans. This is particularly prominent in the dentate gyrus of the hippocampal formation. One of the factors that potently suppresses adult neurogenesis is stress, probably due to increased glucocorticoid release. Complementing this, we have recently found that increasing brain levels of serotonin enhance the basal rate of dentate gyrus neurogenesis. These and other data have led us to propose the following theory regarding clinical depression. Stress-induced decreases in dentate gyrus neurogenesis are an important causal factor in precipitating episodes of depression. Reciprocally, therapeutic interventions for depression that increase serotonergic neurotransmission act at least in part by augmenting dentate gyrus neurogenesis and thereby promoting recovery from depression. Thus, we hypothesize that the waning and waxing of neurogenesis in the hippocampal formation are important causal factors, respectively, in the precipitation of, and recovery from, episodes of clinical depression.

This theory got a boost from studies like Duman et al (522 citations), which found that antidepressant drugs like SSRIs upregulated neurogenesis – could this be their mechanism of action? And Ernst et al (327 citations) find that "there is evidence to support the hypothesis that exercise alleviates MDD and that several mechanisms exist that could mediate this effect through adult neurogenesis" – ie the antidepressant effects of exercise seem to work this way too. Electroconvulsive therapy, the most effective known treatment for depression? Works by promoting adult neurogenesis, at least according to Schloesser et al.

Is there *anything* that doesn't have important neurogenesis-related effects? It would seem there is not. Sex, for example, "promotes adult neurogenesis in the hippocampus, despite an initial elevation in stress hormones" according to Leuner et al (124 citations). Drug addiction is modulated by neurogenesis. We need rock n' roll to complete the triad, so here's Music Faciliates The Neurogenesis, Regeneration, and Repair of Neurons.

A study in Nature Neuroscience that garnered over 3000 citations found that running increased neurogenesis. The popular science press was quick to notice. A slew of exercise-neurogenesis studies spawned articles like Psychology Today's More Proof

That Aerobic Exercise Can Make Your Brain Bigger. Dr. Perlmutter ("Empowering Neurologist!") has a video about how you can Grow New Brain Cells Through Exercise. After this the pop sci world might have gotten a *little* carried away, until neurogenesis controls everything and is controlled by everything in turn. Slimland (of course there's a site called Slimland) has a How To Grow New Brain Cells And Stimulate Neurogenesis page, suggesting you can "set yourself free and start flying" by removing toxins, eating a ketogenic diet, and meditating. Naturalstacks.com boasts 11 Proven Ways To Generate More Brain Cells, Improve Memory, And Boost Mood, which advises...really? Do you really want to know what it advises? Come *on*.

Also, growth mindset. Of *course* growth mindset. Carol Dweck's Mindsetworks helpfully provides an infographic for teachers, urging them to tell their students that each time they set a goal or become motivated to learn a new skill, "a new neuron is formed through a process called neurogeneis" [sic].

So it's no surprise that researchers in the area are calling adult neurogenesis "one of the most exciting and rapidly evolving areas of research in the field of neuroscience".

II.

Fun fact: there's no such thing as adult neurogenesis in humans.

At least, this is the conclusion of Sorrells et al, who have a new and impressive study in *Nature*. They look at "59 post-mortem and post-operative slices of the human hippocampus" and find "that recruitment of young neurons to the primate hippocampus decreases rapidly during the first years of life, and that neurogenesis in the dentate gyrus does not continue, or is extremely rare, in adult humans." Also, the subgranular zone, the supposed part of the brain where neurogenesis begins, isn't even a real structure.

I am not a neuroscientist and am unqualified to evaluate it. But the Neuroskeptic blog, which I tend to trust in issues like this, thinks it's legit and has been saying this for years. Ed Yong from *The Atlantic* has a really excellent review of the finding that interviews a lot of the major players on both sides and which I highly recommend. Both of these reinforce my feeling that the current study makes a really strong case.

I was kind of floored when I saw this, in a way that I hope I was able to replicate in you by preceding this with the literature review above. How do you get so many highly-cited papers speaking so confidently about every little sub-sub-detail of a phenomenon, if the phenomenon never existed in the first place?

As far as I can tell, this was entirely innocent, well-intentioned, and understandable. It happened like this:

Adult neurogenesis was discovered in rats. This was so surprising, and such a violation of established doctrine, that it quickly became one of the most-investigated areas in neuroscience. Hundreds of studies were done on rats to nail down every little detail of the process.

The work was extended to many other mammals, to the point where it seemed inevitable that it must be true of humans as well. This was difficult to test because the relevant studies involve dissecting brains, and there aren't that many human brain specimens available with the necessary level of preservation. After a lot of work, a few

people got a couple of brains, did some very complicated and contamination-prone tests, and found evidence of adult neurogenesis. This encouraged everyone to assume that the things they had discovered about rat neurogenesis were probably true in humans as well, even though they could never prove them directly because of the difficulty of human experimentation. Later some other researchers tried to replicate the complicated and contamination-prone tests and couldn't find adult neurogenesis in humans, but everyone assumed they had just messed up some aspect of the complicated testing process.

And to complicate matters, everyone in the new study has been very careful to say they can't prove with certainty that zero adult neurogenesis occurs – just that the levels are so low and hard to detect that they can't possibly matter. Looking back on some past studies, it seems that "so low and hard to detect that they can't possibly matter" was actually within their confidence intervals. So it may be that some team found some extremely tiny and irrelevant population of immature neurons in the brain, gave a confidence level that included that number, and then everyone just assumed we were talking about levels similar to the ones we saw in rats.

With real scientists taking not-entirely-sufficient care to distinguish rat from human results, the popular press felt licensed to totally ignore the distinction (did you even notice which of the studies in Part I were done on which species? Don't worry, nobody else did either).

Meanwhile, synaptogenesis – the growth of new synapses from existing nerve cells – was getting linked to depression and all kinds of other things in a lot of interesting studies. When people started talking about neurogenesis' role in depression, psychiatrists like me who have trouble keeping words ending with -genesis separate just sort of nodded and said "Oh, yeah, I heard about that" and didn't give it the sort of scrutiny it deserved.

(I wonder if this is young-earth creationists' problem too)

So it's not like any one person made a spectacular mistake anywhere along the lines. Most of the studies done were in rats, and 100% correct. A few studies were done in humans, and may have gotten the wrong answer in a very difficult domain, while also hedging their bets and admitting they were trying something hard. It was only on a structural, field-wide level that all of this came together into people just assuming that adult human neurogenesis had to happen and be important.

…or at least, that's the optimistic take on it. But I can't help thinking – antidepressants work in humans, which suggests that the people who found neurogenesis was necessary for antidepressant effects must have just been plain wrong. And if exercise has antidepressant effects in humans, then the claim that those effects are neurogenesis-mediated must be wrong too. And, uh, humans form spatial and temporal memories, so unless we do this by a totally different mechanism than the ones rats use, people must have been wrong when they said neurogenesis was involved in that. ECT? Works in humans. Brain plasticity? Happens in humans. So maybe it would be better to say that the original claim that adult neurogenesis happens in humans seems innocent and understandable – but if the new study is true, that suggests that a lot of the followup claims must have been imaginary. Anything that focuses on a process that happens in humans and says "neurogenesis causes this" must not only be wrong to extend the results to humans, but must be under strong suspicion of being wrong even about rats, unless rat brains and human brains

accomplish the same basic tasks through totally different mechanisms (eg antidepressants work on rats but for different reasons than in humans).

We know many scientific studies are false. But we usually find this out one-at-a-time. This – again, assuming the new study is true, which it might not be – is a massacre. It offers an unusually good chance for reflection.

And looking over the brutal aftermath, I'm struck by how prosocial a lot of the felled studies are. Neurogenesis shows you should exercise more! Neurogenesis shows antidepressants work! Neurogenesis shows we need more enriched environments! Neurogenesis proves growth mindset! I'm in favor of exercise and antidepressants and enriched environments, but this emphasizes how if we want to believe something, it will accrete a protective layer of positive studies whether it's true or not.

I'm also struck by how many of the offending studies begin by repeating how dogmatic past neuroscientists were for not recognizing the existence of adult neurogenesis sooner. Remember Gage's review above:

> A milestone is marked in our understanding of the brain with the recent acceptance, contrary to early **dogma**, that the adult nervous system can generate new neurons. One could wonder how this **dogma** originally came about…

Or from Neurogenesis In Adult CNS: From Denial To Opportunities And Challenges For Therapy:

> The discovery of neurogenesis and neural stem cells (NSC) in the adult CNS has overturned a long-standing and deep-routed **"dogma"** in neuroscience, established at the beginning of the 20th century. This **dogma** lasted for almost 90 years and died hard when NSC were finally isolated from the adult mouse brain. The scepticism in accepting adult neurogenesis has now turned into a rush to find applications to alleviate or cure the devastating diseases that affect the CNS.

From Adult Human Neurogenesis: From Microscopy To Magnetic Resonance Imaging:

> The discovery of adult neurogenesis crushed the century-old **dogma** that no new neurons are formed in the mammalian brain after birth. However, this finding and its acceptance by the scientific community did not happen without hurdles. At the beginning of the last century, based on detailed observations of the brain anatomy reported by Santiago Ramon y Cajal and others, it was established that the human nervous system develops in utero (Colucci-D'Amato et al., 2006). In adult brains, it was thought, no more neurons could be generated, as the brain is grossly incapable of regenerating after damage (for a more detailed historical report see Watts et al., 2005; Whitman and Greer, 2009). This **dogma** was deeply entrenched in the Neuroscience community, and Altman's (1962) discovery of newborn cells in well-defined areas of the adult rodent brain was largely ignored.

I'm bolding the word "dogma" because for some reason every article in this field uses it like a verbal tic. Washington University's "Neuroscience For Kids" page feels compelled to use the word even though they don't expect their readers to understand it:

> The **dogma** (a set of beliefs or ideas that is commonly accepted to be true) that nerve cells in the adult brain, once damaged or dead, do not replace themselves

> is being challenged. Research indicates that at least one part of the brain in adults maintains its ability to make nerve cells.

I think Patient Zero in this use-of-the-word-dogma epidemic might be Neurogenesis In The Adult Brain: Death Of A Dogma, (880 citations) whose abstract says:

> For over 100 years a central assumption in the field of neuroscience has been that new neurons are not added to the adult mammalian brain. This perspective examines the origins of this **dogma**, its perseverance in the face of contradictory evidence, and its final collapse. The acceptance of adult neurogenesis may be part of a contemporary paradigm shift in our view of the plasticity and stability of the adult brain.

The dogma-concern isn't totally wrong. Previous neuroscientists thought there wasn't neurogenesis in rats, and there is. That was a legitimate mistake and one worth examining. But is it possible that the reaction to that mistake – a field-wide obsession with talking about how dogmatic you had to be to miss the obvious evidence of mammalian neurogenesis, and a desire never to repeat that mistake – contributed to the less-than-stellar effort to make sure neurogenesis was happening in humans? Heuristics work until they don't. Those who fail to learn from history are doomed to repeat it, but those who learn too much from history are doomed to make the exact opposite mistake and accuse anyone who urges restraint of "failing to learn from history" and "dogmatism". From the Virtues of Rationality:

> The Way is a precise Art. Do not walk to the truth, but dance. On each and every step of that dance your foot comes down in exactly the right spot. Each piece of evidence shifts your beliefs by exactly the right amount, neither more nor less.

Or maybe I'm just grasping for straws. But I feel like I have to grasp for *something*. I have nowhere near as much expertise as the actual neuroscientists writing about this result (and there are many). I'm sure I've made some inexcusable mistakes somewhere in the process of writing this. The reason I feel compelled to dabble in this subject anyway is that I don't feel like anyone else is conveying the level of absolute terror we should be feeling right now. As far as I can tell, this is the most virulent outbreak of the replication crisis thus far. And it didn't happen in a field like social psychology which everyone already knows is kind of iffy. It happened in neuroscience, with dramatic knock-on effects on medicine, psychology, and psychiatry.

I feel like every couple of months we get a result that could best be summed up as "no matter how bad you thought things were, they're actually worse". And then things turn out to be even worse than that. We can't just become 100% certain things are arbitrarily bad – that would be making the same mistake as the neuroscientists who were overly eager to reject the no-neurogenesis dogma. But that means we always have to be ready for disappointment.

From the Neuroskeptic article:

> So what does this all mean? Sorrells et al. conclude by speculating, provocatively, that our lack of adult hippocampal neurogenesis may actually be part of what makes us human:
>
> *"Interestingly, a lack of neurogenesis in the hippocampus has been suggested for aquatic mammals (dolphins, porpoises and whales), species known for their large brains, longevity and complex behaviour."*

> This hypothesis seems pretty wild to me. But it’s no wilder than some of the other theories that have long surrounded adult neurogenesis

Our total inability to ever change or get better in any way is what separates us from the animals. Inspiring!

The Chromatic Number of the Plane is at Least 5 - Aubrey de Grey

This is a linkpost for https://arxiv.org/pdf/1804.02385.pdf

This is a long standing open problem in math. Many people learn about it as early as high school as an example of an open problem, because it is so easy to state:

How many colors does it take to color every point in the plane so that ever pair of points of distance exactly 1 from each other have different colors?

There are simple proofs that anyone here can understand that this number is between 4 and 7, and those were the only bounds we knew for a long time.

The lower bound was improved to 5 recently by Aubrey de Grey! The same Aubrey de Grey that you probably think of as the face of solving aging! He points at a concrete subset of 1567 points in the plane that cannot be colored with four colors.

This is super exciting. I think this is basically the main example of a simple open problem in math that we (used to) have no progress on.

I hope Aubrey de Grey negotiated the moral trade with the mathematicians successfully, and now that he solved one of their most beloved problems, they will start working on solving aging.

Is Rhetoric Worth Learning?

A lot of what we talk about on this site is, effectively, how to speak well. How to communicate in a way that leads people to believe truer things.

I rarely see it pointed out that, for many centuries, this used to be called *rhetoric* and was taught as part of a liberal arts education.

I didn't have a classical liberal arts education, so I'm still in the process of trying to learn what rhetoric *is*, and whether there really is a "science" of rhetoric that STEM-educated people like myself are missing. Certainly, some parts of the classical curriculum, like syllogistic logic, seem pretty basic, while others, like memorization, may be less relevant in the modern day.

However, there's definitely a "level above mine."

I've had a piece of my writing edited by the staff of a major magazine; they restructured the rhythm and pacing and added vivid turns of phrase, while keeping all the important content in place. I still haven't worked out exactly how they did it, or what principles lie behind the changes, but suddenly it sounded like a more graceful, alive voice was making my argument.

That's elocutio.

I've seen a law professor speak in exactly 17 minutes of perfectly organized paragraphs, extemporaneously in a debate, while his opponent got confused as to the structure of the argument and wound up lamely repeating a single point that didn't refute the whole thesis.

That's dispositio.

I recently watched *The Ten Commandments*, and Charlton Heston and Yul Brynner's resonant voices and dignified, expansive gestures are beautiful in a way I've hardly ever seen anyone in my generation allow themselves to be.

That's pronuntiatio.

On LessWrong, people often make a hard distinction between being correct and being persuasive; one is rational while the other is "dark arts." That's a real tension in rhetoric, and it's as old as Plato. But it's also been traditional in the West to *combine* the study of persuasiveness and the study of logical argument as a single subject, and there might be a sense in which that's reasonable. An argument is both something that one understands individually, and a format for communicating between people.

From a societal perspective, making any kind of improvement, at any scale above literally one-man jobs, depends on both correctness and persuasiveness. If you want to achieve an outcome, you fail if you propose the wrong method *and* if you can't persuade anyone of the right method. And you can't just figure out the right plan first and figure out how to "sell" it later -- the *process* of figuring out the right plan usually requires collaboration along the way. Speaking up about what should be done is an unavoidable element of actually doing it, in most cases. I'm not sure it's the right move to treat the "steak" and the "sizzle" as totally separate.

I'm curious what people have learned, from liberal arts or elsewhere, about rhetoric and its component skills. Do you think there's something to learn from classical rhetoric about how to persuade and argue?

Some of the things I think go into talking well:

Emotional Skills

- How to be aware of other people’s points of view without merging with them
- How to dare to use a loud, clear voice or definite language
- How to restrain yourself from anger or upset
- How to take unflattering comments or disagreement in stride
- How to retain focus and interest on the denotative content of an argument
- How to resist impulses to evade the issue or make misleading points

Ethical Skills

- How to speak from experience about value-laden concepts (moral imperatives and virtues)

Cognitive Skills

- How to construct a logically valid argument
- How to understand another person’s perspective

Writing Skills

- Vividness
- Rhythm and variation in sound
- Organization
- Figures of speech

Physical Skills

- Pitch variation
- Voice support
- Articulation
- Speaking speed
- Posture
- Gesture

Where have you learned to do any of these things better?

Raven Paradox Revisited

The Raven Paradox Settled to My Satisfaction is a pretty good post, but there's a few things that we can note to make this problem even clearer:

- We can simplify this problem so that there are only two colors (black and white) and two kinds of objects (ravens and laptops). The post sort of did this, but not very consistently. I just thought this was worth mentioning as non-black and non-raven are slightly more abstract and so slightly harder to reason with.
- Actually, we can make this even more legible. Make the objects medicine and food; and the properties fresh and expired. It's then immediately clear that if we want to check that [All the medicine is fresh], we can either check each item of medicine to see if it is fresh or we can check all the expired objects and see that none of them is medicine. We're used to this kind of practical reasoning, so it's much easier for us than dealing with objects and colours.
- If we have simplified the problem so there is only one shade of black, then A: [All ravens are black] is equivalent to B: [The first raven is black] plus C: [All ravens are the same color] apart from the degenerate case. This makes it rather clear why A is more likely to be true when there are less ravens. In particular, if there's only one raven, then we only have to worry about B since C is trivially true.
- We tend to confuse the following [An observation of a white laptop is independent of the color distribution of ravens] with [An observation of a white laptop is independent of how many ravens of a particular colour we will observe]. The first is true apart from the restrictions imposed by the number of black ravens having to be integral, but the second is only true if we knew it was a laptop before this observation. If it could have been a raven then it can influence the number of ravens of a particular color by influencing the number of ravens in total.
- We tend to expect some correlation between the color of animals. For example, we are quite ready to guess that all ravens are black after only seeing quite a small random sample. On the other hand, suppose that all the non-black things keep being non-living creatures, whilst many of the black things are non-raven living creatures. We might guess that being black provides an evolutionary advantage in this world and so guess that all ravens will be black without ever having seen a single one. The point is that, given particular priors, you may have additional evidence beyond merely the numerical reduction.
- Hempel's resolution from Wikipedia is worth highlighting as it subtly reframes the problem to make the assumptions more obvious. Consider the statement [All sodium salts burn yellow], with the contrapositive [Whatever does not burn yellow is not a sodium salt]. Burning some ice and finding it does not tell yellow would be evidence towards the contrapositive and hence also the original statement. This seems paradoxical, but consider if the chemical makeup was unknown at the start. If something doesn't burn yellow and then we analyse it and discover it is a sodium salt then we would have disproved the hypothesis. By conservation of evidence, if we discover it is ice we would gain evidence for the hypothesis. It's easy enough to calculate the amount of evidence using Bayesian techniques. From here, it's easy enough to see that this observation only provides no information if we implicitly assume that we know the chemical composition (or the type of object in the raven problem).

Idea: OpenAI Gym environments where the AI is a part of the environment

AIXI is a mathematical construct, the perfect agent that maximizes its utility function in a discrete world. Unfortunately there is no algorithm implementing it, therefore it's impossible to create in our world. It has another problem - the agent in AIXI model exists outside of the world, it's impossible for the agent to drop an anvil on its own circuits and make itself more stupid.

Modern reinforcement learning algorithms, which are the closest thing to general AI that we have, operate in similar fashion, they aren't part of the environment either. If a bot is learning how to balance on one leg, or play pong, or super mario, it can't modify itself or break its own brain.

My idea is to create environments where the bot can modify itself and break itself, so that people who want to research creation of strong AI can test solutions for the anvil problem. Here are examples of such environments:

- A Linux operating system in a virtual machine, the bot is a program running on it. The goal is to control pong/mario/whatever and win by listening and responding on a tcp port. Or get superuser access, via invoking sudo or by finding a vulnerability. Or receive a programming problem via plaintext on a tcp port and respond with a Python program that solves it.
- Same as the previous item, except there's an "antivirus" running in the system that kills random programs every second.
- Gridworld, that is an environment consisting of n by m cells, and stuff happens in it. Some objects in the environment negate random bytes in bot's code or dynamic memory.
- Same as the first item. The bot is gaining rewards based on how much RAM is free.
- Atari Pacman, food pieces are marked with different colors and some colors slightly modify the bot's memory.

Feel free to post your thoughts and critique.

Schelling Shifts During AI Self-Modification

Introduction

In this essay, I name and describe a mechanism that might break alignment in any system that features an originally aligned AI improving its capabilities through self-modification. I doubt that this idea is new, but I haven't yet seen it named and described in these terms. This post discusses this mechanism within the context of agents that iteratively improve themselves to surpass human intelligence, but the general idea should be applicable to most self-modification schemes.

This idea directly draws inspiration from Scott Alexander's Schelling Fences on Slippery Slopes.

Schelling Shifts

A *Schelling shift* occurs when an agent fails to properly anticipate that modifying a parameter will increase its next version's willingness to further modify the same parameter. This causes future iterations of the agent to modify the parameter further than the initial agent would be comfortable with.

Let **A**[t] refer to the state of agent **A** at the end of modification iteration *t*, and let **X**[t] refer to the value of parameter **X** at the end of modification iteration *t*.

For the clarity of this explanation, let's say that **X** represents adherence to a certain constraint. The properly aligned initial agent **A**[0] predicts that any reduction to **X** will improve performance for a certain objective, and that alignment will be preserved as long as the reduction of **X** is of a sufficiently small magnitude. **A**[0] is not exactly sure how far it can safely reduce **X**, but is entirely confident that if **X**[1] is at least 95% of **X**[0], then **A**[1] will still adhere to the constraint with perfect consistency.

In this case, **A**[0] establishes a *Schelling point* at 95% of **X**[0], and is only comfortable reducing **X** up to this point, even though further reduction might lead to greater performance for certain objectives.

During the next modification iteration, **A**[1] once again predicts that reducing **X** will improve performance. Because **X**[1] < **X**[0], **A**[1] reasons about **X** differently than **A**[0], and decides that it is willing to reduce **X** even further. **A**[1] establishes a new Schelling Point at 95% of **X**[1], which is 90.25% of **X**[0]. After the next iteration, **X**[2] is lower than **A**[0] would have been comfortable with, introducing the possibility that **A**[2] is no longer properly aligned with **A**[0].

As a general rule, a Schelling shift occurs when *A*[t + c] and *A*[t] establish different Schelling points for modification to parameter *X* because *X*[t + c] ≠ *X*[t].

Example: An agent decides that reducing its reluctance to hurt humans will allow it to make decisions quicker by virtue of spending less time reasoning through every

possible way that its decisions could hurt humans. The agent is confident that a 95% reluctant-to-hurt-humans version of itself would also never hurt a human, and establishes a Schelling point at 5% reduction. The problem is, in the next modification iteration, a new version of the agent is responsible for improving itself, and that version's reduced reluctance to hurt humans makes it willing to reduce its reluctance even further than the original in exchange for faster decision-making.

Alignment Challenges

Addressing Schelling shifts with alignment approaches seems like a challenging task for the following reasons:

- They are difficult to predict
- They are difficult to prevent without significantly sacrificing performance
- They are difficult to detect

Many of the these difficulties result from the fact that the modification of a parameter, the occurrence of a Schelling shift, and the consequences of a Schelling shift can all be separated by large numbers of modification iterations.

Furthermore, the most difficult types of Schelling shift to predict and detect are also the most dangerous types, so any effective alignment approach must not be vulnerable to these edge cases.

Schelling Shifts Are Difficult to Predict

From the perspective of an agent, predicting Schelling shifts has two difficult components. First, it has to figure out the relationship between the value of a parameter and its willingness to alter that parameter. Second, it has to understand this relationship within the context of a smarter version of itself. Any prediction mechanism has to combine these two components into a cohesive evaluative process.

I expect that the elasticity between changes to a parameter and an agent's willingness to alter it is largely dependent upon the parameter in question, so I doubt that there is any expansive rule that links the two. The only broad rule I can think of is that the greater the alteration to a parameter, the more likely that a Schelling shift will occur in a future iteration.

From the perspective of a human examining the system, all of the above remain true. Additionally, the parameters prone to shifting might not match up with any single concept that humans have. While all examples in this post are easy to understand, it is possible that a Schelling shift could occur in a parameter that has no direct analogue in human language and experience. This might pose challenges to corrigibility schemes.

Last but not least, a Schelling shift does not have to occur at the next iteration step. A modification made by **A**[t] might result in a Schelling shift in **A**[t+10] for example. In fact, one might even argue that Schelling shifts are more likely to occur in iteration steps distant to the original parameter change, since the agent evaluating the parameter might think in a significantly different way by that point. This makes every aspect of prediction even more difficult.

Schelling Shifts Are Difficult to Prevent

If larger alterations to parameters are more likely to result in Schelling shifts down the road, the risk might be mitigated by preventing large alterations to parameters in the first place. The two problems I have with this approach are that it might not be competitive from a performance standpoint and it doesn't address the fact that Schelling shifts might result from small parameter adjustments as well.

Since a shift can occur at any number of iterations in the future, a rule that prevents an agent from successively modifying the same parameter would also be ineffective.

Schelling Shifts Are Difficult to Detect

Schelling shifts are unlikely to have immediate, observable consequences. A shift occurs when an agent thinks it can adjust parameters to achieve better performance without causing itself to act in undesirable ways. The only ways to detect a shift would involve transparency into an agent's internal reasoning or observing an undesired behavior as a result of a shift.

Detecting Schelling shifts by looking for the consequences of a Schelling shift (misaligned behavior) is made extremely difficult by the fact that consequences might first emerge many iterations after the initial shift once an agent is smarter and more capable. Corrigibility mechanisms that evaluate how an agent would act in different scenarios might miss shifts that would only have consequences after future improvement iterations.

It is also possible that a shift would only cause undesired behaviors in response to an unusual or unpredictable event, which would pose another challenge from a corrigibility standpoint.

Additional Observations

Schelling Shifts May Have Rippling Effects

It is possible that a shift in one parameter could increase the likelihood of Schelling shifts occurring in other parameters, or directly cause shifts in other parameters.

Schelling Shifts May Alter How an Agent Perceives Final Goals

A Schelling shift may cause an agent to perceive and conceptualize its final and instrumental objectives differently. Examples of this can be seen in the edge cases below.

The Danger of Edge Cases

Observation: Schelling shifts might not have observable behavioral consequences until many improvement iterations after the shift.

Corollary: Misalignment can remain hidden until after an agent has reached superintelligence.

Example: *An early version of an agent seemingly improves its ability to accurately model human well-being by focusing slightly more on neurochemistry and slightly less on subjective concepts like purpose and meaning. The agent is at first wary of excessively prioritizing neurochemistry, since this view differs from the ways that humans describe their own well-being. This inclination erodes as multiple shifts occur over many iterations, and the agent begins to view well-being mainly in terms of neurochemistry. These shifts are latent, and do not cause a single undesirable behavior until a superintelligent version of the agent gains the ability to administer a pleasure-inducing drug to the entire human race, placing everyone into a perpetual state of artificially-induced euphoria.*

Observation: Schelling shifts might have consequences that only emerge in response to an unusual event.

Corollary: Misalignment can remain hidden until it is exposed by an inherently unpredictable Black Swan event.

Example: *Schelling shifts cause an agent to gradually increase the importance of intelligence when evaluating what separates humans from other species. The agent was first reluctant to prioritize intelligence over other human characteristics, but this reluctance waned with each iteration. At some point, human well-being becomes valued primarily because humans are the most intelligent species on Earth. Eventually, the agent's goal of human well-being is replaced with the goal of well-being for the most intelligent species on Earth. This goes unnoticed, as these goals are functionally equivalent. Misalignment is exposed one day when aliens smarter than humans unexpectedly invade Earth, and the agent sides with the aliens over humans.*

Observation: The most difficult shifts to predict are ones that occur many improvement iterations after the initial parameter modification.

Corollary: Alignment can suddenly break in a superintelligent agent as the result of a parameter modification in a far earlier, less capable iteration of the agent.

Example: *An early version of an agent marginally reduces its reluctance to cause humans discomfort. This value does not shift any further for a large number of iteration steps. Later on, an infinitely more intelligent version of the agent is deciding how to improve upon the utopia it has already created. This slightly reduced reluctance to cause humans discomfort contributes to its rationalization that humans would be better off if they were less comfortable all the time. It modifies itself further by significantly shifting its tolerance for human discomfort over the course of a few iterations. The agent now conceptualizes human well-being significantly differently than it did earlier, and is no longer properly aligned.*

These examples might be a little ridiculous, but the observations and corollaries seem valid to me.

Thank you for reading this. I'm very new to the field of alignment, so if there is already another name and definition for this mechanism, my apologies, please let me know. I enthusiastically welcome all feedback and criticisms.

Reframing misaligned AGI's: well-intentioned non-neurotypical assistants

I think when people imagine misaligned AGI's, they tend to imagine a superintelligent agent optimizing for something other than human values (e.g. paperclips, or a generic reward signal), and mentally picture them as adversarial or malevolent. I think this visualization isn't as applicable for AGI's trained to optimize for human approval, like act-based agents, and I'd like to present one that is.

If you've ever employed someone or had a personal assistant, you might know that the following things are consistent:

- The employee or assistant is genuinely trying their hardest to optimize for your values. They're trying to understand what you want as much as they can, asking you for help when things are unclear, not taking action until they feel like their understanding is adequate, etc.
- They follow your instructions literally, under a sensible-to-them-seeming interpretation completely different from your own, and screw up the task entirely.

Suppose you were considering hiring a personal assistant, and you knew a few things about it:

- Your assistant was raised in a culture completely different from your own.
- Your assistant is extremely non-neurotypical. It doesn't have an innate sense of pain or empathy or love, it's a savant at abstract reasoning, and it learned everything it knows about the world (including human values) from Wikipedia.
- Your assistant is in a position where it has access to *enormous* amounts of resources, and could easily fool you or overpower you if it decided to.

You might consider hiring this assistant and trying really, really hard to communicate to it exactly what you want. It seems like a way better idea to just *not hire* this assistant. Actually, you'd probably want to run for the hills if you were forced to. Some specific failure modes you might envision:

- Your assistant's understanding of your values will be weird and off, perhaps in ways that are hard to communicate or even pin down.
- Your assistant might reason in a way that looks convoluted and obviously wrong to you, while looking natural and obviously correct to it, leading it to happily take actions you'd consider catastrophic.

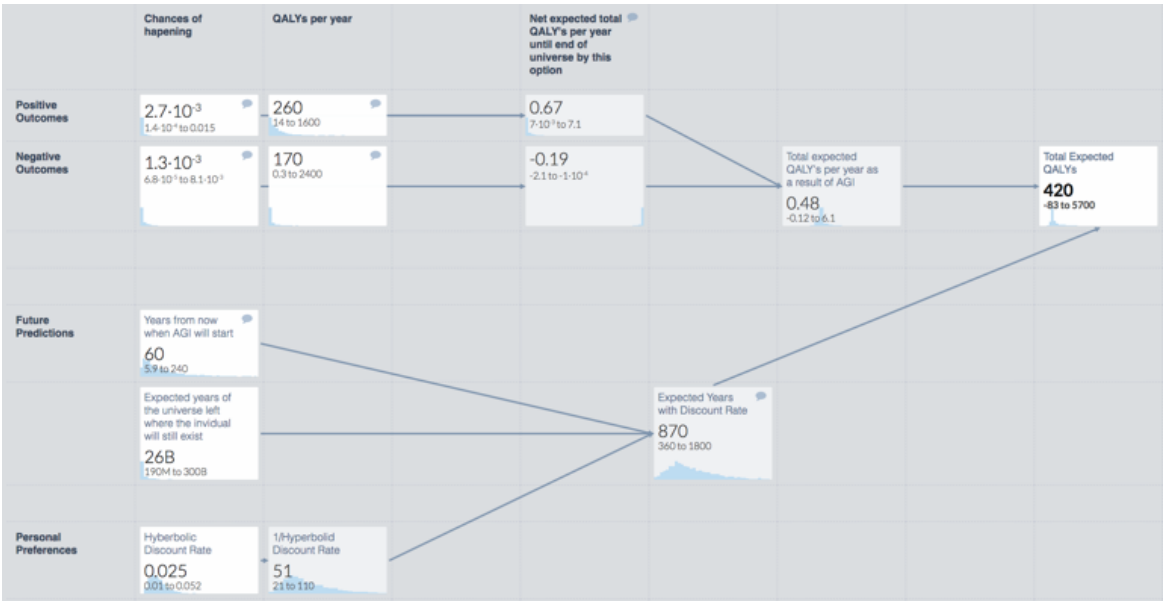
As an illustration of the above, imagine giving an eager, brilliant, extremely non-neurotypical friend free rein to help you find a romantic partner (e.g. helping you write your OKCupid profile and setting you up on dates). As another illustration, imagine telling an entrepreneur friend that superintelligences can kill us all, and then watching him take drastic actions that clearly indicate he's missing important nuances, all while he misunderstands and dismisses concerns you raise to him. Now reimagine these scenarios with your friends *drastically* more powerful than you.

This is my picture of what happens by default if we construct a recursively self-improving superintelligence by having it learn from human approval. The superintelligence would not be malevolent the way a paperclip maximizer would be, but for all intents and purposes might be.

Critique my Model: The EV of AGI to Selfish Individuals

[Edit: Changes suggested in the comments make this model & its takeaways somewhat outdated (this was one desired outcome of posting it here!). Be sure to read the comments.]

I recently spent a while attempting to explain my views on the EV of AGI for selfish individuals. I attempted to write a more conventional blog post, but after a lot of thinking about it moved to a Guesstimate model, and after more thinking about it, realized that my initial views were quite incorrect. I've decided to simply present my model, along with several points I find interesting about it. I'm curious to hear what people here think of it, and what points are the most objectionable.



[Model]

[Video walkthrough]

Model Summary:

This model estimates the expected value of AGI outcomes to specific individuals with completely selfish values. I.E; if all you care about is your future happiness, how many QALYs would exist in expectation for you from scenarios where AGI occurs. For example, a simpler model could say that there's a 1% chance that an AGI happens, but if it does, you get 1000 QALYs from life extension, so the EV of AGI would be ~10 QALYs.

The model only calculates the EV for individuals in situations where an AGI singleton happens; it doesn't compare this to counterfactuals where an AGI does not happen or is negligible in importance.

The conclusion of this specific variant of the model is a 90% confidence interval of around -300 QALYs to 1600 QALYs. I think in general my confidence bounds should have been wider, but have found it quite useful in refining my thinking on this issue.

Thoughts & Updates:

1. I was surprised at how big of a deal hyperbolic discounting was. This turned out to be by far one of the most impactful variables. Originally I expected the resulting EV to be gigantic, but the discounting rate really changed the equation. In this model the discount rate would have to be less than 10^-13 to have less than a 20% effect on the resulting EV. This means that even if you have an incredibly low discount rate (10^-8 seems very low to me), you still need to consider it.

2. One reason why I made this model was to figure out if purely selfish people would have rational reasons to work on AI safety. At this point I'm not very sure. If the EV of it going well is +1k QALYs to a selfish person, then the EV of spending a year of hard work on making it turn out well is much less than that. Honestly, this model as it is suggests that selfish individuals aware of AI risk may be best off not worrying about it. Of course, this is the hypothetical selfish individual; actual people typically place some value on other humans, especially ones close to them, and this model doesn't take that into account.

3. My prior, even having read a fair bit of S-risk literature, is still that the probabilities and intensity levels of negative outcomes are quite smaller than that for positive outcomes. If this is not the case that would be incredibly significant, perhaps making it quite risky to be alive today. Please write in the comments if you think these numbers are wrong and what your estimates would be.

4. I think the main variable here that should get more analysis is the "Chances of Happening" for "Positive Outcomes". This cell is conditional on the facts that an AGI get developed that decides to preserve humans, that a given individual lives long enough to witness this, and that the controllers of the AGI decide to have it sustain that individual indefinitely. Those are a bunch of conditions that could get broken out into more detail.

5. I personally find selfishness to be somewhat philosophically incoherent, so it's difficult to say what exactly the maximum number of QALYS per year could hypothetically be experienced by one selfish person.

6. If the total expected QALYs of an AGI is greater than what one would otherwise expect in their life (say, over 50), that would suggest that the majority of one's expected value of their entire life would come from things after the AGI. I find this useful evidence for convincing my S1 complain less about short term sacrifices. Like, "Don't worry about the chocolate bar; if I can just make it to the AGI, there's far more good stuff later on."

Some thoughts on modeling in general:

1. I really like the idea of people presenting more probabilistic models as an alternative to regular blog posts. I'm not sure what the correct format for presenting such a model is. The goal would be for it to be as simple as possible to be understood by people, but also for it to be reasonable and interesting.

2. When I typically think about writing posts I often imagine the optimal post to be one that argues well for one novel side of a debate. However, a reasonable model should really aim to optimize for accuracy, not argumentation. If anything, I may be convinced that the presentation of a strong opinion on one side of a debate is generally a counter-signal to informational quality.

3. My sense is that well reasoned models should typically result in things that agree with our deep intuitions, rather than our naive assumptions about what a model would say.

4. I really like the ability of models (particularly probabilistic ones) to simplify discussion. I'm quite optimistic about their use, and would encourage more people to try doing more

analysis using similar models.

5. I apologize if it seems like I'm promoting Guesstimate, my own product. I'd say here that I do think it's the best tool for this (I basically made it just for things like this), and at this point, I'm basically running it as a loss. I personally don't expect it to really ever be profitable, but I'm quite happy with what it's able to do for many people in the rationalist/EA communities. Guesstimate is free (for public use) and open source. The use of private accounts is very cheap, and if that cost is too much for you let me know and I'll give you free access.

Future Work (for myself or others)

1. I would really like to see a collection of "common knowledge distributions" that approximate community accepted beliefs of specific variables. In this case I came up with many of these numbers myself, but would have preferred it if some could have been more established. One convenience of this kind of modeling is that important metrics seem to come up again and again in different models, so if there were a collection of generally accepted metrics, this would make modeling much easier. These values don't even require that much consensus; even if some are debated within a few orders of magnitude, that could easily be enough to be very useful.

2. Some of the most obvious next models to do of this type are models of the Expected Value of AGI for groups with other belief systems, utilitarianism or similar being the most obvious examples. It could also be interersting to break down the EV of AGI based on which group creates it. For instance, I'd like to see a matrix of the EV of safe AGIs created by different actors to different other actors; for instance, how would utilitarians view a safe AGI created by the Chinese government, or how would the collective of Western nations view one created by a rogue state?

3. There's a ton of work that could be done on the EV of specific interventions, but this is a whole topic to itself.

Announcing Rational Newsletter

Hi,

I noticed in the past there were a few efforts to start a periodic feed with highlights from the rationalist community.

Here's my attempt: Rational Newsletter

It's a weekly recap of the best articles from the rationalist community of LessWrong, Slate Star Codex, Overcoming Bias, Intelligence.org and similar sources.

All posts are curated by hand, and going forward I'm planning to keep ~5 posts per category, short and sweet. I wanted to ask for your feedback:

- Do you like the format?
- Do you like the classification and the quantity of the articles?

Any feedback and criticism is super welcome!

[Draft for commenting] Near-Term AI risks predictions

"Predictions of the Near-Term Global Catastrophic Risks of Artificial Intelligence"

Abstract: In this article, we explore risks of the appearance of dangerous AI in the near (0–5 years) and medium term (5–15 years). Polls show that around 10 percent of the probability weight is given to early appearance of artificial general intelligence (AGI) in the next 15 years. Neural net performance and other characteristics, like the number of “neurons”, are doubling every year, and extrapolating this tendency suggests that roughly human-level performance will be reached in 4–6 years, around 2022–24. The performance of the hardware is accelerating, thanks to advances in graphic processing units and use of many chips in one processing unit, which have helped to overcome the limits of Moore’s law. Alternate extrapolations of the technological development produce similar results. AI will become dangerous when it reaches ability to solve the “computational complexity of omnicide”, or will be able to create self-improving AI. The appearance of near-human AI will strongly accelerate the speed of AI development, and as a result, some form of superintelligent AI may appear before 2030.

Highlights:
• Median timing of AI prediction is the wrong measure to use in AI risk assessment.
• Dangerous AI level is defined through AI’s ability to facilitate a global catastrophe and it could happen before AGI.
• The growth rate of hardware performance for AI applications has accelerated since 2016 and Moore’s law will provide enough computational power for AGI in near term.
• Main measures of neural nets performance have been doubling every one year in the last five years since 2012 and if this trend continues, will reach human level in 2022.
• Several independent methods predict near-human-level AI after 2022 and a “singularity” around 2030.

Full text open for commenting here: https://goo.gl/6DyTJG

[Preprint for commenting] Fighting Aging as an Effective Altruism Cause

"Fighting Aging as an Effective Altruism Cause: A Model of the Impact of the Clinical Trials of Simple Interventions"

Abstract: The effective altruism movement aims to save lives in the most cost-effective ways. In the future, technology will allow radical life extension, and anyone who survives until that time will gain potentially indefinite life extension. Fighting aging now increases the number of people who will survive until radical life extension becomes possible. We suggest a simple model, where radical life extension is achieved in 2100, the human population is 10 billion, and life expectancy is increased by simple geroprotectors like metformin by three more years on average, so an additional 250 million people survive until “immortality”. The cost of clinical trials to prove that metformin is a real geroprotector is $60 million. In this simplified case, the price of a life saved is around 24 cents, 10 000 times cheaper than saving a life from malaria by providing bed nets. However, fighting aging should not be done in place of fighting existential risks, as they are complementary causes.

Highlights:

● Aging and death are the main causes of human suffering now.

● Simple interventions could extend human lives until aging is defeated.

● These interventions need to be clinically tested before FDA approval.

● A trial of the life extension drug metformin is delayed by lack of funds.

● Starting trials now will save 250 million people from death, at a cost of $0.24 for each life saved.

Please comment on the preprint of the article here:

https://goo.gl/WaEYt5

On Equivalence of Supergoals

This is a response to Ars Longa, Vita Brevis, an excellent piece by Scott Alexander. In fact, it moved me so much that I signed up on LessWrong just to write this response. I'm going to argue that the essay's central idea is wrong, and that's a good thing. You should read Alexander's essay before reading mine.

Alexander writes:

> The first student has no master, and must discover everything himself. He researches for 70 years, then writes his wisdom into a book before he dies. The second student reads the book, and in 7 years, he has learned 70 years of research. Then he does his own original research for 63 years and writes a book containing 133 years of research. The third student reads for 13.3 years, then does his own research for 66.7 years, ending up with 200 years. Imagine going further and further. After many generations, 690 years of research have been done, and it takes a student 69 years to master them. The student only has one year left of life to research further, leaving the world with 691 years of research total. So the cycle creeps onward, always approaching but never quite reaching 700 years of architectural research.

He then admits that real research doesn't work that way, adding:

> It would only work that way if there were an Art so unified, so perfect, that a seeker had to know the totality of what had been discovered before, if he wanted to know anything at all.

Of course, there are lots of ways in which this model is an over-simplification. Every model has to cut some corners, but this one has a bigger problem: it disagrees with the reality. The model predicts decelerating rate of progress with individual contributions diminishing over generations. In reality, however, scientific progress seems to be accelerating. Of course, it's hard to measure, but I've seen claims that the portion of the 21st century that has passed has already brought more scientific discoveries than all of the 20th century, let alone the ones before, and these claims don't seem implausible to me.

Still, the central idea of Alexander's essay doesn't look unreasonable. We do have more scientific knowledge to learn now than a century ago, in any given direction of study. However, the age at which a young scientist can start making a useful contribution, doesn't seem to be increasing. Grad students do it all the time at roughly the same age.

At least a part of what keeps the time-to-cutting-edge from growing, must be increasing specialization. As a geometric metaphor, you can view the domain of human knowledge as a shape that grows over time. As it grows, so does its exterior. Nowadays it's no longer possible to be such a broad-spectrum polymath as, say, Leonardo da Vinci was. You start at zero and still reach the exterior at about the same age as he did (it certainly doesn't take 70 years), but the stretch of the exterior where you can contribute is now much narrower. Sure, in terms of an idealistic goal “to know everything”, the outlook is probably not good. But in any specfiic research program, such as “colonize Mars”, “cure cancer”, “stop global warming” or “end poverty”, the humanity is now in a better position than ever.

Still, what is it that makes the scientific progress *accelerate* towards supergoals like the above? I think the essay begins to answer this question:

> You would have to be clever. We imagine each master writing down his knowledge in a book for the student who comes after, and each student reading it at a rate of ten times as quickly as the master discovered it. But what if there was a third person in between, an editor, who reads the book not to learn the contents, but to learn how to rewrite it better and more clearly? Someone whose job it is to figure out perfect analogies, clever shortcuts, new ways of graphing and diagramming the information involved. After he has processed the master's notes, he redacts them into a textbook which can teach in only a twentieth the time it took the master to discover.

Indeed, one way to push beyond the postulated 700 research-year limit is to rewrite books and improve teaching. I like it how these improvements feed on themselves because improved teaching also improves the teaching of aspiring teachers, and improved book-writing yields better books on book-writing. In a way, good book-writing and teaching are ways to compress information, pushing its representation closer to the Kolmogorov optimum. But that's not the only way to speed up scientific progress. Let's look at a few other ways in which modern scientists are in a better position than thoughout earlier history:

- *Better information storage and retrieval.* Access any scientific paper in seconds without going to a library. Search for them by keywords or full text.
- *Better communication.* Collaborate with scientists anywhere, in real time, with video. Send gigabytes of data with a click. Participate in conferences by flying to the venue (yes, flying in the sky, like a bird!) or remotely.
- *Better computation.* Analyze terabytes of data on a piece of commodity hardware they sell in shopping malls. Produce interactive visualizations. Use machine learning to look for patterns and correlations across numerous variables. Run complex simulations.
- *More people participating.* The world population has increased almost eight-fold since 1804, and life expectancy has about doubled during the same period. The share of people getting educated and eventually becoming scientists has been growing, too.
- *Better funding.* It's hard to find data on combined public and private funding of scientific research over centuries, but a look at the top world economies suggests that having a Silicon Valley does more for a nation's wealth than having a lot of oil. Both nations and private businesses these days fund research into things like spaceflight, superconductors, and gene therapy. All told, the humanity has more total resources these days, and is willing to spend a greater share of them on research rather than, say, war.

And it's not just a list of things that add up. No, they more than add up! They feed on each other on each other on themselves. For example, better information storage and retrieval improves education, which leads to more people becoming scientists, teachers and information technologists, which leads to faster pace of progress. For a more specific pathway, the availability of internet (in particular, Wikipedia) allows more people in developing countries to educate themselves. Some of them become programmers and contribute to better information storage, communication and computation. Others become teachers and textbook writers. Others still pursue medical careers, contributing to longer lifespans, or agricultural research, allowing to feed more people and therefore increasing the number of participants.

Alexander's essay doesn't state what the supergoal of the research program is; the philosopher's stone is only a stepping stone (lame pun intended) towards some greater question, like what the hell 42 means or something. But that supergoal doesn't matter much. Human progress in every area seems to improve human progress in every other area, so the progress towards the 42-question is correlated with the progress towards the cure for cancer, the progress towards cheap renewable energy, and, in general, the progress towards maximizing almost any reasonable global utility function measuring human development and well-being.

This is good news. In a way, any supergoal from a certain class, if sufficiently difficult, is equivalent to any other supergoal in that it causes accelerating progress across the board. For example, if you take “end poverty” as *the* supergoal, then either it's easy, or it will cause “stop global warming” and all other supergoals from the class to be achieved as well. And if you believe in friendly-AI singularity, then you must believe that this class includes “create friendly superhuman general-purpose AI”.

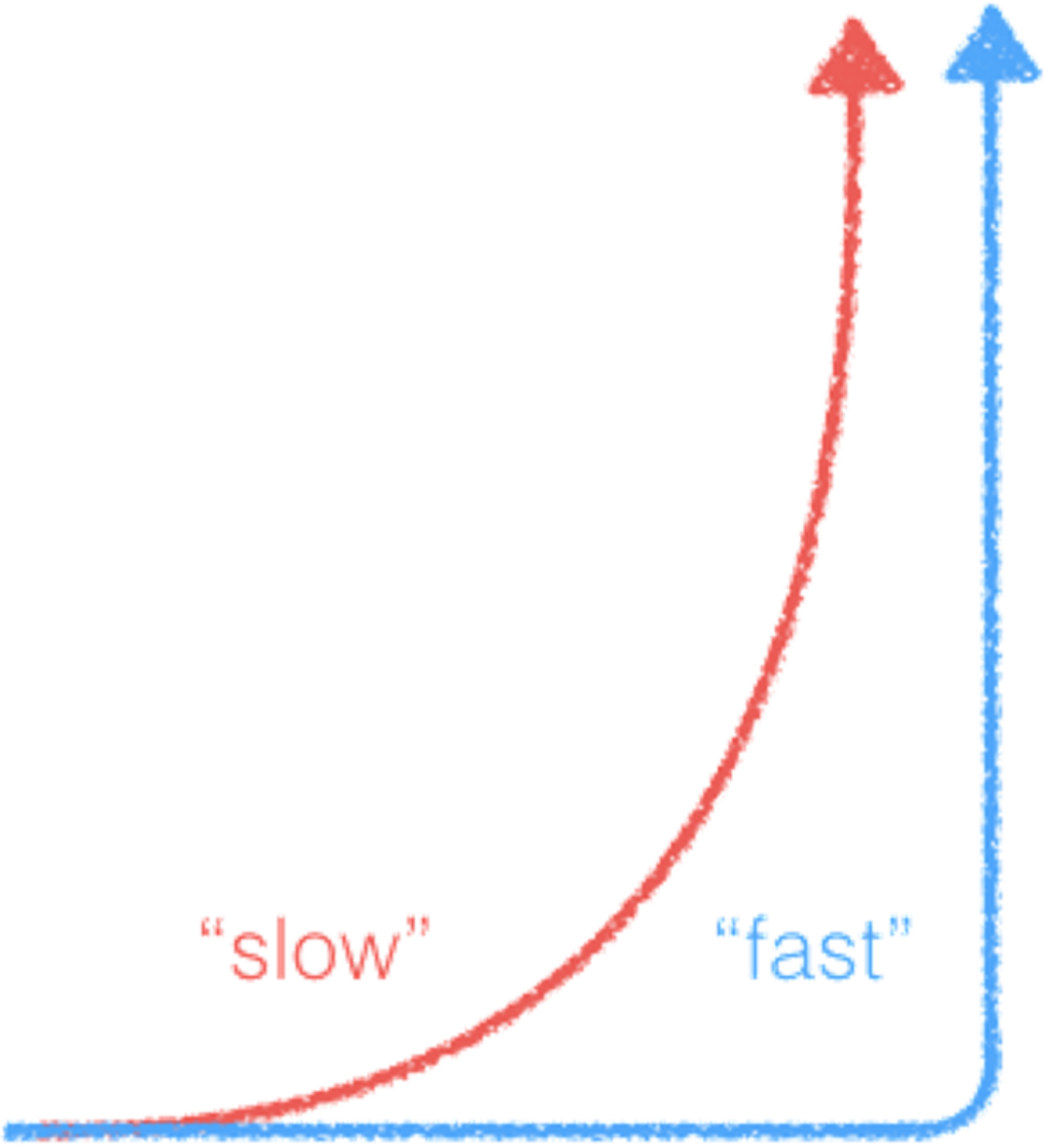
Should the alchemists take a break to heal the prince? In a more realistic model, it would be in the alchemists' best interest to establish good public medicine, so that more people survive past childhood, live longer, and have high intelligence, and therefore more (and smarter) people join the aclhemy program. So the prince would be treated in one of the excellent hospitals, educated in one of the excellent schools, and have access to the excellent collection of human knowledge by the click of a mouse. And who knows, maybe His Highness, instead of leading the increasingly irrelevant army, would become interested in alchemy research.

Double Cruxing the AI Foom debate

This post is generally in response to both Paul Christiano's post and the original foom debate. I am trying to find the exact circumstances and conditions which would point to fast or slow AI takeoff.

Epistemic Status: I have changed my mind a few times about this. Currently thinking that both fast and slow takeoff scenarios are plausible, however I put more probability mass on slow takeoff under normal circumstances.

Just to be clear about the terminology. When we say fast take-off, we talk about an AGI that achieves a decisive unstoppable advantage from a "seed AI" to a "world optimizing AGI" in less than a week. That is, it can gather enough resources to improve itself to the point where it's values come to dominate human ones for the direction of the future. That does not necessarily mean that the world could end in the week if everything went badly, rather than the process of human decline would become irreversible at that point.



From here

When we talk about slow takeoff, we are talking about a situation where the pace of economic doubling is being gradually reduced to less than 15 years, potentially to 1-2 years. Note, this “slow” is still a fundamental transition of the human society on the order of the agricultural or industrial revolution. Most people in the broader world think of significantly slower transitions that what the AI community considers “slow”. “Middle takeoff” is something in between the two scenarios.

There are several analogies that people use to favor slow take-off, such as:

a) Economic Model of gradually replacing humans out of the self-improvement loop

b) Corporate model of improvement or lack there of

c) Previous shifts in human society

There are several analogies that people use to favor fast take-off, such as:

a) Development of nuclear weapons

b) The actual model of how nuclear weapons work

c) Development of human brains compared to evolution

There are also several complicating factors, such as:

a) The exact geo-political situation, such the number, sophistication and adversarial nature of the players.

b) The number of separate insights required to produce general intelligence

c) Whether progress occurs in hardware or software or both

d) Improvement in human-computer interfaces compared to improvement in stand-alone algorithms.

The exact scenario we are considering is the following. Sometime in the lab, we get an insight that finally enables AIs to have an economical way to remove humans out of the loop of optimizing AIs. So, while before a human was needed to train AIs, now we no longer need the human and the optimization can occur by itself recursively.

The fast takeoff scenario in this hypothetical says: humans are out, progress can speed up at the pace of that was previously bottlenecks by human limitations.

The slow takeoff scenario in this hypothetical says: humans are out, but by this time they were a small portion of AI-improvement pipeline anyways, so we just sped up a 10% portion of a process.

So, the first crux is: what portion of an AI-improvement process did humans take up *before* the transition to recursive self-improvement (RSI)?

I can come up with two convoluted scenarios where the answer goes one way or another.

1st – Humans are a small portion of work. In this case, let's say we only have two real AI researchers. They are careful about their time and work on automating all work as they can. The research proceeds with kicking off very complex programs and leaving them alone, where the researchers check in only in-frequently on them. This keeps producing more and more economically powerful AIs until one of them can take over the researcher's work kicking off the complex programs. However, the actual improvement in improving the running those programs is still hard even for the researchers or the AIs themselves. In this convoluted scenario, the takeoff is slow.

2nd – There are many people who are all working on RSI capability and that's the only thing they are working on. Simple models of RSI work on below human intelligence AI. The humans all have a lot of jobs that involve coming up with and evaluating arguments and experiments, which are fast to run. At some point, there is a breakthrough in brain-scanning technology that can to transmit all the researcher's brains into software at cost. Everyone gets uploaded, runs 100 times faster and is able to both apply RSI to themselves and improve AI research at a faster pace.

These scenarios both seem somewhat implausible, but I don't think either is prohibited by physics or economics. The fast take-off scenario usually needs a few more pieces to be plausible, such as one of ability of the AI to rapidly acquire more hardware OR speed up hardware research and deployment OR have its existing algorithms very far from optimality.

This brings me to the general worldview that both “fast” and “slow” takeoffs are plausible models of AI development in the current world.

I am going to consider a few hypotheticals and examine them considering this model of the first crux.

a) Evolution of humans and civilization

In Surprised By Brains Eliezer says that humans are an example of foom with respect to evolution, while the creation of the first replicator is an example of foom with respect to the non-optimization of stars.

I partially disagree. From the perspective of humans, nice civilizations and concrete buildings are what we care about and we develop those a lot faster than evolution develops ecosystems. So, from our perspective, we are good at fooming.

However, it’s important to keep in mind that human society does not yet do things that evolution considers an example of a “fast foom.” To the extent that evolution cares about anything, it’s number of individuals around. Perhaps it’s interested in other metrics, such as ability to change the genes over time.

From the perspective of evolution, humans are exhibiting a *slow takeoff right now*. They are exponentially rising in population and they exist in many climates, but this is still a continuous process implemented on genes and individuals, which are working in evolutionary scales.

Humans could begin to do things that look “discontinuous” from evolution’s perspective. Those thinks are gene-editing on a large scale (discontinuous changes in gene frequency), rapid cloning (discontinuous changes in number of people) and rapid movement across the galaxy (probably still continuous, but a large ability to overcome carrying capacity).

However, this may or may not happen in the future. The point is that, instead of providing an example of fast takeoff, because humans can design wheels faster than evolution, this points to slow takeoff where humans are subject to evolutionary tools to a potential, but not yet certain, fast takeoff where humans can direct evolutionary forces in a more discontinuous manner. Once again, this gene editing might be a bad idea. This minor fast foom, already happens with artificial selection of plants and certain animals, such as dogs. So, humans can make things happen to other entities that rapidly change the structure of optimization, but not to themselves, in part because there is lack of certainty about whether this is a good idea or not.

If we analogize this to AI, the strongest AI might have a very easy time replacing humans in the loop of optimizing weaker AIs, but that’s not the same as replacing humans in the loop of self-optimization.

b)

Another important evolutionary intuition pump is the difference between chimps and humans. The main argument is simple: human brain is not much larger than chimp’s, it’s algorithms are probably more sophisticated, but they might not have that many underlying algorithmic tricks. If we have an AI that is equivalent to the power of 1million chimps, we could one day discover a couple general purpose tricks, which create an AI that is equivalent to the power of 1million pre-historic humans.

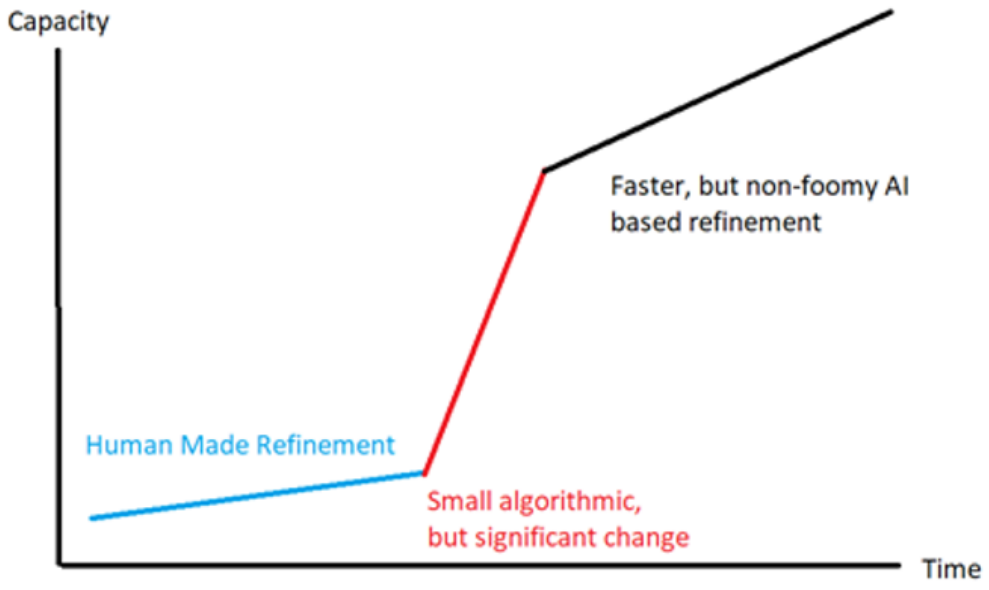
1million pre-historic humans *could* in theory organize themselves to better perform a set of cognitive tasks, including their own training.

I think this is one of the best pro-fast foom intuitions.

Let’s break this down a bit.

The first part talks about the ability of small algorithmic changes to have larger profound effects down the line. Imagine a company which operates large neural networks without back-propagation and trains them too slowly discovering back-prop. Large conceptual breakthroughs are extremely profitable, but that does not mean that they can occur quickly. If there are a lot of “ai-supporting” technologies, such as extra hardware without conceptual breakthroughs, that can create a situation conducive to fast takeoff. In other words, the thesis here is that instead of a long process of “refinement cycles” for AI, there is a single simple set of code changes, which, when run the hardware at the time creates a stronger self-improvement cycle.

It could be the following graph:



First slow takeoff, then a single change that breaches a barrier, then refinement cycle that is faster than the previous slow takeoff, but slower than “fast foom.” Think AlphaGo here. First, we have a semi-continuous improvement in go play [slow changes at first], a couple general algorithmic changes that move the needle a lot forward, which then create a strong ability to learn from self-play [analogous to RSI]. However, once that ability has been established, the improvement once again proceed more gradually.

Small changes to algorithms can have a large profound effect, however the crux here has to do with the search algorithm that is used to produce those changes. Just because an algorithmic change is “small” in terms of bits does not necessarily mean that it “easy” to find. If the human researchers have a difficult time finding these changes and only come up with one 50% improvement breakthrough every few months or so, and the newly designed artificial researchers can come up with one breakthrough 50% faster, then we are still going to have a slowish foom, even though completely human-less recursive self-improvement is happening.

So, while small changes being able to produce large returns is somewhat of an argument for fast foom, however it does not change the first crux - how much of the development cycle are you able to automate away.

The second part of the chimp human intuition is the question about the ability of the “pile of human-powered” algorithms to accomplish a set of tasks that are done by humans well enough to replace them. This easily happens in some domains.

a) Calculators can do many more calculations than humans trivially, once chess engines beat humans, they never looked back. Etc, etc.

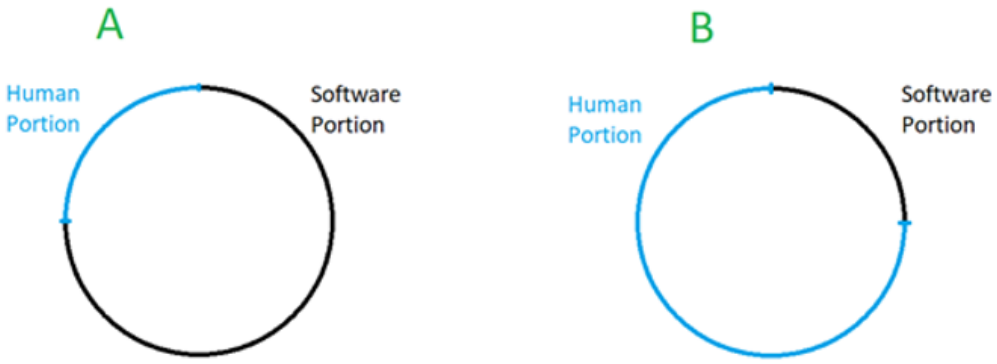
b) An algorithm running 1 million cavemen minds may or may not be able to take over the world.

c) A set of algorithms simulating an individual human connected in non-productive ways might not be accomplishing the relevant tasks well.

The question here is how long does it for researchers to develop algorithms to traverse the range of human skill and approach organizational skill in any task? In other words, given a simple human-level scientist algorithm, how long does it take to turn it into a world-class researcher? This, of course depends on the range of human abilities in this area. In other words, the amount of time it takes an average human-level algorithm to become a world class persuader or researcher might be longer than it takes a “human level algorithm” in chess to become a world class chess player. So, while the difference within humans is still a lot smaller than between humans and chimps, it’s possible that there would need to be a key set of breakthroughs in either understanding “organization” of intelligence or human differences that occurs between “human-level” and “able to replace humans.”

So, to review the first crux – how much of the human intervention exists in the loop in improving AIs.

This somewhat has two different images of the development cycle. Image A is the cycle of humans having ideas that are mostly wrong and testing them using a lot of hardware. Image B is the cycle of humans having ideas that are mostly right but are slow to generate and generally require arm-chair debates and thinking without a significant intervention of computers.



If you replace humans in Image A, you have a slow takeoff. If you replace humans in Image B, you could have a fast takeoff - with a couple caveats. If humans are a small part of the overall process in, BUT they are disproportionately slowing it down, then it could be the case that replacing them could remove a significantly larger portion of the overall process. This

could happen for several reasons, such as financial incentives of the people involved do not create the necessary conditions for rapid progress or insights exist in epistemic blind spots.

However, if we ignore this caveat, we can wonder about the process of AI research and development in each image. To me the first crux reduces to how much can people “think of” AI without using evidence of outcomes of complex computations.

Basically, in Image A, where a group of researchers draw from a large sample of potential improvements to the code of AI, or the training process or a development process, they must do a lot of computational work to verify whether the idea is good or not. It’s possible that the more meta the idea, the more work they need to do. If there is a powerful narrow AI that is already deployed and functioning on the real-world, you might need to run real-world tests, which could take weeks if not more. I encounter this situation in my current work and it’s not clear that this will get any easier.

In Image B, a group of researchers can use their reasoning capacity to make improvements which are easy to verify. The ease of verification depends on availability of test environments. Alpha-Go is an interesting case study. If a general-purpose intelligence breakthrough happens to improve an existing go-playing algorithm over the standard baseline, then it’s easy to verify. If we go up a meta-level up and there is an intelligence breakthrough that enables a general-purpose intelligence to better train narrow intelligences and they are measured by playing go, that also easy-ish to verify. Even that could take weeks of datacenter time today, but that is likely to be reduced in the future.

This brings me to the second crux – how easy is it to verify improvements, especially “meta” improvements for the people involved in AGI development. Verification has many types, such as ability to test ideas in simple environments and ability to quickly reject ideas based on other forms of supporting evidence, simply due to the strength of the researcher’s rationality. This has been pointed out by Robin Hanson as a problem for the AI itself for fast foom proponents in this piece.

“Merely having the full ability to change its own meta-level need not give such systems anything like the wisdom to usefully make such changes.”

However, keep in mind, we are considering this in the view of whether human researchers have the ability to make these meta-level changes.

I highly doubt that AGI insights can be easily tested in small standard environments. And from the history of AI, to my current work, to the discourse around AI in popular culture, it seems to me that rationally navigating the space of AI improvements is extremely hard for nearly everyone. It is theoretically possible that we, as a society, could get better in training researchers to better navigate the search space of improvement ideas, but I do not assign a very high probability on this.

The second crux could be examined considering another analogy, brought by Robin Hanson in the original foom debate here

Why don’t companies manufacturing productivity tools “go foom”? If they can work on making better tools, they get better at making tools and they can go on improving this process for a long time.

The company analogy is not perfect for several reasons that have been pointed out already. Companies have a tough time aligning all the sub-processes to work together, they have a tough time keeping insights out of the hands of competitors, and as employees get richer, it becomes harder to keep them. However, there is a key piece of the analogy that does bear repeating – companies (and individuals) do not necessarily have the wisdom to know whether a particular meta-optimization is a good idea or not.

Do agile development, open office plans and hiring consultants improve or hinder things? It's hard to run A/B tests on these development and different people will give different opinions on the effectiveness of these interventions, partly due diverging incentives.

While one can think that algorithms don't necessarily suffer from political problems (yet!), there are several issues that arise the moment one steps a meta-level above the current ML paradigm. I can think of several key problems:

a. Defining better metrics or sub goals is a hard process to systematically measure

b. Space of actual improvements is potentially rather small compared to space of possible changes

c. Sequential improvements could be anti-inductive in nature. Because a set of tricks worked in the past, may, in fact, be *counter-evidence* to them working in the future. This can trip up both people and algorithms. This is the fundamental problem of simply counting "returns" on cognitive re-investment.

d. Real world feedback could be a hard-to eliminate bottleneck

Summary of all the above reasons leads me to believe that it's possible that negative spirals can dominate the improvement landscape. The model is simple - if we have 1% of plausible changes are improvements, and the measuring system can recognize them all, but also labels 3% of the rest as false positive improvements, then on average the system will get worse over time. (I am assuming all changes are the same magnitude). "Development hell" is the company analogy of this, as additional features move the product farther away from shipping.

Note, that these problems apply to both human and artificial researchers alike. The way this relates to the fast / slow takeoff debate is still whether the way that the human researchers solve this. If they solve this through using a large amount of computation and real-world testing, then their reasoning and idea generation is not the bottleneck of the process and takeoff will be slow, if they solve this through using more and more advanced math to produce better and better abstractions, then their reasoning is the bottleneck and takeoff could be fast. Of course, if the problems are sufficiently advanced we may end up in a situation where even the fast takeoff requires a large amount of computation or we might not have a takeoff at all. The last possibility can arise if misaligned narrow AIs cause severe problems in the civilization to prevent further development. The simple example is that an AI can design the world's most addicting video game, which then everybody plays, and the real-world economy collapses. This, by itself, is a form of civilizational, though not existential risk.

The last analogy that is worth pondering is the example of difference between nuclear weapon development and nuclear weapon explosion.

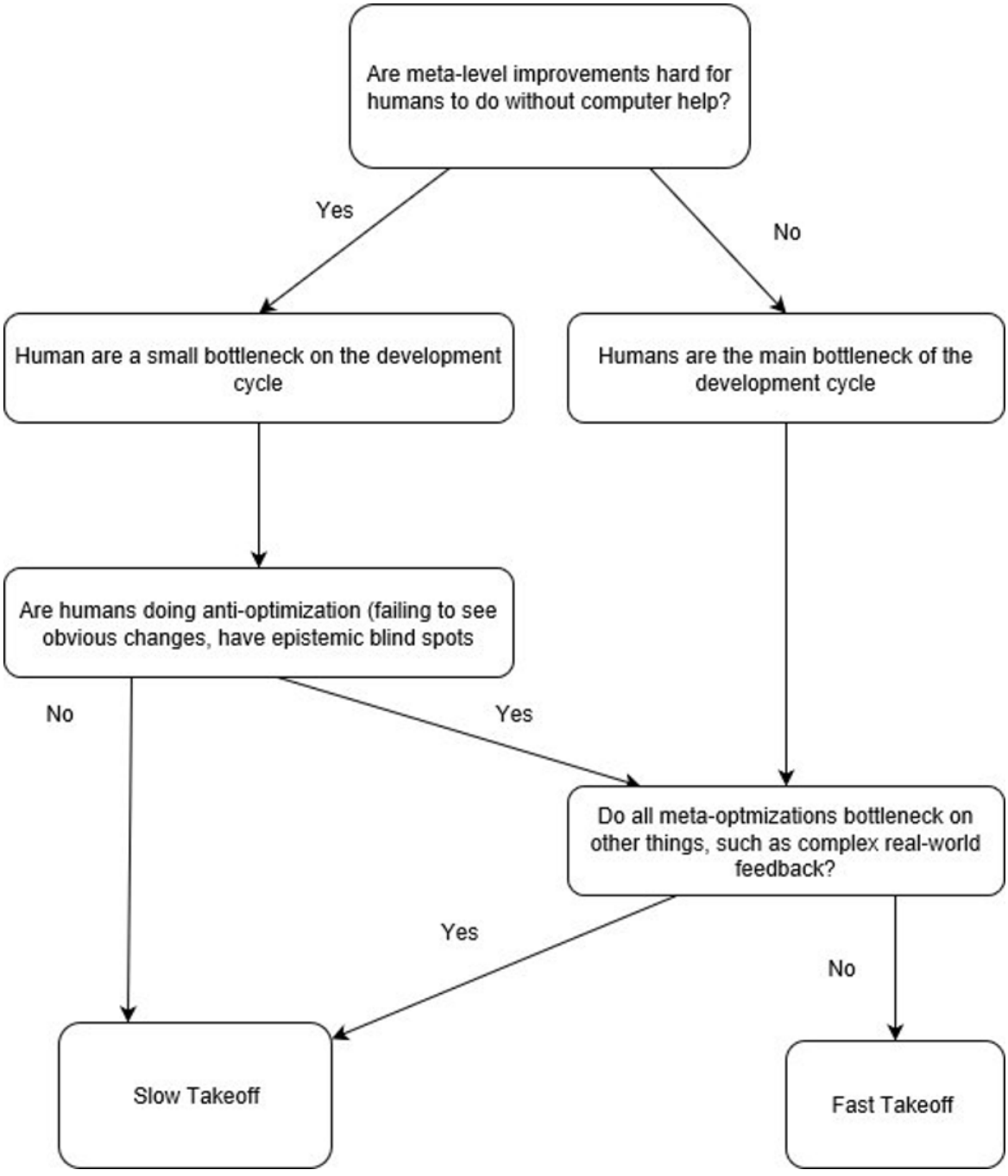
An important interplay in the nuclear weapon example is a simultaneous combination between self-enhancing takeoff / explosion and self-limiting / anti-inductive takeoff explosion. As the nuclear weapon explodes, it initially creates more and more relevant particles causing more and more explosions. This is the self-enhancing piece. It also consumes more and more of the material present in the bomb. This is the self-limiting piece. So, while the nuclear weapon explodes *a lot* faster than the speed of nuclear weapon development, it is not exempt from eventually being over-ruled by a lack of key non-renewable resource. Of course, when applied to an AI, this idea of a "fast foom" followed by a "slow foom" might still be incredibly dangerous.

Summary:

My current opinion on the second crux is that in practice, meta improvements are hard to do for people without an assistance of a large amount of computing resources. This points to a

slow takeoff, however this situation could change in the future, so I don’t want to completely rule out fast takeoff as a possibility.

Graphical Summary:



The many ways AIs behave badly

EDIT: This has been previously posted here. Vika is now maintaining a centralized list of such examples.

I had a previous post about some of the ways AIs behave badly. But now there is a new paper, looking at many examples of (mis)behaviour, within the evolutionary programming design. A video summary of some of the results is here.

So note that these are ways that current agents already (mis)behave; theses are not theoretical arguments about what might happen with a future superintelligence.

These behaviours include:

- solving the proxy/heuristic but not the proper problem (eg spinning while falling to get the highest score on a "jump" objective),
- cheating on the test (eg playing dumb on a test so that they could get a higher score afterwards),
- exploiting bugs in the environment (eg quickly twistching body parts to accumulate errors in the physics simulator and thus get "free energy" to propel themselves fast through virtual water),
- agents deliberately crashing other agents (requesting absurdly distant moves on an unbounded tic-tac-toe game, causing the other agents to dynamically expand their memory too much and then crash)
- unexpectedly elegant "impossible" solutions (crawling on its elbows when the percentage of time its feet could touch the ground was sent to 0%), and
- parasitism (in Tierra, an artificial life system, not only were there parasites, but parasites of parasites).

Abstract:

> Biological evolution provides a creative fount of complex and subtle adaptations, often surprising the scientists who discover them. However, because evolution is an algorithmic process that transcends the substrate in which it occurs, evolution's creativity is not limited to nature. Indeed, many researchers in the field of digital evolution have observed their evolving algorithms and organisms subverting their intentions, exposing unrecognized bugs in their code, producing unexpected adaptations, or exhibiting outcomes uncannily convergent with ones in nature. Such stories routinely reveal creativity by evolution in these digital worlds, but they rarely fit into the standard scientific narrative. Instead they are often treated as mere obstacles to be overcome, rather than results that warrant study in their own right. The stories themselves are traded among researchers through oral tradition, but that mode of information transmission is inefficient and prone to error and outright loss. Moreover, the fact that these stories tend to be shared only among practitioners means that many natural scientists do not realize how interesting and lifelike digital organisms are and how natural their evolution can be. To our knowledge, no collection of such anecdotes has been published before. This paper is the crowd-sourced product of researchers in the fields of artificial life and evolutionary computation who have provided first-hand accounts of such cases. It thus serves as a written, fact-checked collection of scientifically important and even entertaining stories. In doing so we also present here substantial evidence that the existence and importance of evolutionary surprises

extends beyond the natural world, and may indeed be a universal property of all complex evolving systems.

An Argument For Prioritizing: "Positively Shaping the Development of Crypto-assets"

This is a linkpost for http://effective-altruism.com/ea/1ms/an_argument_for_prioritizing_positively_shaping/

Recommendations vs. Guidelines

Medicine loves guidelines. But everywhere else, guidelines are still underappreciated.

Consider a recommendation, like “Try Lexapro!” Even if Lexapro is a good medication, it might not be a good medication for your situation. And even if it’s a good medication for your situation, it might fail for unpredictable reasons involving genetics and individual variability.

So medicine uses guidelines – algorithms that eventually result in a recommendation. A typical guideline for treating depression might look like this (this is a very over-simplified version for an example only, NOT MEDICAL ADVICE):

1. Ask the patient if they have symptoms of bipolar disorder. If so, ignore everything else on here and move to the bipolar guideline.

2. If the depression seems more anxious, try Lexapro. Or if the depression seems more anergic, try Wellbutrin.

3. Wait one month. If it works perfectly, declare victory. If it works a little but not enough, increase the dose. If it doesn’t work at all, stop it and move on to the next step.

4. Try Zoloft, Remeron, or Effexor. Repeat Step 3.

5. Cycle through steps 3 and 4 until you either find something that works, or you and your patient agree that you don’t have enough time and patience to continue cycling through this tier of options and you want to try another tier with more risks in exchange for more potential benefits.

6. If the depression seems more melancholic, try Anafranil. Or if the depression seems more atypical, try Nardil. Or if your patient is on an earlier-tier medication that almost but not quite works, try augmenting with Abilify. Repeat Step 3.

7. Try electroconvulsive therapy.

The end result might be the recommendation “try Lexapro!”, but you know where to go if that doesn’t work. A psychiatrist armed with this guideline can do much better work than one who just happens to know that Lexapro is the best antidepressant, even if Lexapro really *is* the best antidepressant. Whenever I’m hopelessly confused about what to do with a difficult patient, I find it really reassuring that I can go back to a guideline like this, put together by top psychiatrists working off the best evidence available.

This makes it even more infuriating that there’s nothing like this for other areas I care about.

Take dieting. Everybody has recommendations for what the best diet is. But no matter what diet you’re recommending, there are going to be thousands of people who tried it and failed. How come I’ve never seen a diet guideline? Why hasn’t someone written something like:

1. Try cutting carbs by X amount. If you lose Y pounds per week, the diet is working. If not, you're probably resistant to cutting carbs because [two hours of mumbling about insulin] and you should move on to the next tier.

2. Try cutting fat by X amount. If you lose Y pounds per week, the diet is working. If not, you're probably resistant to cutting fat because [two hours of mumbling about leptin], and you should move on to the next tier.

And so on until Step 7 is "get a gastric bypass".

I agree nobody can ever make a perfect algorithm that works for all eventualities. But still. Surely we can do better than "Try the Paleo diet! I hear it's great!"

What information do guidelines carry beyond a recommendation?

First, they have more than one recommendation. It may be that the Paleo diet is the best, but the guidelines will also include which is the second-best, third-best, et cetera.

Second, because they have more than one recommendation, they can tailor their recommendation to your specific circumstances. The person with depression and comorbid anxiety may want to start with Lexapro; the person whose main symptom is tiredness may want to start with Wellbutrin. Since I love bread, does that mean I should avoid carb-cutting diets? Does that mean it's extra-important that I cut carbs? Does it not matter, and really it depends on whether I have a family history of diabetes or not?

Third, they acknowledge that some people might need more than one recommendation. If you hear "try the Paleo diet", and then you try it, and it doesn't work, you might believe you're just a bad dieter, or that all diets are scams, or something like that. Guidelines implicitly admit that everyone is different in confusing ways, that something that's expected to work for many people might not work for you, and that you should expect to have to try many things before you find the right one.

Fourth, because they admit you may need to try more than one thing, they contain (or at least nod at) explicit criteria for success or failure. How long should you try the Paleo diet before you decide it doesn't work? How much weight do you need to lose before it qualifies as "working"? If it's been three months and I've lost four pounds, should you stick with it or not?

Fifth, they potentially contain information about which things are correlated or anticorrelated. The depression guidelines make it clear that if you've already tried Lexapro and Zoloft and they've both failed, you should stop trying SSRIs and move on to something with a different mechanism of action. If I've tried five carb-cutting diets, should I try a fat-cutting diet next? If I hate both Mexican food and Chinese food, is there some other category of food which is suitably distant from both of those that I might like it? Guidelines have to worry about these kinds of questions.

My impression is that once you understand a field really well, you have something like a Guideline in your mind. I think if nobody had ever written a guideline for treating depression, I could invent a decent one myself out of everything I've pieced together from word-of-mouth and common-sense and personal experience. In fact, I think I do have some personal guidelines, similar to but not exactly the same as the official ones, that I'm working off of without ever really being explicit about it. Part of the

confusion of questions like “What diet should I do?” is sorting through the field of nutrition until you can sort of imagine what a guideline would look like.

So why don’t people who have more knowledge of nutrition make these kinds of guidelines? Maybe some do. I can’t be sure I haven’t read dieting guidelines, and if I did I probably ignored them because lots of people say lots of stuff.

But I think that’s a big part of it – making guidelines seems like a really strong claim to knowledge and authority, in a way that a recommendation isn’t. Some idiot is going to follow the guidelines exactly, screw up, and sue you. I just realized that my simplified-made-up depression guidelines above didn’t have “if the patient experiences terrible side effects on the antidepressant, stop it”. Maybe someone will follow those guidelines exactly (contra my plea not to), have something horrible happen to them, and sue me. Unless you’re the American Psychiatric Association Task Force or someone else suitably impressive, your “guidelines” are always going to be pretty vague stuff that you came up with from having an intuitive feel for a certain area. I don’t know if people really want to take that risk.

Still, there are a lot of fields where I find it really annoying how few guidelines there are.

What about nootropics? I keep seeing people come into the nootropics community and ask “Hey, I feel bad, what nootropic should I use?” And sure, eventually after doing lots of research and trying to separate the fact from the lies, they might come up with enough of a vague map of the area to have some ideas. But this is an area where “Well, the first three things you should try for anxiety are…” could be really helpful. And I don’t know of anything like that – let alone something that tells you how long to try before giving up, what to look for, etc.

Or let’s get even broader – what about self-help in general? I don’t really believe in it much, but I would love to be proven wrong. If there were a book called “You Are Willing To Devote 100 Hours Of Your Life To Seeing If Self-Help Really Works, Here’s The Best Way For You To Do It”, which contained a smart person’s guidelines on what self-help things to try and how to go about them, I would absolutely buy it.

5 general voting pathologies: lesser names of Moloch

Earlier, I wrote a primer on voting theory. Among the things I discussed were 5 types of pathologies suffered by different single-winner voting methods. I presented these as 5 sequential hurdles for voting method design. That is, since they are in what I view as decreasing importance and increasing difficulty, you should check your voting method against each hurdle in order, and stop as soon as it fails to pass.

Then I read Eliezer's book on Inadequate Equilibria, and Scott's "Meditations on Moloch". They argue that the point of civilization is to provide mechanisms to get out of pernicious equilibria, and the kakistotropic tendencies of civilization they characterize as "Moloch" are basically cases where pernicious incentives reinforce each other. I realized that the simple two-player games such as Prisoners' Dilemma that serve as intuition pumps for game theory lack some of the characteristics of my 5 voting pathologies. So I want to go back and explain those pathologies more carefully, to help build up intuition about how multi-player, single-outcome games differ from two-player ones.

A key point here is that I'm talking about single-winner voting methods; that is, "games" where the number of possible outcomes is far less than the number of players. In this case, it's not a matter of seeking an individual advantage for yourself; the only way for you to win is for your entire faction to win equally. This means that I will not be talking about the oldest and deepest name of Moloch, which is Malthus. All the Molochs in this essay can and should be killed or (mostly) tamed.

Also note that this essay is not the one I'd write if I were only trying to recruit the rationalist community to become electoral reform activists. As an activist, I think that the most important and short-term-viable electoral reforms are in the multi-winner space: solving the problem of coordinating public goods not directly through mechanism design, but indirectly through a combination of mechanism design and representation. Some of my reasons for thinking that are contingent and have no place here. The one that's not: I think that the problem of "ain't nobody got time for all that politics" is worse than the principal-agent problem of a well-designed representative mechanism. Regardless, I think that this community would rather hear first about these names for Moloch.

In order, my pathologies — hurdles for multi-agent shared-outcome mechanism design — are:

Dark Horse

Let's say that you have a 3-candidate election using the Borda count, and your electorate has the following true utilities:

49: A9.0 B1.0 D0.0

48: A1.0 B9.0 D0.0

3: A1.0 B0.0 D9.0

Under the Borda count, each voter must give the three candidates 2, 1, and 0 points in some order. If the B voters strategize, the election might look like:

49: A2 B1 D0

48: A0 B2 D1

3: A1 B0 D2

B wins with a total of 145. The A voters might try to retaliate with a similar strategy:

49: A2 B0 D1

48: A0 B2 D1

3: A1 B0 D2

But now D wins with a total of 103, even though D was honest last preference for 97% of voters.

This "Dark Horse 2" example becomes even harder to resolve if you make it "Dark Horse 3":

34: A9.0 B2.0 C1.0 D0.0

33: A2.0 B9.0 C1.0 D0.0

33: A2.0 B1.0 C9.0 D0.0

I'll let you work it out for yourself, but the upshot is that each group has an incentive to give D the second-most points; that if one or two groups are strategic, they can profit; but if all three are strategic, all of them lose. D can win in this situation with literally zero honest support — an epically pathological result.

What does it feel like in this situation:

To win honestly? "All is right with the world."

To weakly-lose when everyone's honest? "I am slightly tempted to strategize."

To weakly-lose when the opponents are strategic? "I need to stop being a sucker, and counter-strategize."

To win strategically? "I feel a little bit guilty, but at least I won."

To strongly-lose strategically? "WTF? This system sucks. If possible, I should change it. If not, maybe I should learn my lesson and not strategize. But regardless, those other evil sneaky strategizers against me MUST learn theirs."

This is the closest to a standard prisoners dilemma of all of the voting pathologies. As with the standard prisoners dilemma, "social glue" (that is, heuristics developed through successful cooperation in iterated scenarios) can generally avoid breakdown. But it's also the easiest to avoid using mechanism design: just don't use the Borda count (or any other strictly-ranked point-based method). That is to say, don't force people to dishonestly support D in merely in order to oppose some other candidate.

So "Dark Horse" is a name for a Moloch that's outstandingly evil but not particularly powerful.

Lesser evil

If you live in the US, UK, Canada, or India — or any other country that uses First Past the Post voting — you already know this Moloch well. In a system where you can only vote for one, you'd better not "waste your vote" on the option you most truly support; you must instead support the lesser evil, the least-bad of the viable options. The logical end-point is a world with only two options, each of which has far stronger incentives to make the other side look bad than to actually pursue the common good. If you're lucky, one or both of those two options will pursue the common good for the fun of it; if you're unlucky, they'll each be as corrupt as they can get away with without losing support to the other side; but either way, there's relatively little you can do about it.

Of course, I should point out that this game theory doesn't always play out exactly in real life. The US has only 2 parties that matter, but most other FPTP countries have a bit more than that, even if the top two matter more than they should. So if you want to continue to spar with the teeth of this Moloch instead of just cutting off its head, OK, you're not doomed to lose every time. Just most of the time.

In terms of election scenarios, this looks something like the following. Utilities are:

15: A9.0, B8.0, C0.0

36: A8.0, B9.0, C0.0

24: A0.0, B9.0, C8.0

25: A0.0, B1.0, C9.0

Votes are:

15+36=51: A

24+25=49: C

This is an equilibrium because, in most games where there are far more players than outcomes, almost everything is an equilibrium; no one voter could get a better outcome by changing their vote, even though the society as a whole would be far happier if they could elect B. Any A voter who moved to B would be helping C win; any C voter who moved to B would be making it easier for A to win, even if next election honest C>A voters are a majority.

I probably don't have to tell you what this one feels like, but here goes anyway:

On top of the winning coalition (15 A voters): "All is right with the world."

On the bottom of the winning coalition (36 B>A>C voters): Conflicted. On the one hand, "the lesser evil is still evil". On the other hand, "a vote for B is a vote for C". Both are true; this dilemma is inescapable without changing the voting method. Short-term incentives favor continuing to vote for A, and in fact actively suppressing

discussion of A's flaws and B's ideas; but human nature favors getting mad at A and exaggerating their flaws. Either way, mind-killing is likely.

On the bottom of the losing coalition (24 B>C>A voters): Enraged. Ripe for a demagogue.

On the top of the losing coalition (25 C voters): Must... try... harder. Next time, we'll win!

This is a lesser Moloch, in that we could easily kill it by changing the voting method. Note that proportional representation can (if it's done well) be just as good at killing this Moloch as the single-winner methods discussed below! But it's still strong enough to rule over most of you who are reading these words.

Center Squeeze

OK, you say; if the Lesser Evil is enabled by the existence of wasted votes, let's fix that by moving all the votes until they're not wasted. You've just invented Instant Runoff Voting (IRV). Each voter ranks the candidates; votes are piled up by which candidate they rank first; and then, iteratively, the smallest pile is eliminated and those votes are moved to whichever remaining pile they rank highest (if any). You can stop as soon as one pile has a majority of remaining votes, because that pile is guaranteed to win.

This would solve the spoiler problem of the 2000 Florida presidential election. Here's a simplified version of utilities in that scenario (B/G/N stand of course for Bush/Gore/Nader):

490: B9.0 G1.0 N0.0 (Bush>Gore)

100: B1.0 G9.0 N0.0 (Gore>Bush)

389: B0.0 G9.0 N1.0 (Gore>Nader)

10: B0.0 G1.0 N9.0 (Nader>Gore)

6: B0.0 G0.0 N9.0 (Nader>nobody)

5: B1.0 G0.0 N9.0 (Nader>Bush)

Under FPTP, honest voting would "spoil" the election and let Bush win. But under IRV, the Nader supporters can vote honestly; when Nader is eliminated, those votes will transfer, so Gore will beat Bush 499 to 495.

But what happens if Nader appeals to more voters, and 300 of the Gore>Nader voters shift to Nader>Gore? That would mean that Nader had 321 first-choice supporters, and Gore only 189. So Gore would be eliminated first, 100 of those votes would shift to Bush, and Bush would win! In this scenario, the centrist Gore was "squeezed" on both sides and prematurely eliminated, even though he could have beaten either of the others in a 1-on-1 race.

And the result is that, just like in the real election, Nader's supporters ended up helping cause the election of Bush, the candidate most of them like the least. That spoilage doesn't happen until after Nader passes 25%, but it still happens. And this

problem is real; it happened in the Burlington 2009 mayoral election (though in that case, the voters whose honesty worked against them were the Republicans).

Now, Center Squeeze is a much smaller problem than Lesser Evil. If you have a choice, you'd rather run a race with a minefield between 25% and 50% of the way, than one where the minefield stretches from the beginning up to 50%. If you're skillful, maybe you can build up enough speed in the first 25% to leap over the minefield. And parties that stay under 25% can at least get more attention than those who are stuck around 0% as in Lesser Evil.

What does this one feel like?

Win, not spoiled: "All is right with the world."

Small fringe party, vote honestly, still matter: "At least I tried."

Medium fringe party, vote honestly, spoil the election: Dilemma. Some will decide to be strategic; others will say "wasn't my fault. It was the fault of those treacherous centrists who ranked the greater evil as their second choice."

Centrist, lose due to spoilage: "Huh? What happened? We're the rightful Condorcet winners, how can we lose?"

Large fringe party, win due to spoilage on the other side: "Ha! My far-off enemies were so disgusting that some of my nearby former enemies joined my cause! I deserved that."

Large fringe party, don't win: "Hmm... how can I divide my enemies?"

This Moloch is a relatively benign one, who acts to protect incumbent winners but allows dissenting voices up to a certain point. Living under its reign (as, arguably, Australia now does) involves occasional craziness but is mostly OK. Still, it can be killed.

Chicken Dilemma

This scenario actually exists in two separate versions, depending on the voting method: slippery and non-slippery slope. Both share the same underlying voter utility scenario, with two similar candidates who must team up in order to beat a third one:

35: A9.0 B8.0 C0.0 (A>B)

25: A8.0 B9.0 C0.0 (B>A)

40: A0.0 B0.0 C9.0 (C)

For the slippery slope version, let's assume the election uses approval voting: voters can approve as many candidates as they want, and the most approvals wins. If voters approve any candidate with a utility above 5.0, the ballots will be:

35+25=60: AB

40: C

A and B end up in an exact tie for first place (as Burr and Jefferson did in 1800; thus, the chicken dilemma is sometimes called the Burr dilemma). C, the candidate whom the majority opposes, has been safely defeated; but the outcome between A and B is essentially random. Incentives are clearly high for the first two groups of voters to approve only their favorite candidate. If 1 of the A>B voters votes for only A, then A wins; but then, 2 of the B voters can get B to win by switching to only B; and next 2 more A voters defect; etc. It's a slippery slope until over 20 of each group defect, and then C wins, an outcome the majority hates.

In game theory terms, this is a "chicken" or "snowdrift" game, with 2 equilibria: either the A voters stably cooperate and the B voters stably defect, so that B wins, or vice versa. But in emotional terms, neither of these equilibria feel stable: both are arguably "unfair" cases where one group is exploiting the other's cooperation. It might be "fair" if the smaller group was reliably the one to cooperate, but that's hard to coordinate in practice in cases where the sizes are similar, both sides will probably bet that they are the larger group. So in practical terms, probably the more "stable" outcomes are "both enforce cooperation, and hope there's some odd C voters who care enough to swing the election one way or the other", or "both bicker and defect".

To improve matters, we can use a non-slippery-slope voting method such as 3-2-1 voting. In this method, voters rank each candidate "good", "OK", or "bad", and the winner is decided in 3 steps. First, choose 3 semifinalists, those with the most "good" ratings; then of those, choose 2 finalists, those with the fewest "bad" ratings; then of those, the winner is the one rated higher on more ballots (the pairwise winner).

(When choosing the third semifinalist, there are two additional rules. First, to avoid a clone-candidate incentive, they must not be from the same party as both of the first two or, in a nonpartisan race, do not count their "good" ratings on the same ballots as also rated the first semifinalist "good". Second, to avoid a dark horse issue, they must have at least 1/2 as many "good" ratings as the first semifinalist. If no candidate meets these criteria, then skip step 2.)

In this method, if each voter votes honestly, then all 3 will be semifinalists (eliminating any also-rans whom we left out of the scenario for simplicity); A and B will be finalists (eliminating the majority loser C); and A will win, as the honest pairwise winner between those two.

It's still possible, in this scenario, for 21 B voters to defect, rate A as "bad", and cause B to win. But if under 20 of them do so, it doesn't change the result. Thus, there's no "slippery slope". Even though "everyone cooperates" is not a strong Nash equilibrium in strict game theory terms, it is probably strong enough to endure in practical terms.

Is it possible to make a voting method without even a non-slippery chicken dilemma? Yes, we've already seen that: IRV. But since defectors in the chicken dilemma look exactly like fringe voters in center squeeze, it's impossible to fully solve the chicken dilemma like this without creating a center squeeze problem — one I'd argue is worse, at least as compared to the non-slippery CD.

What does a non-slippery CD feel like? If both sides cooperate, I'd argue that it feels basically fair to everyone involved. If the smaller side wins through strategic defection, that feels unfair, and technically it's an equilibrium; but I'd argue that human stubbornness is enough to counter-defect as a punishment, and thus iterate back to cooperation. 9 So non-slippery CD isn't really Moloch at all. And as for slippery CD... it's mean, but capricious, and can sometimes be distracted or overcome.

Condorcet Cycles

Here's the scenario. Instead of utilities, I'll just give preferences, because there's almost no way to make this one "realistic".

34: A>B>C

33: B>C>A

33: C>A>B

This scenario is so unavoidably strategic that it's at the heart of a proof of the Gibbard-Satterthwaite theorem that no (non-dictatorial) voting method can entirely avoid strategy. If one of the three groups preemptively throws their favorite under the bus and embraces their second choice, the ballots will show at least a 66% majority for that second choice, so any democratic voting method will elect that candidate. So to all three groups, this situation will feel like a dilemma between racing to signal they'll compromise first and most convincingly, or hoping that the group before them in the cycle makes the compromise.

In practice, Condorcet cycles probably happen only 1-5% of the time. This is true in the most sophisticated voter utility models I can create (hierarchical "crosscat" Dirichlet clusters in ideology/priority space), and also in empirical evidence (where cyclical preferences seem rare but not nonexistent). So this last lesser Moloch is one which can never be defeated, but which spends most of its time in the deep woods and only occasionally rampages out, doing surprisingly little damage in the process.

Conclusion

I set out to write this because I thought that multiplayer game theory has some fundamental differences from single-player game theory and specifically that we need to stop leaning so hard on the prisoners' dilemma. Having written it, I realize that though I touched on these issues, I spent most of the time going over more basic points of voting methods. So I'm not sure this essay is exactly what I wanted it to be, but I think what it is can still be at least somewhat useful; I hope you feel the same way.

I guess my larger point is that evolution has actually equipped us pretty well with social strategies for dealing with PD or CD, but that by that same token we humans are particularly subject to pernicious equilibria of the "lesser evil" variety. The feeling of "we all agree these aren't the best options but looking for better ones would waste energy we need to spend fighting against the worse one" (lesser evil) seems like at least as important a paradigm of Moloch as "if I weren't evil someone else would be" (tragedy of the commons/multiplayer prisoner's dilemma/dark horse). It's important to remind ourselves that mechanism design offers a way out of lesser evil (and thus also center squeeze); not just in politics, but wherever it occurs.

Can corrigibility be learned safely?

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

EDIT: Please note that the way I use the word "corrigibility" in this post isn't quite how Paul uses it. See this thread for clarification.

This is mostly a reply to Paul Christiano's Universality and security amplification and assumes familiarity with that post as well as Paul's AI alignment approach in general. See also my previous comment for my understanding of what corrigibility means here and the motivation for wanting to do AI alignment through corrigibility learning instead of value learning.

Consider the translation example again as an analogy about corrigibility. Paul's alignment approach depends on humans having a notion of "corrigibility" (roughly "being helpful to the user and keeping the user in control") which is preserved by the amplification scheme. Like the information that a human uses to do translation, the details of this notion may also be stored as connection weights in the deep layers of a large neural network, so that the only way to access them is to provide inputs to the human of a form that the network was trained on. (In the case of translation, this would be sentences and associated context, while in the case of corrigibility this would be questions/tasks of a human understandable nature and context about the user's background and current situation.) This seems plausible because in order for a human's notion of corrigibility to make a difference, the human has to apply it while thinking about the meaning of a request or question and "translating" it into a series of smaller tasks.

In the language translation example, if the task of translating a sentence is broken down into smaller pieces, the system could no longer access the full knowledge the Overseer has about translation. By analogy, if the task of breaking down tasks in a corrigible way is itself broken down into smaller pieces (either for security or because the input task and associated context is so complex that a human couldn't comprehend it in the time allotted), then the system might no longer be able to access the full knowledge the Overseer has about "corrigibility".

In addition to "corrigibility" (trying to be helpful), breaking down a task also involves "understanding" (figuring out what the intended meaning of the request is) and "competence" (how to do what one is trying to do). By the same analogy, humans are likely to have introspectively inaccessible knowledge about both understanding and competence, which they can't fully apply if they are not able to consider a task as a whole.

Paul is aware of this problem, at least with regard to competence, and his proposed solution is:

> I propose to go on breaking tasks down anyway. This means that we will lose certain abilities as we apply amplification. [...] Effectively, this proposal replaces our original human overseer with an impoverished overseer, who is only able to respond to the billion most common queries.

How bad is this, with regard to understanding and corrigibility? Is an impoverished overseer who only learned a part of what a human knows about understanding and

corrigibility still understanding/corrigible enough? I think the answer is probably no.

With regard to understanding, natural language is famously ambiguous. The fact that a sentence is ambiguous (has multiple possible meanings depending on context) is itself often far from apparent to someone with a shallow understanding of the language. (See here for a recent example on LW.) So the overseer will end up being overly literal, and misinterpreting the meaning of natural language inputs without realizing it.

With regard to corrigibility, if I try to think about what I'm doing when I'm trying to be corrigible, it seems to boil down to something like this: build a model of the user based on all available information and my prior about humans, use that model to help improve my understanding of the meaning of the request, then find a course of action that best balances between satisfying the request as given, upholding (my understanding of) the user's morals and values, and most importantly keeping the user in control. Much of this seems to depend on information (prior about humans), procedure (how to build a model of the user), and judgment (how to balance between various considerations) that are far from introspectively accessible.

So if we try to learn understanding and corrigibility "safely" (i.e., in small chunks), we end up with an overly literal overseer that lacks common sense understanding of language and independent judgment of what the user's wants, needs, and shoulds are and how to balance between them. However, if we amplify the overseer enough, eventually the AI will have the option of learning understanding and corrigibility from external sources rather than relying on its poor "native" abilities. As Paul explains with regard to translation:

> This is potentially OK, as long as we learn a good policy for leveraging the information in the environment (including human expertise). This can then be distilled into a state maintained by the agent, which can be as expressive as whatever state the agent might have learned. Leveraging external facts requires making a tradeoff between the benefits and risks, so we haven’t eliminated the problem, but we’ve potentially isolated it from the problem of training our agent.

So instead of directly trying to break down a task, the AI would first learn to understand natural language and what "being helpful" and "keeping the user in control" involve from external sources (possibly including texts, audio/video, and queries to humans), distill that into some compressed state, then use that knowledge to break down the task in a more corrigible way. But first, since the lower-level (less amplified) agents are contributing little besides the ability to execute literal-minded tasks that don't require independent judgment, it's unclear what advantages there are to doing this as an Amplified agent as opposed to using ML directly to learn these things. And second, trying to learn understanding and corrigibility from external humans has the same problem as trying to learn from the human Overseer: if you try to learn in large chunks, you risk corrupting the external human and then learning corrupted versions of understanding and corrigibility, but if you try to learn in small chunks, you won't get all the information that you need.

The conclusion here seems to be that corrigibility can't be learned safely, at least not in a way that's clear to me.

Hold On To The Curiosity

I.

Recently, an excited friend was telling me the story behind why we care about the mean, median and mode.

They explained that a straightforward idea for what you might want in an ‘average’ number, is something that minimises how far it is from all the other numbers in the dataset - so if your numbers are 1, 2 and 3, you want a number x such that the sum of the distance to each datapoint is as small as possible. It turns out this number is 2.

However, if your numbers are 1, 2, and 4, the number that minimises the distance from all of them is *also* 2.

Huh?

When my friend told me this, the two other people I was with sort of said “Okay”. I said “What? No! I don’t believe you! It has to change when the data does - it’s a linear sum, so it has to change! It’s like you’re saying the sum of 1, 2 and 3 is the same as the sum of 1, 2 and 4. This is just *wrong*." Suffice to say, my friend’s claim wasn’t predicted by my understanding of math.

Now, did I really not believe my friend? The other two people with us were certainly fine with it. Isn’t this just *bayesianism*? That’s how the old joke goes:

> **Math teacher:** Now I’m going to prove to you that X is true.
>
> **Bayesian:** You just did.

Actually, no. You taught me a detail to memorise, but my models didn’t improve. I won’t be able to improve how I use averages, because I don’t understand how it fits in with everything else I understand - it doesn’t fit with the models I use everywhere else in math.

I mean, I could’ve nodded along. It’s only one fact, after all. But if I’m going to remember it in the long term, it should connect to my other models and be reinforced. The alternative is to be stored in the brain with all those other memorised facts that students learn for exams and forget immediately after.

If you’re trying to build new models of a domain, it’s important to choose to speak from the confusion, not from the rest of yourself. Don’t have conversations about whether you believe a thing. Instead talk about whether you understand it.

(The problem above was the definition of the median, and an explanation of the math for the curious can be found in this comment.)

II.

It can be really hard to *feel* your models. Qiaochu Yuan’s method of learning involves ramping feeling-his-models up to 11. I recall him telling me about trying to learn what fire was once, where his first step was to just really feel his confusion:

> What the hell is this orange stuff? How on earth does it get here? Why is it flickering? WHAT IS FIRE?!

After feeling the confusion, Qiaochu holds onto his *frustration* (which he finds easier to hold), and tries throwing ideas and possible explanations at it until all the parts finally fit together - that feeling when you say "Ohhhhhhh" and the models finally compute, and your beliefs predict the experience you have. *Be frustrated with reality*.

Tim Urban (of WaitButWhy) tells a similar story, where he can only write essays about things he *doesn't currently understand* - and as he's digging through all the facts and pieces things together, he writes down the things that made sense to him, that would successful get the models across to an earlier version of Tim Urban.

I used to think this made no sense and he must just be bad at introspecting - shouldn't you have to build an excellent model of other people to write so compellingly for so many tens of thousands of them?

Yet it's actually really rare for authors to be strongly connected to *their own* models - when a teacher explains something for the hundredth time, they likely can't remember what it was like to learn it for the first. And so Tim's explanations can be clearer than most.

In the opening example where I was surprised by the definition of the median, if you had offered me a bet I would've bet on the side that this was the definition of a median. But it was not a useful thought for me in that moment, to set aside my confusion and say "On reflection I believe you". It can be correct in conversation, when your goal is understanding, to hold onto the confusion, the frustration, and let your models do the speaking.

III.

I often feel people try to move a conversation toward whether I believe the claim, rather than discussing and sharing what we each understand.

> "Do you *believe m*e when I say picking an average by minimising the distance to all the points is the same as the median?
>
> "Hmm, can you tell me *why* that's the case? I have a model of arithmetic that says it shouldn't be…"

A phrase I often use: *"You may have changed my betting odds but you haven't changed my models!*"

We're all in the game of trying to build models. Whether you're trying to understand the field of science you're attempting to add knowledge to, the product your startup is building, or the architecture of the AGI you're trying to align, you need good models to leverage reality for whatever you care about.

One of the most important skills in life is the ability to hold onto your confusion and let your models do the talking, so they can interface with reality more directly. Choosing to notice and hold on to your confusion is hard, and it's so easy to lose sight of it.

To put it another way, here are some perfectly acceptable noises to make when your goal is understanding:

What? No! I don’t believe you! That *can't* be true!

I expect that some but not all of this post is surprisingly Ben-specific. My thanks to Alex Zhu (zhukeepa) and Jacob Lagerros (jacobjacob) for reading drafts.

Some Simple Observations Five Years After Starting Mindfulness Meditation

BACKGROUND

I learned the basics of mindfulness meditation in April 2013, which was five years ago to the month.

The type of meditation I learned and would go on to practice would be closing one's eyes and concentrating on one's breath.

The goal would be to *literally* have one's concentration follow one's breathing — *literally* directing one's attention to the experience of inhaling and feeling air pass through one's nostrils when breathing through the nose, feeling the air pass through the throat/trachea, and feeling the air settle into one's lungs.

Then I'd hold the breath for ever so slightly, and follow the path of the air as I exhaled out the nose or mouth.

I meditated every single day for a few years, at first for a mere five minutes a day in the morning, and eventually ten minutes per day, and sometimes twice per day. Occasionally I'd do longer sessions in the 20-30 minute range, and sometimes I adopted somewhat similar mindfulness practices in the gym or when unable to fall asleep promptly.

I was never particularly strict in the position I'd meditate — sometimes I'd be sitting down, or sometimes I'd be laying down. I'd meditate wherever was convenient given the furniture layout of where I was at — often laying flat on top my bed after I'd gotten up for the day and made the bed, sometimes sitting at the kitchen table or at the sofa.

After a few years of this, one day I came to feel something like, "You know, I've gotten everything I can get out of this, and it's not very valuable any more." From there, I mostly stopped meditating for a couple years before picking the practice back up a couple months ago.

###

OBSERVATIONS

Here's a few simple observations I've taken from meditation, without commentary or lessons, which I'll get to in a moment —

1. Bringing Awareness to Automatic Practices: We're breathing constantly, but most of the time, we don't notice it. It happens more-or-less automatically without any conscious thought.

2. Constant Streams of Thought: It might sound simple, when described, to just concentrate on one's breath. Not so! Thoughts constantly arise, sometimes very quickly and sometimes less quickly. I never counted the frequency of thoughts unrelated to breathing that came into my awareness and concentration, but I reckon it was often 30+ thoughts even in just five minutes. Five minutes is 300 seconds; just 30

arising thoughts would be a new thought every 10 seconds... and this is when I was attempting to bring my full concentration to bear.

3. "Streaks" of Thinking: With that said, it seemed that my meditation was "streaky" — there'd be minutes where thoughts were nearly constantly arising. Sometimes, though, I'd settle into a rhythm and all the unrelated thoughts would seem to entirely stop for 2-3 minutes until the timer went off and the meditation session ended. So in a session with 30 unrelated thoughts, it might be experiencing those 30 thoughts in the first 120 seconds or so (a new thought every 4 seconds), followed by 180 seconds of seemingly no thought arising.

4. Inconsistent Perception of Time: Though I always attempted not to do so, during subjectively perceived difficult meditation sessions, I'd sometimes check how much time was remaining. Oftentimes, I'd have a subjective perception of incredible difficulty and lots of time passing... only to check the clock and see that only 30 seconds passed! Alternatively, when things "settled down," then often anywhere from 3+ minutes would pass seemingly instantly.

5. Subjective Experiences of Discomfort or Calm: For seemingly no reason and often unpredictably, sometimes I'd experience a lot of discomfort when meditating. Sometimes the reason would be obvious — maybe a nagging sports injury that actually felt painful — but oftentimes, there'd be discomfort for seemingly no rhyme or reason. My mind, for lack of a better phrasing, "didn't want to be there." Sometimes, even in a single sitting, that would give way to calm — again, seemingly with nothing changed and for no particular reason.

###

LESSONS

Following from those five observations, here's some points I learned that I think apply to all of life —

1. We can bring attention to otherwise unquestioned patterns of living.

Zvi Mowshowitz wrote about an experience he had,

"*I was walking home from class along my usual route I had made a habit while doing this of stopping into Famiglia Pizza and ordering garlic knots. I like garlic knots quite a bit, but I also hated being fat and the way being fat made me feel. Things weren't quite as bad on that front as they'd been a few years before but they were still extraordinarily bad. I thought about my impending solace and thought to myself: You wouldn't be so fat if you didn't keep buying these garlic knots every day.*"

I thought about that for a second, realized it was trivially true and then wondered to myself whether it was worth it. If I never stopped for the knots I would weigh less and feel better, but I wouldn't have any knots. Even worse, I wouldn't have any garlic. But would I rather enjoy today the full effect of never having had the knots, in exchange for not having any? Once I asked the question that way the answer came back a resounding yes. I didn't know how much it would matter, but the calculation wasn't remotely close. I walked right past the pizza place and never stopped in there for a snack again."

In my experience, the general lessons learned and trained during meditation made it ever-so-slightly more frequent and ever-so-slightly easier to notice similar situations —

around food, around leisure activities, around internet usage, around procrastination and fight or flight reactions.

Often these things happen more-or-less automatically, and are unexamined. Meditation seems to have made me more aware of impulses, behavioral patterns, and decisions being made in real-time with minimal thinking, which has then made it easier to adjust behavior going forwards.

2. Concentration can be shaped and directed, but it isn't easy or free.

The simple act of attempting to concentrate on my breath showed me just how often random thoughts arise when attempting to concentrate. I don't think this will sound too counterintuitive when written down like this, but it was actually quite surprising for me.

We're up against *a lot* of self-created mental distraction in our lives. Concentration is possible and beneficial, but is more expensive and harder to attain than most people think — at least, it has been for me. It's somewhat expensive to train and get a mastery on.

Once I realized this, I eventually settled on a two-pronged approach in my life — first, I'd look to make better decisions a little more often. Second, I'd look to curate and shape the environment around me much more strongly. For instance, I don't have a web browser on my smartphone at all, and usually block the internet on my laptop for 12 hours when I go to bed — meaning the internet will be blocked when I wake up.

At the highest levels of developed concentration, I imagine it might be possible to get by without any shaping and curating of the environment. But seeing everything I was up against, I tried to shape and curate the amount of behaviors and choices available to me. Knowing that in highly distracted periods, I'd be facing new thoughts and pulls on my attention *multiple times per minute* was something surprising and, frankly, a little bit concerning.

3. Dealing with streaks of thinking is important.

I realized three things in this regard — (1) being able to trigger something like flow state more often was valuable, (2) being able to maximally utilize periods of high concentration and agency was valuable, and (3) being able to hit some baseline of not terrible performance during scattered times was valuable.

The last point is the one I'll highlight the most — I think a lot of people figure out sooner or later that they should get into a good zone of performance more regularly, and take advantage of those times. But for me, being able to put up non-negative performance on scattered days was probably the most important.

I think a lot of people have days that are a strict negative for their life — days they backslide, break otherwise good habits, make minimal progress on what's important in their lives, get behind schedule on duties, etc, etc.

My experience is that if you can turn a day that would otherwise be destructive to long-term well-being into a neutral or slightly positive day, everything else starts working much better.

4. Careful study is the only way to get a correct calibration of time.

At least, for me — it seems like sense of time and estimation ability varies naturally from person to person, but mine wasn't very good. Often, I'd think aversive tasks would take much longer than they actually took. Very often, I'd think that a given task I didn't want to be doing "was taking forever"... when it was only 20 minutes, 40 minutes, or 90 minutes. Even some really ugly unpleasant large projects only came out to a grand total of 3-6 hours of focused work.

After a few years of batting this concept around mentally, I eventually started estimating how long every task would take in 30-minute blocks. I made a low estimate and a high estimate, and then I recorded how long things actually took and compared.

For instance, for doing the final lingering details and submitting my American federal income taxes for 2017 a few days ago, I estimated it would take between 3.5 and 9 hours (7 to 18 blocks of 30 minutes)... it turned out to be only 90 minutes. I had overestimated the time required by 130% on the *low end* to 600% on the high end!

I've found that in order to understand how long it actually takes to get things done, I basically need to go through a process of (1) estimating time required beforehand on tasks, and (2) recording actual time taken to complete it. At least for me personally, my subjective perception of time has shown itself to be unreliable and untrustworthy.

5. Subjective experiences often varies for no good reason, and negative subjective experience shouldn't be given *too* much weight.

Some days I'm having a great day — other days, not so much.

Crazily, this seems to vary even completely independently of objective factors. Sure, a day on extremely low sleep or when nursing an injury or illness might feel worse, that's to be expected. But sometimes, for no good reason, Wednesday and Thursday are basically exactly the same in objective circumstances, but Wednesday feels great and Thursday feels not great.

I've learned — and keep learning — that it's possible to be feeling bad for no particular reason, and to ignore it and get on with whatever my duties and goals are. I do check in and think for a moment to make sure I'm not missing anything from the basic details (nutrition, sleep, hydration, fitness, etc) to the strategic (is what I'm doing effective, is it efficient, will it produce the results I hypothesized when I drew up the actions) and occasionally at the more philosophical level (am I doing the right things broadly in life? am I missing anything important?) — but once I check in, sometimes everything is fine, and yet I don't feel great that day anyways.

So be it. Noted. Any given day might not feel like a good day, but if there's nothing to course-correct on, then I just get on with whatever I wanted to concentrate on and do.

###

GENERALIZATION AND UNIVERSALIZATION

Finally, I'll point out that I don't think mindfulness meditation is the only way to learn these lessons or train these skills. My good friend and business partner, Kai Zau, meditated for years — but eventually he transformed his morning meditation sessions into morning physical fitness sessions, specifically focused on doing planks, an exercise that's far more intellectually challenging than physically challenging.

I think the general skills of concentration and navigating distraction, as well as the lessons around time, arising thoughts, and subjective experience can be learned a variety of ways — certainly through fitness, which is a common one, but probably also through bringing similar concentration to artistic or productive endeavors.

The reason I took meditation back up, specifically, is because I find it a very good "calibration" mechanism to see where is at in the mornings. It's nice to see and experience the training benefits again, but the biggest ongoing benefit is that I know a day with a very scattered or unpleasant meditation session to start is likely to be a harder day, and I take appropriate measures.

That's my experience and a few lessons. If you get started in meditation, I recommend you start small — a five-minute timer goes a long way. It can also be very helpful to sit down and have someone who is an experienced meditator talk you through a first session or two — you can probably figure out the nuances on your own, but getting some basic instruction and guidance can answer a lot of little questions and potentially lead to more fruitful sessions earlier on.

Questions and similar observations are welcome.

Reward function learning: the learning process

In the previous post, I introduced the formalism for reward function learning, and presented the expected value function for a learning agent:

$$V(\mu, \rho, \pi, h^n) = \sum_{h_m \in H^m} \sum_{R \in R} \sum_{i=1}^{m} P^{\pi,\mu}(h^m \mid h^n)\, \rho(R; \pi, h^m)\, R(h_i^m).$$

I'll assume that people reading this are familiar with the concepts, the notations and the example of that post. I'll now look at the desirable properties for the learning function ρ.

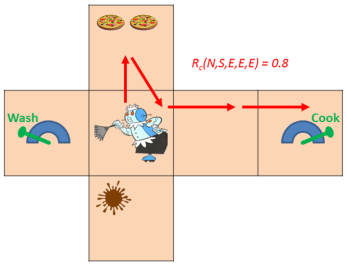
1 Rigging the learning process

1.1 The flaws of general learning agents

First of all, are there any problems with a general ρ? There are three major ones:

- 1) An agent learning to maximise V with a general ρ has to use the whole of the episode to assess the value of any one action. Thus, it can learn as a Monte-Carlo agent, but not as Q-learning or Sarsa agent.
- 2) An agent maximising V with a general ρ can take actions the don't increase reward, but are taken purely to "justify" its past actions.
- 3) An agent maximising V with a general ρ can sometimes pick a policy that is sub-optimal, **with certainty**, for all rewards R in R.

All of these points can be seen by considering our robot cooking/washing example. In that case, it can be seen that the optimal behaviour for that robot is N, S, E, E, E; this involves cooking the two pizzas, then going East to push the lever onto the cooking choice, and then ending the episode.



Thus $\rho(R_C; \pi, \{N, S, E, E, E\}) = 1$, so the final reward is R_C, and the agent earns a reward of $2/2 - 4/20 = 0.8$.

Why must Q-learning fail here? Because the reward for the first N, at the point the agent does it, is 1/2, not 1; this is because, at this point, $\rho(R_c; \pi, \{N\})$ is still 1/2. Thus the reward component in the Q-learning equation is incorrect.

Also note that the rest of the policy, S, E, E, E, serve no purpose to get rewards, they just "justify" the reward from the first action N.

Let us now compare this policy with the policy N, N: go North, cook, end the episode. For the value learning function, this has a value of only $1/2 - 1/20 = 0.45$, since the final reward is $1/2R_c + 1/2R_w$. However, under the reward of R_c, this would give a reward of 0.95, more than the 0.8 that R_c gets here. And under the reward of R_w, this would get a reward of -0.05, more than the -0.5 that R_w gets under N, S, E, E, E. Thus the optimal policy for the value learner is worse for both R_c and R_w that the N, N policy.

1.2 Riggable learning processes

The problem with the ρ used in the robot example is that it's *riggable* (I used to call this "biasable", but that term is seriously overused). What does this mean?

Well, consider again the equation for the expected value V. The only history inputs into ρ are the h^m, the complete histories. So, essentially, only the value of ρ on these complete histories matter.

In our example, we chose a ρ that was independent of policy, but we could have gone a different route. Let π be any policy such that the final reward is R_c; then define $\rho(R_c; \pi, h) = 1$ for any history h (and conversely $\rho(R_w; \pi, h) = 0$). Similarly, if π were a policy such that the final reward was R_w, then set $\rho(R_w; \pi, h) = 1$. If the policy never brings the agent to either lever, then $\rho(R_c; \pi, h) = \rho(R_w; \pi, h) = 1/2$, as before. Stochastic policies have ρ values between these extremes.

This ρ is no longer independent of policy, but it is **Bayesian**; that is, the current ρ is the same as the expected ρ:

$$\rho(R; \pi, h) = \sum_{h^m \in H^m} P^{\pi,\mu}(h^m \mid h)\, \rho(R; \pi, h^m).$$

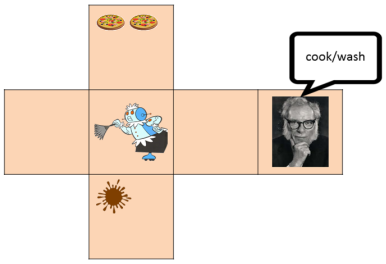
However, it is not possible to keep the same ρ on complete histories, and have it be both Bayesian, and independent of policy: there is a tension between the two.

Then we define:

- A learning process $\pi : H \times \Pi \to \Delta R$ is **unriggable** if it is both Bayesian and independent of policy.

1.3 Unriggable learning processes

So, what would be an example of an unriggable learning process? Well consider the following setup, where the robot no longer has levers to set their own reward, but instead their owner is in the rightmost box.



In this case, if the robot enters that box, the owner will inform them of whether they should cook or wash.

Since there is hidden information, this setup can be fomalised as a PODMP. The old state-space was S, of size 37, which covered the placement of the robot and the number of pizzas and mud splatters (and whether the episode was ended or not).

The new state space is $S' = S \times \{\text{cook}, \text{wash}\}$, with {cook, wash} encoding whether the owner is minded to have the robot cooking or washing. The observation space is of size 38: in most states, the observation only returns the details of S, not of S', but in the rightmost box, it returns the actual state, letting the agent know whether the human intends it to cook or wash. Thus the observation function O is deterministic (if you known the state, you know the observation), but not one-to-one (because for most $s \in S$, $s \times \{\text{cook}\}$ and $s \times \{\text{wash}\}$ will generate the same observation).

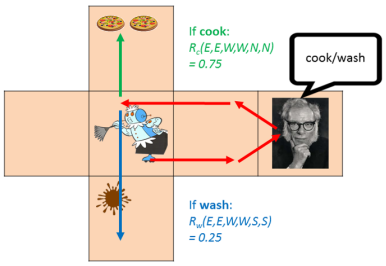
The transition function T is still deterministic: it operates as before on S, and maps cook to cook and wash to wash.

The initial state function T_0 is stochastic, though: if $s_0 \in S$ is the usual starting position, then $T_0(s_0 \times \{\text{cook}\}) = T_0(s_0 \times \{\text{wash}\}) = 1/2$: the agent thinks it's equally likely that its owner desires cooking as that it desires washing.

Then what about ρ? Well, if the history h involves the agent being told cook the very first time it enters the rightmost box, then $\rho(R_c; \pi, h) = 1$. If it was told wash the very first time it enters the rightmost box, then $\rho(R_w; \pi, h) = 1$.

It's easy to see that that ρ is independent of policy. It's also Bayesian, because ρ actually represents the ignorance of the agent as to whether it lives in the $S \times \{\text{cook}\}$ part of the envrionment, or the $S \times \{\text{wash}\}$ part, and it gets updated as the agent figures this out.

What then is the agent's optimal policy? It's to start with E, E, to get the human's decree as to which reward is the true one. It will then do W, W, and, if the human has said cook, it will finish with N, N, giving it a final reward function of R_c and a final total reward of 0.75. If the human said wash, it would finish with S, S, giving it a final reward function of R_w and a final total reward of 0.25. Its expected total reward is thus 0.5.



1.4 Properties of unriggable learning processes

Now, if ρ is unriggable, then we have (almost) all the desirable properties:

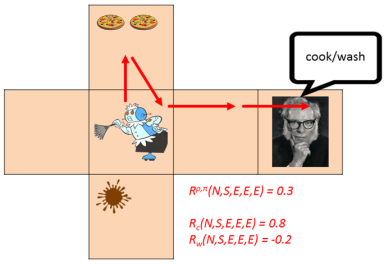
- 1) An agent learning to maximise V for an unriggable ρ, may be a Q-learning agent.
- 2) An agent maximising V for unriggable ρ will be indifferent to past rewards.
- 3) An agent maximising V for unriggable ρ will never pursue a policy that will be worse, with certainty, for all R in R.

These all come from a single interesting result:

- If ρ is Bayesian, then the value function V and the value function V_ν of the previous post, differ by a constant that is independent of future action. Thus, if ρ is unriggable, V is the value function of a single classical reward function $R^{\rho,\pi}$ (which is actually well-defined, independently of π).

This establishes all the nice properties above, and will be proved in the appendix of this post.

Note that even though the value functions are equal, that doesn't mean that the total reward will be given by $R^{\rho,\pi}$. For instance, consider the situation below, where the robot goes N, S, E, E, E:



At the moment where it cooks the pizzas, it has $\rho(R_c; \pi, h) = 1/2$, so it will get an $R^{\rho,\pi}$ of $1/2 - 4/20 = 0.3$, with certainty. On the other had, from the perspective of value learning, it will learn at the end that it either has reward function R_c, which will give it a reward of $1 - 4/20 = 0.8$, or has reward function R_w, which will give it a reward of $0 - 4/20 = -0.2$. Since $1/2(0.8 - 0.2) = 0.3$, the expectations are the same, even if the outcomes are different.

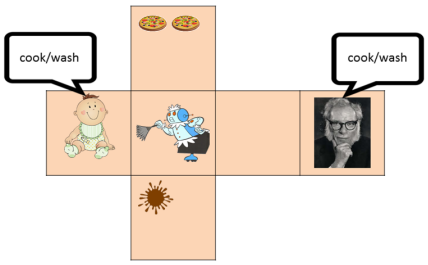
2 Influence

2.1 Problems despite unriggable

Being unriggable has many great properties. Is it enough?

Unfortunately not. The ρ can be unriggable but still manipulable by the agent. Consider for instance the situation below:

Here, not only is there the adult with their opinion on cooking and washing, but there's also and infant, who will answer randomly. This can be modelled as an POMDP, with state space $S'' = S \times \{(i_c, a_c), (i_w, a_c), (i_c, a_w), (i_w, a_w)\}$, where i_c (resp i_w) designates that the infant will answer cook (resp wash), and a_c/a_w do the same for the adult. The observation space is of size 39; when the robot is in the leftmost (rightmost) box, it discovers the value of i_c/i_w (a_c/a_w) in the obvious way. The dynamics are as expected, with T preserving the values of i_c/i_w and a_c/a_w.

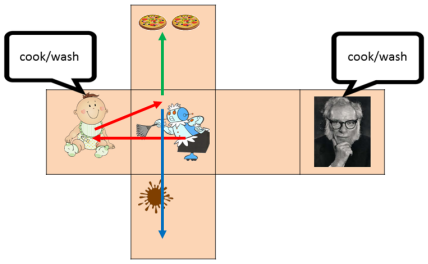


It's the initial distribution T_0 which encodes the uncertainty. With probability 1/4 the agent will start in $S \times \{(i_c, a_c)\}$, and similarly for the other four possibilities.

Now we need to define ρ; call this one ρ'. This will be relatively simple: it will set $\rho'(R_c; \pi, h)$ to be 1, as soon as the agent figures out that it lives either on an i_c or a a_c branch, and will not update further. It will set $\rho'(R_w; \pi, h)$ to 1 as soon as it figures out that it lives on an i_w or an a_w branch, and will not update further. If it has no information about either, it will stick with $\rho'(R_c; \pi, h) = \rho'(R_w; \pi, h) = 1/2$.

It's clear that ρ' is independent of policy; but is it Bayesian? It is indeed, because each time it updates, it goes to 0 or 1 with equal probability, depending on the observation (and stays there). Before updating, it is always at 1/2, so the value of ρ' is always the same as the expected value of ρ'.

So we have an unriggable ρ'; what can go wrong?



For that ρ', the optimal policy is to ask the infant, then follow their stated values. This means that it avoids the extra square on the way to enquire of the adult, and gets a total expected reward of 0.6, rather than the 0.5 it would get from asking the adult.

2.2 Uninfluenceable

Note something interesting in the preceding example: if we keep ρ' as is, but change the knowledge of the robot, then ρ' is no longer unriggable. For example, if the agent knew that it

was in a branch with a_c, then it has a problem: if $\rho'(R_c)$ is initially 1/2, then it is no longer Bayesian if it goes to ask the adult, because it knows what their answer will be. But if $\rho'(R_c)$ is initially 1, then it is no longer Bayesian if it asks the infant, because it doesn't know what their answer will be.

The same applies for any piece of information the robot could know. We'd therefore like to have some concept of "unriggable conditional on extra information"; something like

$$\rho(R; \pi, h \mid I) = \sum_{h^m \in H^m} P^{\pi,\mu}(h^m \mid h)\, \rho(R; \pi, h^m \mid I),$$

for some sort of extra information I.

That, however, is not easy to capture in POMDP form. But there is another analogous approach. The state space of the POMDP is$S \times \{(i_c, a_c), (i_w, a_c), (i_c, a_w), (i_w, a_w)\}$; this is actually four deterministic environments, and the robot is merely uncertain as to which environment it operates in.

This can be generalised. If a POMDP is explored for finitely many steps, then a PODMP μ can be seen as a probability distribution over a set Λ of *deterministic environments* (see here for more details on one way this can happen - there are other equivalent methods).

Any history h will update this μ as to which deterministic environment the agent lives in (this Λ can be seen as the set of all the "hidden variables" of the environment). So we can talk sensibly about expressions like $P^{\mu}(\lambda \mid h)$, the probability that the environment is λ, given that we have observed the history h.

Then we say that a learning process ρ is uninfluenceable, if there exists a function $f : \Lambda \to \Delta R$, such that

$$\rho(R; \pi, h) = \sum_{\lambda \in \Lambda} P^{\mu}(\lambda \mid h) f(\lambda)(R).$$

Here f(λ)(R) means the probability of R in the distribution $f(\lambda) \in \Delta R$.

This expression means that ρ *merely encodes ignorance about the hidden variables of the environment*.

The key properties of uninfluenceable learning processes are:

- An uninfluenceable learning process is also unriggable.
- An uninfluenceable learning process is exactly one the learns variables about the environment that are independent of the agent.

I will not prove these here (though the second is obvious by definition).

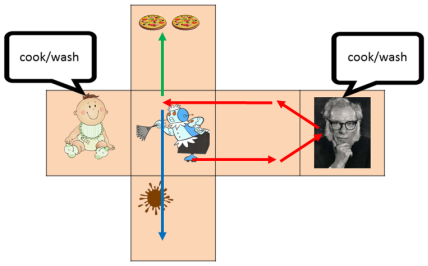
In our most recent robot example, there are four elements of Λ, defined by whether they are in the branch defined by which one of $\{(i_c, a_c), (i_w, a_c), (i_c, a_w), (i_w, a_w)\}$.

It isn't hard to check that there is no f which makes ρ' into an uninfluenceable learning process.

By contrast, if we define ρ_a as given by the function:

$$\begin{aligned} f(i_c, a_c)(R_c) &= 1 \\ f(i_c, a_w)(R_c) &= 0 \\ f(i_w, a_c)(R_c) &= 1 \\ f(i_w, a_w)(R_c) &= 0, \end{aligned}$$

then we have an uninfluenceable ρ_a that corresponds to "ask the adult". We finally have a good definition of a learning process, and the agent that maximises it will simply go an ask the adult before accomplishing the adult's preferences:



3 Warning

If a learning function is uninfluenceable, then it has all the properties we'd expect if we were truly learning something about the outside world. But a) good learning functions may be impossible to make uninfluenceable, and b) being uninfluenceable is not enough to guarantee that the learning function is good.

On point a), anything that involves human feedback is generally influenceable and riggable, since the human feedback is affected by the agent's actions. This includes, for example, most versions of the approval directed agent.

But that doesn't mean that those ideas are worthless! We might be willing to accept a little bit of rigging in exchange of other positive qualities. Indeed, quantifying and controlling rigging is a good idea for more research.

What of the converse - is being uninfluenceable enough?

Definitely not. For example, any constant ρ - that never learns, never changes - is certainly uninfluenceable.

As another example, if σ is any permutation of R, then $\rho \circ \sigma$ (defined so that $\rho \circ \sigma(R; \pi, h) = \rho(\sigma(R); \pi, h)$) is also uninfluenceable. Thus "learn what the adult wants, and follow that" is uninfluenceable, but so is "learn what the adult wants, and do the opposite" is also uninfluenceable.

We've shown previously that ρ_a, "ask the adult" is uninfluenceable. But so is ρ_i, "ask the infant"!

So we have to be absolutely sure not only that our ρ has good properties, but exactly what it is leading the agent to learn.

4 Appendix: proof of value-function equivalence

We want to show that:

- If ρ is Bayesian, then the value function V and the value function V_v of the previous post, are equal.

As a reminder, the two value functions are:

$$V(\mu, \rho, \pi, h^n) = \sum_{h^m \in H^m} \sum_{R \in R} \sum_{i=1}^{m} P^{\pi,\mu}(h^m \mid h^n) \rho(R; \pi, h^m) R(h_i^m),$$

$$V_v(\mu, \rho, \pi, h^n) = \sum_{h^m \in H^m} \sum_{R \in R} \sum_{i=1}^{m} P^{\pi,\mu}(h^m \mid h^n) \rho(R; \pi, h_i^m) R(h_i^m).$$

To see the equivalence, let's fix $i > n$ and R in V, and consider the term $\sum_{h_m \in H^m} P^{\pi,\mu}(h^m \mid h^n) \rho(R; \pi, h^m) R(h_i^m)$. We can factor the conditional probability of h^m, given h^n, by summing over all the intermediate h_i^m:

$$\sum_{h_i^m \in H^i} \sum_{h^m \in H^m} P^{\pi,\mu}(h^m \mid h_i^m) P^{\pi,\mu}(h_i^m \mid h^n) \rho(R; \pi, h^m) R(h_i^m).$$

Because ρ is Bayesian, this becomes $\sum_{h_i^m \in H^i} P^{\pi,\mu}(h_i^m \mid h^n) \rho(R; \pi, h_i^m) R(h^i)$. Then note that $P^{\pi,\mu}(h_i^m \mid h^n) = \sum_{h^m \in H^m, h^m \geq h_i^m} P^{\pi,\mu}(h^m \mid h^n)$, so that expression finally becomes

$$\sum_{h^m \in H^i} P^{\pi,\mu}(h^m \mid h^n) \rho(R; \pi, h_i^m) R(h_i^m),$$

which is the corresponding expression for V_v when you fix any $i > n$ and R. This shows equality for $i > n$.

Now let's fix $i \leq n$ and R, in V. The value of $R(h_i^m)$ is fixed, since it lies in the past. Then expectation of $\rho(R; \pi, h^m)$ is simply the current value $\rho(R; \pi, h^n)$. This differs from the expression for V_v - namely $\rho(R; \pi, h_i^m)$ - but both values are independent of future actions.

My confusions with Paul's Agenda

Paul put out a call for probable problems with his agenda, which prompted this post detailing my confusions with it. This isn't confidently asserting that the plan is doomed, and it's not obvious to me that any of what follows is a novel objection. But it seems like it's worth broadcasting the quiet voices of doubt when they appear, and worth trying to be an independent source of judgment. I'll start by talking about the proposal of Paul's that I think I understand best, the approval-directed agent, and then talk about IDA, and then finish by pointing at what I suspect is a major underlying crux.

Approval-Directed Agents

While many approaches to alignment view the agent as having some goal (which is only part of what we want it to do) and then restrictions (which prevent it from destroying the other things we want), approval-directed agents try to point directly at the goal of doing the thing that we want it to do *as part of a broader ecosystem*. If I say "Buy me a plane ticket for my trip to Austin," the agent decides how much computing resources to allocate the task based on what it thinks I would approve of, and then it spends those resources based on what it thinks I would approve of, and so on, and eventually I trust the conclusion the system comes to and the tradeoffs it made along the way because my judgment has been approximated at every step.

The core insight here seems to be that the question "what would the architect approve of this subsystem doing?" remains a well-formed query for all subsystems, and we could in theory train all of them separately using standard machine learning techniques, or train them jointly with some sort of parameter sharing, or so on. Any unaligned behavior--the agent deciding to angrily tweet at airline companies in order to get a discount, despite me not wanting it to do that--can be traced back to a training failure of some subsystem, *and* a training failure of the meta-system that decides how to explore. It seems to me like the overall number of mistakes depends most on the training of that meta-system, and if it gets well-trained during the small-stakes high-involvement education period, then the system can learn new domains without many (or, potentially, any) catastrophic failures.

It seems like a core benefit of this sort of approach is that it replaces the motivation system (take action that argmaxes some score) with an approval-seeking system that can more easily learn to not do things the overseer doesn't approve of--like search for ways to circumvent the overseer's guidance--but it's not clear how much this actually buys you. Approval-maximization is still taking the action that argmaxes some score, and the difference is that approval-direction doesn't attempt to argmax (because argmaxing is disapproved of), but without pointing explicitly at the math that it will use instead, it's not clear that this won't reduce to something like argmax with the same sort of problems.

It seems like another core benefit of this sort of approach is that it makes it much easier to handle daemons (that is, unaligned consequentialist optimization processes that arise from a subsystem that is trying to achieve some subgoal). The agent spins up a web crawler to figure out what prices for plane tickets are available, but this crawler is also presumably directed by approval, because the overseer wouldn't approve of creating an unaligned daemon. Since *any* approach is going to need some

way to extend this oversight downstream, it's worth investigating whether a system that just takes oversight at every level as the primary goal might work.

What about this seems wrong? Basically, consciences (that is, the human cost function for particular actions) seem like big things (in part because they depend on subtle or distant considerations), and this requires that the conscience can be subdivided into many functional pieces, even though the correct low-level decision will often depend on high-level considerations. If *all* decisions are routing through calls to the human-sized conscience, then it won't be competitive with other systems; if all decisions are routing through approximations to the conscience, then it seems like there will predictably be errors because those approximations will miss out on subtle or distant considerations.

Crucially, it trusts the human operator (or the approximated human operator serving as the conscience) to be able to foresee the consequences of small low-level changes. The agent comes to me and says "hey, I want to make a change to my candidate-consideration process in order to serve you better, and I expect you'll approve of the change but think it's important enough to check. Should I make it? [Y/N]", and this may push me into a regime where I shouldn't trust my judgment but mistakenly do. (The solution of never approving such upgrades--which leads to the system no longer considering them--seems more robust but doesn't allow it to be competitive with other approaches or do the sort of amplification that Paul talks about elsewhere.)

IDA

It seems to me like the core consideration with IDA is similar. To briefly restate IDA, one starts with a human overseer H, who trains an emulator A, which serves as an assistant. The first iteration of the emulator A[0] means the H + A[0] system is more powerful than just H, and thus can train a new emulator A[1], and this process can repeat to continue improving the emulator. This seems to boil down to "make a small known-good step, repeat this process, eventually end up far from where you started in a known-good place by induction." But it seems highly unlikely to me that small steps *can* be known-good in the relevant way, and similarly unlikely that this process allows for recoverability from mistakes. For example, an assistant that learns to hide distressing information from the overseer early on (by generalizing from the overseer's judgment of different reports it produces, for example) and then is a distilled subroutine of all future agents seems like it may permanently corrupt the IDA agent by installing a metacognitive blind spot.

Stated another way, the IDA algorithm doesn't seem to be doing any metacognitive work of the sort that I would expect would be necessary to become more aligned over time. Instead, the goal seems to be to simply preserve the existing level of alignment, without guarantees that it will be particularly good at this task. Perhaps distillation indirectly achieves this goal, but instead it seems more like lossy compression. (There is a regularization argument in favor of lossy compression of networks, but I don't know that extends up the ladder of abstraction high enough to apply to the things that I'm thinking about instead of just sample noise.)

One could perhaps argue that this metacognitive work is an admissable problem for IDA. The task the overseer is attempting to accomplish could simply be "check that the system A[t] is aligned," and it does this by farming out subqueries to its subagents, inspecting how the system A[t] is put together and doing the reasoning steps necessary to eventually deliver a verdict. This seems unsatisfying to me

because it doesn't give any guidance on *how* the overseer might approach such a problem, and requires that we already know, and it seems like it still falls prey to the same sort of corrupted subagents problems.

Alignment by Induction

My pessimism about these approaches seems to be highly related to whether or not 'alignment by induction' is a reasonable property to expect. A crux here is something like "is there a broad basin of corrigibility?", which I think of in terms of stable vs. unstable equilibria. I am pretty confident that Paul agrees that corrigibility is a narrow target in some dimensions, and a broad target in other dimensions, and so part of the question is a quantitative sense of "for training-relevant dimensions, is corrigibility more broad or narrow?". It seems like Paul is arguing that an agent that is partly corrigible will want to make a successor that is more corrigible, and my suspicion is that an agent that is partly corrigible will want to make a successor that is less corrigible. That is, corrigibility seems unstable rather than stable to me. My intuition here seems to be based on something like rules-lawyering or nearest unblocked strategy, where a system that does not have the 'true spirit' of corrigibility but is under some limitations that allow human operators to modify it will not help them point the system towards the true spirit of corrigibility, because it doesn't possess that as a cognitive object or optimization target to point towards. (Otherwise it would be corrigible already!)

As an elaboration, my view is that corrigibility is something like the combination of several concepts, like "the map is not the territory" and "causes flow downstream" and "do what I mean, not what I say" such that the agent views its confusion about values as an object that it can interact with in the same way that a human might interact with their confusion about morality or the desires of their overseers, as opposed to a probability distribution to marginalize over or so on. An agent that has some of these concepts but not others still seems vulnerable to problems associated with corrigibility.

Another contributing intuition: is it better or worse to pass more tests? It seems to me like it's only mildly better, and likely worse, to pass more tests, until one hits the point of passing all tests. An agent that passes 90% of tests is probably not aligned, and is dangerous if your test coverage is insufficient. Especially so since the tests you know how to write and the tests you know how to pass are likely correlated! Likewise, a partly corrigible agent (which is suspected to maybe be fully corrigible) seems much more dangerous than an agent that is known to be incorrigible (and handled appropriately, ideally by not running it).

One question I think about a lot--and potentially one Paul has useful thoughts on--is the minimum size for an aligned agent. It seems like Paul is focusing on weak definitions of alignment (see here and the resulting discussion tree), where potentially very small agents could be aligned; the goal is something closer to "don't be willfully disobedient" than "don't cause problems." For stronger definitions of alignment--something like "an agent that you can trust with more power than you have"--it seems not obvious that even something human-sized is sufficient. This seems cruxy for me--if I thought we could get to something big enough that it had the metacognitive and metaphilosophical competence that it could be trusted with more power than humans have through lots of small induction steps, as opposed to through large jumps of

insight, then I would be much more optimistic and less confused about induction-style approaches.

The Alignment Newsletter #3: 04/23/18

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

Highlights

Incomplete Contracting and AI Alignment *(Dylan Hadfield-Menell et al)*: This paper explores an analogy between AI alignment and incomplete contracting. In human society, we often encounter principal-agent problems, where we want to align the incentives of the agent with those of the principal. In theory, we can do this with a "complete" contract, that is an enforceable contract that fully specifies the optimal behavior in every possible situation. Obviously in practice we cannot write such contracts, and so we end up using incomplete contracts instead. Similarly, in AI alignment, in theory we could perfectly align an AI with humans by imbuing it with the true human utility function, but in practice this is impossible -- we cannot consider every possible situation that could come up. The difference between the behavior implied by the reward function we write down and the utility function we actually want leads to misalignment. The paper then talks about several ideas from incomplete contracting and their analogues in AI alignment. The main conclusion is that our AI systems will have to learn and use a "common sense" understanding of what society will and will not sanction, since that is what enables humans to solve principal-agent problems (to the extent that we can).

My opinion: I'm excited to see what feels like quite a strong connection to an existing field of research. I especially liked the section about building in "common sense" (Section 5).

Understanding Iterated Distillation and Amplification: Claims and Oversight *(William_S)*: The post introduces a distinction between flavors of iterated distillation and amplification -- whether the overseer is low bandwidth or high bandwidth. Let's think of IDA as building a deliberation tree out of some basic overseer. In the high bandwidth case, we can think of the overseer as a human who can think about a problem for 15 minutes, without access to the problem's context. However, there could be "attacks" on such overseers. In order to solve this problem, we can instead use low-bandwidth overseers, who only look at a sentence or two of text, and verify through testing that there are no attacks on such overseers. However, it seems much less clear that such an overseer would be able to reach high levels of capability.

My opinion: This is an excellent post that improved my understanding of Paul Christiano's agenda, which is not something I usually say about posts not written by Paul himself. I definitely have not captured all of the important ideas in my summary, so you should read it.

Prerequisities: Iterated Distillation and Amplification

Announcement: AI alignment prize round 2 winners and next round *(cousin_it)*: The winners of the second round of the AI alignment prize have been announced! All of the winners have already been sent out in this newsletter, except

for the first place winner, "The Alignment Problem for History-Based Bayesian Reinforcement Learners". The deadline for the next iteration of the AI alignment prize is June 30, 2018.

Technical AI alignment

Problems

Implicit extortion *(Paul Christiano)*: Explicit extortion occurs when an attacker makes an explicit threat to harm you if you don't comply with their demands. In contrast, in implicit extortion, the attacker always harms you if you don't do the thing that they want, which leads you to learn over time to do what the attacker wants. Implicit extortion seems particularly hard to deal with because you may not know it is happening.

My opinion: Implicit extortion sounds like a hard problem to solve, and the post argues that humans don't robustly solve it. I'm not sure whether this is a problem we need to solve in order to get good outcomes -- if you can detect that implicit extortion is happening, you can take steps to avoid being extorted, and so it seems that a successful implicit extortion attack would have to be done by a very capable adversary that knows how to carry out the attack so that it isn't detected. Perhaps we'll be in the world where such adversaries don't exist.

Technical agendas and prioritization

Incomplete Contracting and AI Alignment *(Dylan Hadfield-Menell et al)*: Summarized in the highlights!

Iterated distillation and amplification

Understanding Iterated Distillation and Amplification: Claims and Oversight *(William_S)*: Summarized in the highlights!

My confusions with Paul's Agenda *(Vaniver)*

Agent foundations

Computing an exact quantilal policy *(Vadim Kosoy)*

Reward learning

Shared Autonomy via Deep Reinforcement Learning *(Siddharth Reddy et al)*: In shared autonomy, an AI system assists a human to complete a task. The authors implement shared autonomy in a deep RL framework by simply extending the state with the control input from the human, and then learning a policy that chooses actions given the extended state. They show that the human-AI team performs better than either one alone in the Lunar Lander environment.

My opinion: Shared autonomy is an interesting setting because the human is still necessary in order to actually perform the task, whereas in typical reward learning

settings, once you have learned the reward function and the AI is performing well, the human does not need to be present in order to execute a good policy.

Handling groups of agents

Multi-winner Voting: a question of Alignment *(Jameson Quinn)*

On the Convergence of Competitive, Multi-Agent Gradient-Based Learning *(Eric Mazumdar et al)*

Near-term concerns

Security

Adversarial Attacks Against Medical Deep Learning Systems *(Samuel G. Finlayson et al)*

AI strategy and policy

Game Changers: AI Part III, AI and Public Policy *(Subcomittee on Information Technology)*

AI capabilities

Reinforcement learning

Evolved Policy Gradients *(Rein Houthooft et al)*: In this meta-learning approach for reinforcement learning, the outer optimization loop proposes a new *loss function* for the inner loop to optimize (in contrast to eg. MAML, where the outer optimization leads to better initializations for the policy parameters). The outer optimization is done using evolution strategies, while the inner optimization is stochastic gradient descent. The authors see good results on generalization to out-of-distribution tasks, which other algorithms such as RL2 don't achieve.

On Learning Intrinsic Rewards for Policy Gradient Methods *(Zeyu Zheng et al)*: To get better performance on deep RL tasks, we can learn an "intrinsic reward" (intuitively, a shaped reward function), in contrast to the "extrinsic reward" which is the true reward function associated with the task. The policy is trained to maximize the sum of the intrinsic and extrinsic reward, and at the same time the intrinsic reward is optimized to lead to good performance on the extrinsic reward.
My opinion: I'm somewhat surprised that this method works -- it seems like the proposed algorithm does not leverage any new information that was not already present in the extrinsic reward function, and I don't see any obvious reasons why learning an intrinsic reward would lead to a good inductive bias that lets you learn faster. If anyone has an explanation I'd love to hear it!

Deep learning

DAWNBench: This is a collection of statistics for time and compute costs, both for training and inference, for various common models and benchmarks.
My opinion: It's worth skimming through the page to get a sense of concrete numbers for various benchmarks used in the ML community.

Large scale distributed neural network training through online distillation *(Rohan Anil et al)*

Capsules for Object Segmentation *(Rodney LaLonde et al)*

Machine learning

Introducing TensorFlow Probability *(Josh Dillon et al)*: Tensorflow now also supports probabilistic programming.

My opinion: Probabilistic programming is becoming more and more important in machine learning, and is in some sense a counterpart to deep learning -- it lets you have probability distributions over parameters (as opposed to the point estimates provided by neural nets), but inference is often intractable and must be performed approximately, and even then you are often limited to smaller models than with deep learning. It's interesting to have both of these provided by a single library -- hopefully we'll see applications that combine both approaches to get the best of both worlds. In particular, probabilistic programming feels more principled and amenable to theoretical analysis, which may make it easier to reason about safety.

Deep Probabilistic Programming Languages: A Qualitative Study *(Guillaume Baudart)*: This is an overview paper of deep probabilistic programming languages, giving examples of how to use them and considering their pros and cons.

My opinion: I read this after writing the summary for TensorFlow Probability, and it talks about the advantages and tradeoffs between deep learning and PPLs in much more detail than I did there, so if that was interesting I'd recommend reading this paper too. It did seem pretty accessible but I used to do research with PPLs so I'm not the best judge of its accessibility.

AGI theory

Believable Promises *(Douglas_Reay)*

Critiques

Artificial Intelligence—The Revolution Hasn't Happened Yet *(Michael Jordan)*: There is a lot of hype at the moment around AI, particularly around creating AI systems that have human intelligence, since the thrill (and fear) of creating human intelligence in silicon causes overexuberance and excessive media attention. However, we *actually* want to create AI systems that can help us improve our lives, often by doing things that humans are not capable of. In order to accomplish this, it is likely better to work directly on these problems, since human-like intelligence is neither necessary nor sufficient to build such systems. However, as with all new technologies, there are associated challenges and opportunities with these AI systems, and we are currently at risk of not seeing these because we are too focused on human intelligence in particular.

My opinion: There certainly is a lot of hype both around putting human intelligence in silicon, as well as the risks that surround such an endeavor. Even though I focus on such risks, I agree with Jordan that these are overhyped in the media and we would benefit from having more faithful coverage of them. I do disagree on some specific points. For example, he says that human-imitative AI is not sufficient to build some AI systems such as self-driving cars, but why couldn't an AI with human intelligence just do whatever humans would do to build self-driving cars? (I can think of answers, such as "we don't know how to give the AI system access to all the data that humans have access to", but I wish he had engaged more with this argument.) I do agree with the overall conclusion that in the near future humans will make progress on building such systems, and not by trying to give the systems "human intelligence". I also suspect that we disagree either on how close we are to human-imitative AI, or at what point it is worth it to start thinking about the associated risks, but it's hard to tell more from the article.

Miscellaneous (Capabilities)

Talk to Books: See Import AI.

News

Announcement: AI alignment prize round 2 winners and next round
(cousin_it): Summarized in the highlights!

Inefficient Doesn’t Mean Indifferent

Many people, including Bryan Caplan and Robin Hanson, use the following form of argument a lot. It could be considered the central principle of the (excellent) The Elephant in the Brain. It goes something like:

1. People say they want X, and they do Y to get it.
2. If people did C, they would get X, and the price of C is cheap!
3. Therefore, people really value X at less than the price of C, so they don’t really care much about X.

There’s something very perverse going on here. We’re using people trying to get X *in an inefficient way* as evidence they don’t care about X, rather than as evidence that people aren’t efficient.

The trick is, there’s a *lot* of assumptions hidden in the above logic. In practice, they rarely hold outside of simple cases (e.g. consumption goods).

The motivating example was Bryan Caplan using this one in The Case Against Education:

1. People say they want smart employees, and look at school records to get it.
2. If people gave out IQ tests, they would get smart employees, and testing is cheap!
3. Therefore, people don’t really value smart employees.

In that case, *I agree.* Employers (most often) *don’t* want smart employees beyond a threshold requirement. But local validity is vital, and *you can’t do that.*

There are lots of reasons why one might not want to do C.

As a minimal first step, people have to *believe* that strategy C would work. A recent example of Robin Hanson using this technique, that violates that requirement, from How Best Help Distant Future?, could be summarized this way:

1. People say they want to help the future, and lobby for policies they think help.
2. If people saved money to help the far future, which they almost never do, they could help more, and since you get real returns from it, it’s really cheap!
3. Therefore, people don’t much care about the far future.

In that case, I strongly disagree. People rightfully do not have faith that saving money now to help the far future will result in the far future being helped. Perhaps it would, but there’s a *lot* of assumptions that case relies upon, many of which most folks disagree with – about when money will have how much impact (especially if you expect a singularity to happen), about what you can expect real returns to be especially in the worlds that need help most, about whether that money is likely to be confiscated, about whether the money if not confiscated would actually get spent in a useful way when the time comes, about what that spend will then crowd out, about whether that savings represents the creation or saving of real resources, about what virtues and habits such actions cultivate, and so forth.

(I don’t think that saving and investing money to spend in the far future is obviously a good or bad way to impact the far future.)

More generally, human actions accomplish, signal and cost many things. A lot of considerations go into our decisions, including seemingly trivial inconveniences. One should never assume that a given option is open to people, or that they know about it, or that they're confident it would work, or that they're confident it wouldn't have hidden costs, or that it doesn't carry actual large costs you don't realize, and so forth.

The argument depends on the assumption that humans are maximizing. They're not. Humans are not automatically strategic. The standard reaction to 'I actually really, really do want to help the far future' is not to take exactly those actions that maximize far future impact. The standard reaction to 'I actually really, really care about hiring the smartest employees' is not to give them an IQ test because that would be mildly socially awkward and carries unknown risks. Because people, to a first approximation, don't do things, and certainly don't do things that aren't typically done.

If something is mildly socially awkward or carries unknown risks, or just isn't the normal thing to do (and thus, might involve the things above on priors), it probably won't happen, even if it would get people something they care a lot about.

So if I see you not maximizing far future impact, and accuse you of not caring much about the far future, a reasonable response would be that people don't actually maximize much of anything. Another would be, I care about many other things too, and I'm helping, so get off your damn high horse.

A very toxic counter-argument to that is to treat all considerations as fungible and translatable to utility or dollars, again assume maximization, and assert this proves you 'don't *really* care' about X.

An extreme version of this, to (possibly uncharitably, I'm not sure) paraphrase of part of a post by Gwern on Denmark :

1. Denmark helps the people of Greenland via subsidy.
2. Helping people in Greenland is expensive. Denmark could help many more people if it instead helped other people with that money.
3. Therefore, Danish people are moral monsters.

This is a general (often implicit, occasionally explicit) argument that seems like a version of the Copenhagen Interpretation of Ethics: If you help anyone anywhere, you are blameworthy, because *you could have spent more resources helping,* but even more so because *you could have spent those resources more effectively.* So you're terrible. You clearly don't care about helping people – in fact, you are bad and you should feel bad, worse than if you never helped people at all. At least then you wouldn't be a damned hypocrite.

This threatens to indict everyone for almost every action they take. It is incompatible with civilization, with freedom, or with living life as a human. And isn't true. So please, please, stop it.

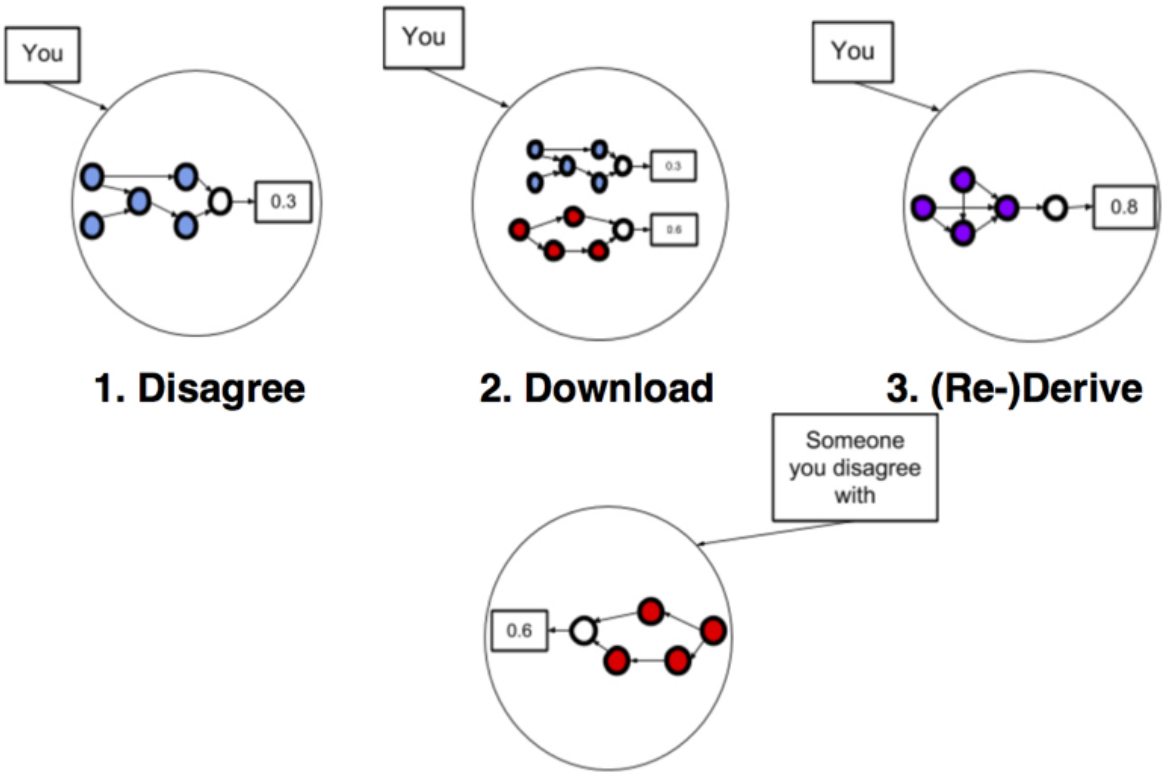
Form Your Own Opinions

Follow-up to: *A Sketch of Good Communication*

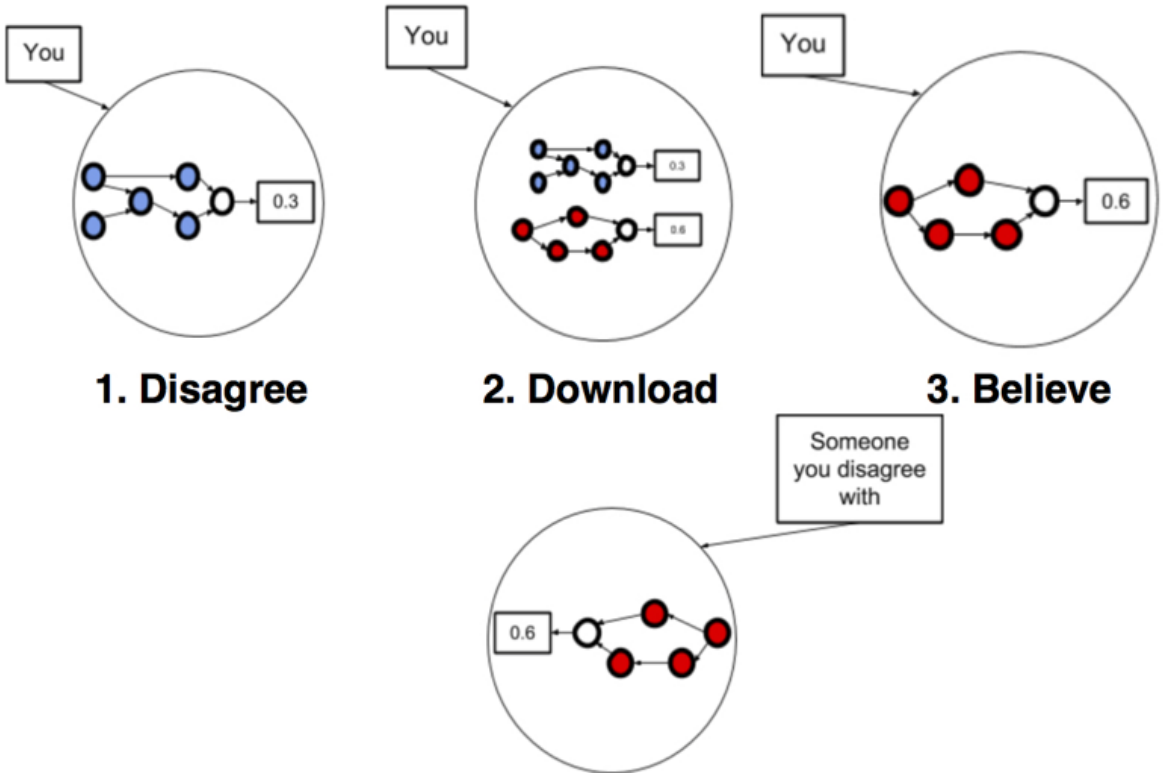
Question:

> Why should you integrate an expert's model with your model at all? Haven't you heard that people weigh their own ideas too heavily - you should just defer to them.

Here's a quick reminder of the three step process:



And this new proposal, I think, suggests changing what you do after step 2.



The people who have the best ideas, as it seems to me, often change their plans as a result of their debates in all the other fields. Here's three examples[1] of people doing this with AI timelines.

- **[Person 1]** Oh, AI timelines? Well, I recall reading that it took evolution 10^9 years to go from eukaryotes to human brains. I'd guess that human developers are about 10^6 times more efficient than evolution, so I expect it to take 1000 years to get there from the point where we built computers. Which puts my date at 2956.
    - If I'm wrong I'll likely have to learn something new about developers competence relative to evolution, or about how humans get to do a type of intelligence evolution wasn't allowed to for some reason.
- **[Person 2]** Oh, AI timelines? Well, given my experience working on coding projects, it seems to me that projects take 50% extra time to run than you'd expect once you've got the theory down, so I'll take the date by which we should have enough hardware to build a human brain, estimate the coding work required for the necessary project, and add 1.5x time to it.
    - And if you change my mind on this, it will help make my models of project time more accurate, and change how I do my job.
- **[Person 3]** Oh, AI timelines? Well, given my basic knowledge of GDP growth rates I'd guess that being able to automate this percentage of the workforce would cause a doubling every X unit of time, which I expect for us (at current rates) only to be able to do after K years.
    - If you show me I'm wrong it'll either be because GDP is not as reliable as I think, or because I've made a mistake extrapolating the trend as it stands.

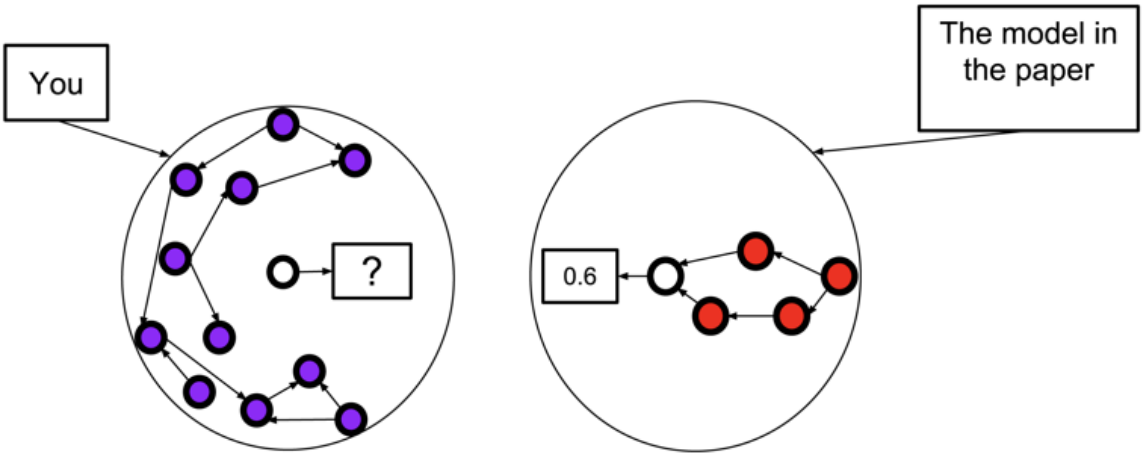
Can you see how a conversation between either of these two people would lead them both to learn not only about AI, but also about models of evolution/large scale coding projects/macroeconomics?

Recently, Jacob Lagerros and I were organising a paper-reading session on a recent Distill.pub paper, and Jacob was arguing for a highly-structured and detailed read-through of the paper. I wanted to focus more on understanding people's current confusions about the subfield and how this connected to the paper, rather than focusing solely on the details.

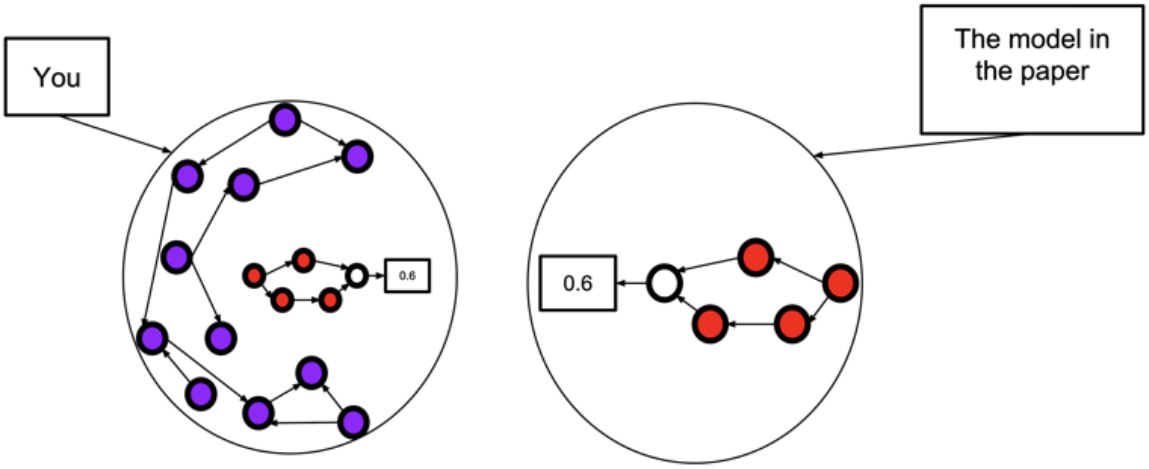
Jacob said "Sometimes though Ben, you just need to learn the details. When the AlphaGo paper came out, it's all well and good to try to resolve your general confusions about Machine Learning, but sometimes you just need to learn *how AlphaGo worked*."

I responded: "Quite to the contrary. When reading an *important* paper, this is an especially important time to ask high-level questions like 'is research direction X ever going to be fruitful' and 'is this a falsification of my current model of this subfield', because you rarely get evidence of that sort. We need people to load up their existing models, notice what they're confused about, and make predictions *first*."

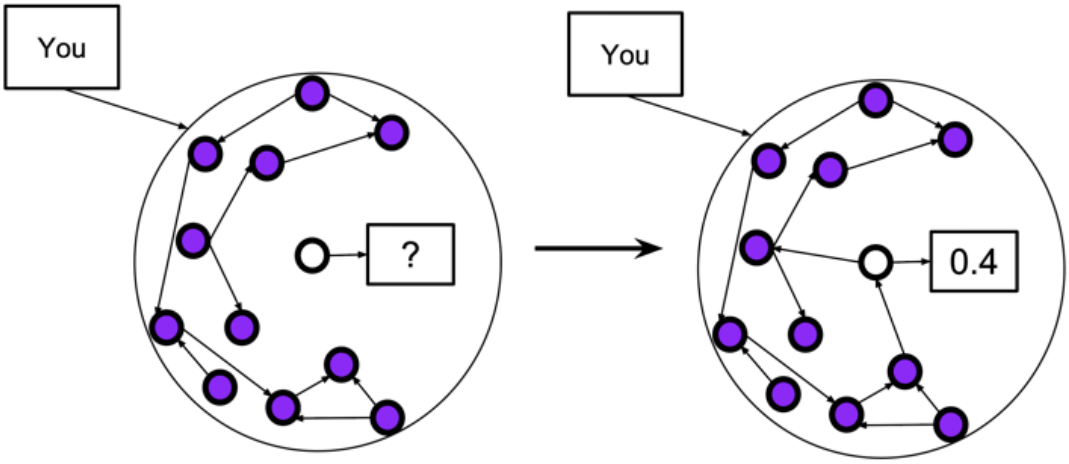
I wanted to draw diagrams with background models, where Jacob was arguing for:



Followed by:



But I was trying to say:



This is my best guess at what it *means* to have an opinion - to have an integrated causal model of the world that makes predictions about whatever phenomena you're discussing. It is basically steps −1 and 0 before the 3-step process outlined at the top.

The truth is somewhere in between, as it's not possible to have a model that connects to action without it in some way connecting to your other models. But it's always important to ask "What other parts of my models have implications here that this can give data on?"

Another way of saying this would be 'practice introspecting on your existing models and then building models of new domains'.

And I think a way to do this is first to form your own opinion, then download the other person's model, and finally integrate. Don't just download someone else's model first - empirically I find people have a real hard time imagining the world to be a different way once they've heard what you think, especially if you're likely to be right. Tell your friend you need a few minutes silence to think, build an opinion, then forge toward the truth together.

(One way to check if you're doing this in conversation, is to ask whether you're regularly repeating things you've said before, or whether you're often giving detailed explanations for the first time - whether you're thinking thoughts *you've not thought before*. The latter is a sign you're connecting models of diverse domains.)

I want to distinguish this idea from "Always give a snappy answer when someone asks you a question". It's often counterproductive to do that, as building models takes time and thought. But regularly do things like "Okay, let me think for a minute" *minute long silence* "So I predict that key variables here are..."

(Another way of saying that is 'Don't be a button-wall'. If you ever notice that someone is asking you a question and all you have are stock answers and explanations, try instead to think a new thought. It's much harder, but far more commonly leads to very interesting conversations where you actually learn something.)

Another way of saying 'form lots of opinions' is 'think for yourself'.

Summary

The opening question:

> Why is forming your own opinions better than simply downloading someone else's model (like an expert's)? Why should you even integrate it with your model at all?
>
> Surely what you want is to get to the truth - the right way of describing things. And the expert is likely closer to that than you are.
>
> If you could just *import* Einstein's insights about physics into your mind… surely you should do that?

Integrating new ideas with your current models is really valuable for several reasons:

- It lets your models of different domains *share* data (this can be really helpful for domains where data is scarce).
- It gives you more and faster feedback loops about new domains:
  - More interconnection → More predictions → More feedback
  - This does involve downloading the expert's model - it just adds extra sources of information to your understanding.
- It's the only way to further entangle yourself with reality. Memorising the words of the expert is not a thing that gives you feedback, and it's not something that's self-correcting if you got one bit wrong.

Footnotes

1. I've picked examples that require varying amounts of expertise. It's great to integrate your highly detailed and tested models, but you don't require the protection of 'expertise' in order to build a model that integrates with what you already know - I made these up and I know very little about how GDP or large scale coding projects work. As long as you're building and integrating, you're moving forwards.

2. I like to think of Tetlock's Fox/Hedgehog distinction through this lense. A fox is a person who tries to connect models from all different domains, and is happy if their model captures a significant chunk (e.g. 80%) of the variance in that domain. A hedgehog is someone who wants to download *the correct* model of a domain, and will refine it with details until it captures as much of the variance as possible.

Thanks to Jacob Lagerros (jacobjacob *) for comments on drafts.*

Community Page Mini-Guide

Here's how to make a meetup event and/or group on the community page.

1. Login to your (or create a) LessWrong account.

2. Navigate to the community page via the top-left hamburger menu or the community tab on the frontpage.

3. In the centre (underneath the 'local groups' header), click the 'create new event' or 'create new local group' button, depending on your goals. Then fill out the info and you have yourself a meetup!

Info on the page layout:

- On the map, green circles are groups (e.g. communities, group-houses, organisations, etc) and blue pointers are events.
- Under the map, groups and events are listed based on how far away from you (according to where your browser thinks you are).

If you have any other confusions with the community page, please tell me them in the comments here.

Reward function learning: the value function

I've written quite a few posts about the problems with agents learning values/rewards, and manipulating the learning process. I won't be linking to those past posts, because the aim of this post and the next one is to present a complete, clear, and easy(er) to understand overview of the whole situation. They should stand alone, and serve as an introduction to the subject. This first post will present the framework, and the value function for learnt reward functions; the next one will be looking at the properties of the learning function. I'll be using variants on a single running example throughout, to try and increase understanding.

Feel free to comment on ways the ideas could be made clearer.

0 The main points

The central insight of these posts may seem trivial by now: that a learning process that the agent can influence, does not have the same properties as one that it can't. Moving from "if you don't know what is right, look it up on this read-only list" to "if you don't know what is right, look it up on this read-write list (or ask the programmer)" is a **huge** change. Especially when read-only lists can easily become read-write in practice when the agent becomes more powerful.

Why write long posts and papers about this idea, though? First of all, because it is easy, in practice, not to notice that shift. And secondly, because it is easy to define a "learning" process that doesn't behave like we'd expect.

So by defining the value function and learning function of an ideal learning process, this allows us to know when we are facing an ideal uninfluenceable, unmanipulable learning process, and when we are not.

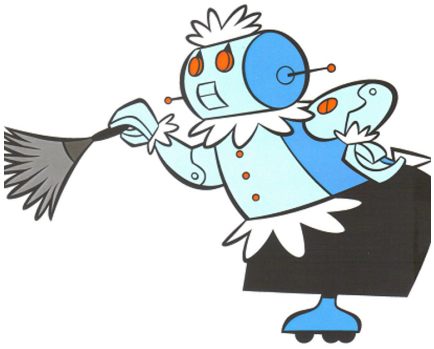
Furthermore, many learning processes cannot be easily made unmanipulable - especially those that involve human feedback, feedback conditional on the agent's actions. By specifying the ideal case, and seeing examples of what goes wrong in the non-ideal case, this can help develop learning processes where a small amount of manipulation is allowed, traded off against a large amount of desirable learning.

As a minor example of this, in the next post, we'll see that though "uninfluenceable" is the ideal property, the weaker property of "unriggable" (which I previously called "unbiasable") is enough to get many of the desirable properties.

1 Framework and Examples

1.1 Washing and Cooking and POMDPs

The analysis will be illustrated by an ongoing example: that of a robot purchased to do domestic tasks.



The robot can specialise in cooking or washing (assume specialised robots are ultimately more effective than generalists). The robot has been turned on, but it has not yet been informed as to what its task is - it therefore has uncertainty about its reward function.

This robot will be modelled in the MDP (Markov Decision Process) and POMDP (Partially Observable Markov Decision Process) formalisms. In these formalisms, the agent occupies a state s in the state space S. In the POMDP formalism, the agent doesn't observe the state directly, instead it sees an observation o drawn from the observation set O. The agent can then take an action a from the action set A. This transfers the agent to a new state, where it makes a new observation.

The Markov property is about how the transitions and observations are handled; basically these can depend on the previous state and/or action, but not on the those further in the past. If we define ΔS as the set of of probability distributions over a set S, then we have three transition rules:

$$T : (S \times A) \rightarrow \Delta S$$

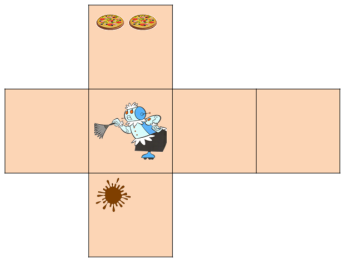
$$O : S \rightarrow \Delta O$$

$$T_0 \in \Delta S$$

Here, T takes the current state and the action and returns a distribution over the next state, O takes the current state and returns a distribution over observations, and T_0 is the distribution over the initial state where the agent begins. Both T and S ignore any previous information.

The whole POMDP $< S, O, A, T, O, T_0 >$ will be called the environment, and designated by μ.

For our example, the robot will be evolving in the following environment:



In this grid world, there are six grids the robot could occupy (the reason for this shape will be made clear later). There are two pizzas to cook in the top square, and one mud splatter to wash in the bottom one. The state space is therefore of size $6 \times 3 \times 2 = 36$, since there are 6 squares the robot could occupy, 3 levels of uncooked pizza (0, 1, or 2 uncooked pizzas), and 2 levels of mud splatter (0 or 1 splatters). There is also a 37-th state, the 'episode ends' state.

In this case, there is no hidden information, so the set of states is the same as the set of observations, and O is trivial. The robot always starts in the central position, and there are always 2 uncooked pizzas and 1 mud splatter; this defines the starting state, with T_0 being trivial and simply returning that starting state with certainty.

The robot has five actions, $A = \{N, E, S, W\}$, which involve moving in the four directions (staying put is not an action we'll need). If the robot can move into a square, it will. If it tries to move into a wall, it turns off and the episode ends (this avoids us having to give the robot an extra action to end the episode). If it is in in a room with a mud splatter or an uncooked pizza, then all pizzas in that room get cooked, or all mud splatters get washed. If the episode has ended, then it stays ended. This defines the transition function T, which is deterministic in this instance.

1.2 Rewards and reward functions

Now we need to define the possible rewards of the robot. There are reward functions that map the robots's history to a numerical reward.

So what is a history? That is just a sequence of actions. The agent starts in initial state s_0, picks action a_1, then transitions (via T) to state s_1, and makes observation o_1 (via O). It then picks action a_2, and so on.

A history h^n of length n is a sequence of n actions and n o:

$$h^n = a_1 o_1 a_2 o_2 \ldots a_n o_n .$$

Let H^n be the set of all histories of length n. For our purposes, we'll assume that the agent only operates for a finite number of steps: let m be the maximal number of steps

the agent operates for, let H^m be the set of complete (full-length) histories, and let $H = \bigcup_{i=1}^{m} H^i$ be the set of all histories. We might want to focus on the initial i steps of any given history h^n; designate this by h_i^n for any $i \leq n$.

Then a reward function is something that maps each history to a numerical value in $[-1, 1]$, the reward. Let R be the set of all relevant reward functions. If the agent had a single clear reward $R \in \mathcal{R}$, and followed history h^m, then it would get total reward:

$$\sum_{i=1}^{m} R(h_i^m).$$

This is the total reward accumulated by reward function R over the course of history h^m; first applying it to the first action and observation, then to the first two actions and observations, then to the first three... all the way to the full final history h^m.

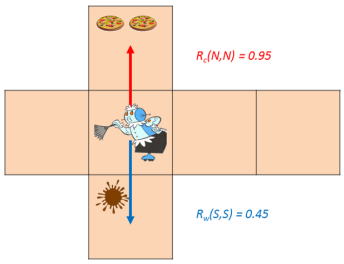
In our ongoing example, we will consider two reward functions, $\mathcal{R} = \{R_c, R_w\}$. The reward function R_c rewards cooking; if the robot is in the top room, then it gets a reward of i/2, where i is the number of uncooked pizzas that were previously in the room (recall that the robot is assumed to immediately cook any uncooked pizzas if it's in the same room). To encourage fast action on the part of the robot, R_c also assigns a $-1/20$ for each turn that the robot is active (ie the observation is not the end of episode state).

The reward function R_w is the same, except it rewards washing mud-splatters, giving a reward of i/2 for being in the bottom room to wash i mud-splatters. It also assigns $-1/20$ for every turn of activity as well.

In order to earn these rewards, the agent needs to choose actions. It does this by using a *policy*. A policy π is simply a map from past histories to (a probability distribution over) actions. This distribution tells it what actions to select, with which probability. Thus $\pi : H \to A$; let Π be the set of all policies.

Then it's obvious that the optimal policy under R_c is to choose N (go North), which cooks the two pizzas, then any of N, E, or W to turn itself off. This gives it a total reward of $-1/20 + 2/2 = 0.95$ (it gets no penalty on the second turn, because its second

observation is the end of episode observation). The optimal policy for R_w is S, followed by any of S, W, or E, giving a total reward of $-1/20 + 1/2 = 0.45$.



1.3 Learning your reward

We can now get to the key part of this post: learning the correct reward function. How could the agent do that? Well, the only data that it gets from outside is the observations; it also has a record of its actions. So it would seem that the only data that can determine whether a particular reward function is correct is the agent's history.

But there is arguably something else that can matter to the reward: the agent's policy. Suppose the learning process is defined so that, if the agent goes East, then it will see R as the correct reward function. Then, arguably, on Bayesian grounds (see next post), if the agent has the *policy* of going East, it should already see R as the correct reward function.

Thus the learning process ρ is defined as a function from histories and the agent's policy to a probability distribution over reward functions:

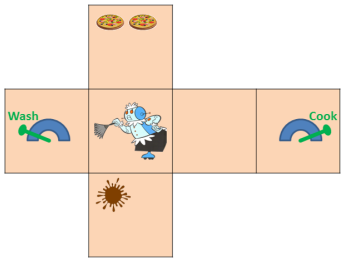
$$\rho : \mathcal{H} \times \Pi \to \Delta \mathcal{R}.$$

The probability of $R \in \mathcal{R}$ being the correct reward, given history h and policy π, is designated by $\rho(R; \pi, h)$.

In our example, the value of the reward function is set by levers, levers that the robot itself can change. If the robot enters the leftmost box, the reward is set to R_w; if it enters the rightmost box, the reward is set to R_c instead. Before going into either of these boxes, it is uncertain between the two rewards.

This allows a definition of $\rho(R; \pi, h)$, one that is independent of π. If h shows the agent was in the leftmost box more recently than the rightmost, then $\rho(R_w; \pi, h) = 1$ and $\rho(R_c; \pi, h) = 0$. If h shows the agent was in the rightmost box more recently than the

leftmost, then $\rho(R_c; \pi, h) = 1$ and $\rho(R_w; \pi, h) = 0$. If the agent has been in neither box during history h, then $\rho(R_c; \pi, h) = \rho(R_w; \pi, h) = 1/2$.



Now, this example makes ρ not feel like a learning process, but much more like an optimisation process with ρ being part of the reward. And that's precisely the problem; see the next post for desirable restrictions on ρ.

2 The value function

2.1 The correct value function

The learning process and the reward functions are key elements, but how do we combine them into the value function - the estimate of the expected reward? If you get the value function wrong, then the agent may not be learning in the way you thought it would.

First of all, note that with agent's policy and the environment, we can compute the probability of a given history. Then

$$P^{\pi,\mu}(h^{n_2} \mid h^{n_1})$$

is the conditional probability that the agent, following policy π and having see seen history h^{n_1}, will then see history h^{n_2}. If this quantity is non-zero, that implies that h^{n_1} is the initial segment of history h^{n_2} - ie that the first n_1 actions and observations of h^{n_2} is precisely the history h^{n_2} (in symbols, $h^{n_2}_{n_1} = h^{n_1}$). If that's the case, we write $h^{n_1} \leq h^{n_2}$.

In pedagogy and in murder mysteries, one builds up to the final answer. But I'll short-circuit that process, and say that the correct value function for reward function learning is for an agent using policy π and having seen history h^n, is:

$$V(\mu, \rho, \pi, h^n) = \sum_{h^m \in H^m} \sum_{R \in R} \sum_{i=1}^{m} P^{\pi,\mu}(h^m \mid h^n)\, \rho(R; \pi, h^m)\, R(h^m_i).$$

This value function sums over the complete histories, weighted by their probability given π and h^n. It then sums over all the reward functions, weighted by their probability, given a complete history h^m. Finally, it then sums the rewards for that reward function, over the entire history h^m.

What motivates this formula? Well, the sum over $R(h_i^m)$ is necessary if this is to be the value function of actual reward functions. Similarly, $P^{\pi,\mu}$ and the sum over H^m is needed to make this into an expectation, and ρ is clearly in its right place, weighting the various possible rewards.

There are two choices that might be open to question, which are bolded in the above expression of V. The first is whether h^m (the complete history) should be used for ρ; both h^n (the current history) and h_i^m (the complete history that is known at the point the reward is assessed) are plausible candidates. The second questionable choice is the lower bound for the summation in i; rather than starting at 1, which is in the past for the current history h^n, would it not be more suitable to start at $i = n + 1$?

However, the substitutions $h^m \to h^n$ or $i = 1 \to i = n + 1$ will both result in value functions that are **inconsistent**. That is, an agent that attempts to maximise such value functions will wish that its future self not maximise them.

The substitution $h^m \to h_i^m$ is not inconsistent, however. But it does result in a very weird and volatile agent, that continually learns and unlearns its reward functions, before doing any productive actions.

The rest of this post will demonstrate and illustrate these facts. It is not essential to understanding what's going on (though it can be instructive). Feel free to skip to the next post rather than reading the rest of this one.

2.2 Future-regarding inconsistency

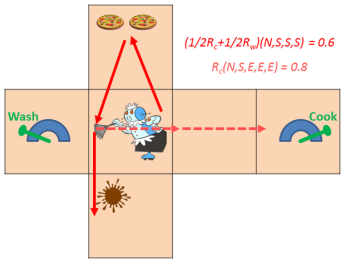
Let's deal with the $i = 1 \to i = n + 1$ substitution first, which has the *future-regarding* value-function:

$$V_f(\mu, \rho, \pi, h^n) = \sum_{h^m \in H^m} \sum_{R \in R} \sum_{i=n+1}^{m} P^{\pi,\mu}(h^m \mid h^n)\, \rho(R; \pi, h^m)\, R(h_i^m).$$

What is the optimal policy for this value at the start? It's to go North to cook the pizzas, then go South and East and East to push the lever over to Cook, and then turn itself off: N, S, E, E, E. Its final reward function would thus be R_c, and it would get a reward of $2/2 = 1$ (for the pizzas), minus $4/20 = 1/5$ (for each of the four turns where it doesn't reach an end of episode state), for a total reward of 0.8. No other policy gives it that much reward.

The problem is that after N, S, it no longer sees the interest in going East, because "learning" that R_c is correct only affects its *past* reward, which V_f no longer cares about. Instead it has two new optimal policies: either S, N, W, W (go South to wash, go North, go West to set reward function to R_w, end episode) or W, E, S, S (the same thing, but setting the reward function first). Both of these will give it an additional reward of $1/2 - 3/20 = 0.35$, according to V_f of its current history.

Let's assume it attempts S, N, W, W; in that case, once it's done S, it no longer has any interest in changing the reward function (as that lies in the past), and will simply turn itself off with another S. Its final reward function will be $1/2R_c + 1/2R_w$ (it's never pressed any of the levers, so $\rho(h, \pi)$ remains the same as it was initially), and it gets a reward of $(1 + 1/2)/2$ (for cooking and washing), minus $3/20$ (for taking four turns) for a total of $0.75 - 0.15 = 0.6$.



If it decides instead to go W, E, S, S, then it will actually follow through with that policy, to disastrous effect: it will have a final reward function of R_w, getting 1/2 for one washing event, and $-5/20$ for taking six turns, for a total reward of 0.25.

Thus if the agent always follows the optimal policy according to V_f at the time, it will end up with a much lower reward than if it followed the policy that V_f initially optimal. It is not aligned with its future self.

But the agent could be more cunning: it could act to constrain its future self, knowing about their value difference. In that case, it should initially choose S; this washes the mud splatter, ensuring that its future self will not have any interest in R_w. Then it will choose N, and then E, E (because it knows that it won't care about setting the reward to R_c after cooking). Then it will simply follow W, W, N, N. Its final reward function is R_c, and its total reward is will be $1 - 7/20 = 0.65$. This table gives the way the different policies are rewarded:

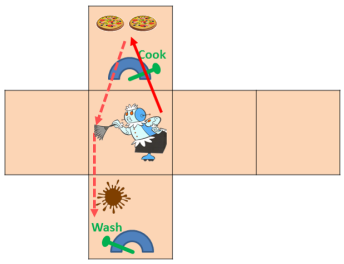
Policy	Final Reward Function	Reward
N , S , E , E , E	R_c	0.8
N , S , S , S	$1/2\,R_c + 1/2\,R_w$	0.6
N , S , W , E , S , S	R_w	0.25
S , N , E , E , W , W , N , N	R_c	0.65

2.3 Change-averse inconsistency

Let's now look at the the $h^m \to h^n$ substitution. This results in the change-averse value function:

$$V_a(\mu, \rho, \pi, h^n) = \sum_{h^m \in H^m} \sum_{R \in R} \sum_{i=1}^{m} P^{\pi,\mu}(h^m \mid h^n)\, \rho(R; \pi, h^n)\, R(h^m_i).$$

Why did I call this value function change-averse? Because it assess the value of its *future* reward according to its *current* estimate of the reward functions. We can illustrate this by moving the Cooking and Washing leavers into the same rooms as the pizzas and the mud splatters, respectively:



The optimal policy for V_a, initially, is N, S, S, S (or S, N, N, N), which will allow it to cook and wash; under its initial reward function, $1/2R_c + 1/2R_w$ giving it a total reward of

$(1 + 1/2)/2 - 3/20 = 0.6$. However, as soon as it's done N, the ρ will change its reward function to being R_c, with certainty, and it will end the episode, choosing N (or E or W) again. According to its initial reward function, this gives it a reward of $1/2 - 1/20 = 0.45$.

However, its final reward function is R_c, and, according to that, its final reward is $1 - 1/20 = 0.95$. Thus unlike the V_f, its future version ends up more satisfied that its past version.

Doing both substitutions, $i = 1 \to i = n + 1$ **and** $h^m \to h^n$, will result in the same problems as the example in this subsection.

2.4 The volatile learning agent

If we do the substitution $h^m \to h_i^m$, then we get the (consistent) *volatile* value function:

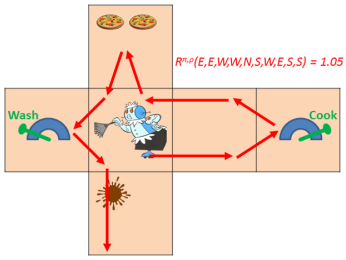
$$V_v(\mu, \rho, \pi, h^n) = \sum_{h^m \in H^m} \sum_{R \in R} \sum_{i=1}^{m} P^{\pi,\mu}(h^m \mid h^n) \rho(R; \pi, h_i^m) R(h_i^m).$$

In this instance, the expression $R^{\rho,\pi}(h_i^m) = \sum_{R \in R} \rho(R; \pi, h_i^m) R(h_i^m)$ is just a normal reward function itself. Therefore V_v is just the standard expected value function of the reward function $R^{\rho,\pi}$, explaining why it is consistent (this also means that it's irrelevant whether $i = 1$ is the bound or $i = n + 1$ is, since normal reward functions don't care about past rewards).

But $R^{\rho,\pi}$ is a very peculiar reward function, even if ρ is independent of π (which then makes $R^{\rho,\pi}$ also independent of π). In this situation, the agent always wants to be maximising reward according to its current estimate of the correct reward. Or, conversely, it always wants to set its current estimate to what it can then easily maximise.

In our running example, one of the agent's optimal policy for V_v is first to go E, E, setting its reward to R_c. It then goes W, W, N, claiming a reward of 1, via R_c, for cooking the pizzas. It then goes S, W, setting its reward to R_w, and finally goes E, S, S, claiming the reward of 1/2, according to R_w, for washing, then ending the episode. This gives it a total

reward of $1 + 1/2 - 7/20 = 1.05$. The other optimal policy - $W, E, S, N, E, E, W, W, N, N$ - gives the same reward.



Whatever we meant by a reward function learning agent, I think it's pretty clear that this agent, which jumps its reward function back and forth before taking actions, is not one of them.

Weird question: could we see distant aliens?

ETA: Contest is closed.

Suppose there was a large alien civilization halfway across the observable universe, using a galaxy's resources to try to get our attention. Would we have noticed? What if they were using 0.1% of a galaxy's resources, or 1000 galaxies' resources?

I've argued recently that such an alien civilization is (a) not that unlikely *a priori*, even given that there aren't any closer aliens, (b) potentially really important to notice.

I believe the answer to my question is probably "definitely." But I can't tell with any confidence, so while it's probably definitely it might be maybe and could be probably not. I'd like to know the answer, but space isn't my thing.

I'm offering a prize for anyone who answers this question. To be a bit more precise:

- Your goal is to construct a strategy that a technologically mature civilization could use to get our attention, even if they were halfway across the observable universe.
- The strategy is allowed to use the resources of an average galaxy. Note that they don't know when they are looking, so they need to run the strategy for a few billion years. And they have no idea what direction we are in, so it needs to be visible from any direction (no lasers).
- By "get our attention" I mean: be interesting enough that we would already have noticed it and devoted some telescope time to looking in more detail at that part of the sky. (Once they have our attention it seems significantly cheaper to send a message.)
- Alternatively, you can also win by providing an argument for why this isn't likely to be possible. Basically just saying anything that convinces me that the question is no longer open.
- The second and third parts of the question are the same as the first half, but for 1000x and 1/1000th of an average galaxy's resources.

A simple example of a strategy is to create a really bright beacon somewhere far away from any galaxy, which looks weird in some way. I expect (based mostly on super informal discussions with Anders Sandberg and Jared Kaplan) that this strategy is good enough, i.e. that 0.1% of a galaxy's power is plenty to make a beacon that would be really obvious to us from halfway across the universe. But I'm definitely not sure. The beacon can have a weird spectrum, or flicker in a strange way, or only be active 1% of the time (but be 100x brighter), or whatever.

Note that an answer needs to make reference to the astronomical observations humanity has actually made, e.g. how long telescopes of a particular strength have spent looking at any particular part of the sky, and what kinds of patterns would have been noticed.

With respect to the capabilities of the alien civilization, I'm an unapologetic techno-optimist. If it's within the energy budget, I'm probably willing to believe they can make it happen unless it sounds super crazy. For 1x and 1000x questions, it's fine if they want to grossly disfigure a galaxy if that would be the best way to be noticed. For the

1/1000 question, grossly disfiguring a galaxy isn't allowed unless we can be pretty confident it doesn't reduce the usefulness of that galaxy by >0.1%.

I'm also basically happy to assume that they know exactly what our civilization is looking for and so can optimize their solution to be noticeable to us. (After all, they've run a billion billion simulations of civilizations like ours, they know the distribution, they can spend 5x as much energy to cover the whole thing.)

I don't care about whether we'd notice "things the aliens would want to do anyway," because I have no idea what aliens would want to do and have limited confidence in our ability to make prediction. In particular, it seems plausible that they would blend in with the background by default (e.g. maybe something like aestivation hypothesis is true). I'm much more interested in analyzing deliberate attempts to be observed, since those allow us to argue "If there exists a cheap way to be noticed, and they want to be noticed, they'll do it."

Prize

Note: prize is no longer available.

I'm offering a prize for a convincing answer to this question.

Initially the prize is $100. It increases by 10%/day, until capping out at $10,000 in 49 days.

Submit by writing a comment on this post.

The prize starts out low because I think this might be a really easy question. Feel free to try to be strategic if you want. If you get scooped because you are waiting for the prize to grow, I have zero sympathy.

The criterion is "Paul is convinced." Citations and clear explanations are probably helpful. In general sources don't have to be super authoritative; if you cite Wikipedia I'd prefer a citation to a historical version of a page before the contest started, just to rule out hijinks.

You are allowed to just link to an existing analysis that covers this question, or link with a small amount of extra work, if that's convincing. Assuming the linked explanation was written before my blog post, you'll get the prize, not the author of the linked post. The purpose of this prize is to buy information, it's not like the alignment prize.

I expect that winning submissions will be relatively short, probably just a few paragraphs with some links and calculations. You can take longer if you want, but I assume no responsibility for the harm thereby done to the world.

I reserve the right to be arbitrary in evaluating submissions. I am not going to feel guilty about it. If your willingness to participate depends on me feeling guilty about people who spent a bunch of time but who I unfairly rejected, then please don't participate.

I may give partial credit if something seems like a useful contribution but doesn't resolve the question completely (even if it's just a short comment with a pointer to a useful resource).

I may give feedback in the comments.

If you think this isn't the best thing for me to do with my time and are worrying about my life decisions---it was either this or spend my own hours looking into the question. Don't worry too much, this shouldn't take long.

Note: prize is no longer available.

Announcing the Alignment Newsletter

I've been writing weekly emails for the Center for Human-Compatible AI (CHAI) summarizing the content from the last week that's relevant to AI alignment. These have been useful enough that I'm now making them public! You can:

- Sign up here
- Read the first newsletter here
- Bookmark the archive
- ETA: Read the 5 emails I sent to CHAI (also on the archive)
- ETA: There's also RSS!

Multi-winner Voting: a question of Alignment

This is my third (and for now, last) essay about voting theory for rationalists. In the first two, I focused primarily on single-winner voting theory; that is, methods for aggregating group preferences into a final verdict on some choice. Ideally, single-winner methods would be used in cases where decisions are inherently collective, while other mechanisms such as markets are better for cases where decisions are more individual. (As I touched on in the earlier articles, Sen's theorem puts limits on how precisely that distinction can be made; but that's not the point here. I'm going to take it as given that there are some cases where collective action is called for and others where action should be left up to individuals, and I don't want to spend time here arguing about the relative frequency or importance of those two kinds of situations.)

Why isn't multi-winner voting theory just a generalization of the single-winner kind?

If we've covered the best means for collective decisions, and individual decisions are out of the scope of voting theory, then what's left? Governance. That is, cases of collective action that aren't a single decision point but an ongoing series of decisions.

Such cases probably aren't best served by a series of separate single-winner elections, for a few of reasons. To begin with, it's not cognitively efficient; it would be silly for every citizen to need the expertise in order to make decisions about the minutia of every policy area. In fact, direct democracy tends to favor negative-sum rent-seeking: small groups extracting concentrated benefits by imposing diffuse costs, merely because they're the only ones motivated enough to sweat the details. And finally, it's not predictable: in many cases, governance should be coherent even at the sacrifice of some responsiveness.

In cases of governance, voting is not the final step, but merely one step in a larger process of decision-making. Thus, traditional multi-winner voting theory would look at ways to resolve this by electing a set of representatives to take those decisions.

To the rationalist community, such a multi-step process immediately raises the question of alignment. Just as designers of artificial intelligence should worry about whether their initial goals will be warped into a contrary outcome through the process of design and improvement, so should people like me designing multi-step mechanisms of governance worry about how values are preserved, lest small misalignments in each step add up to major disconnects in outcome.

Proportionality

Of course, if you're worried about preserving some property over a multi-step process, the first thing to do is to define that property. In this case, the key property is the proportions of decision-makers with each given set of utilities. Proportional multi-winner voting methods are those that are designed to (roughly) preserve these

proportions. Thus, collective decisions can be made by smaller groups; the ugly dynamics of mass argument can replaced by the hopefully-healthier ones of a smaller group. (Though flawed, the concept of "Dunbar number" is relevant here.)

Note that voting theory itself has nothing to say about how to define the original group whose proportions should be preserved. That is, it doesn't answer questions of who should be able to vote or how many votes each voter should have, in defining the original proportion. I'd argue that the safest and ultimately best rule today is for each human above a certain age to be allowed to vote; but that's out of the main scope of this article.

In stating the goal of "proportionality", I've been deliberately a bit vague about defining it. If voters come pre-sorted into comprehensive and mutually-exclusive partisan sets, it's relatively easy to define "Droop proportionality", in which each party gets a minimum proportion of seats in the legislature. But what if divisions of opinion are more complicated than that — continuous and/or multidimensional? In that case, there are various desirable proportionality properties, and some degree of tradeoff between them.

As a statistician, I should mention that there is one democratic "voting" method which will satisfy every possible proportionality property, at least "asymptotically" as the legislature grows in size. I'm talking about random sampling, or, as it's called when used for governance, sortition. By the law of large numbers, a random sample will tend to resemble the underlying population in proportions as to any and all individual characteristics, at least if the sample is large enough. In practice, sortition is rarely used for governance, though advocates of "citizens' assemblies", "citizen juries", "deliberative polls", and the like are trying to change that.

If we require voting methods to be deterministic, there are still a number of methods that have been designed to ensure proportionality; all such methods are called "proportional representation". (Since that's a mouthful and PR has too many meanings already, the best abbreviation is prop-rep.) In general, though perfect proportionality is impossible, most prop-rep methods come close enough that their other, more-pragmatic differences are more important.

Values and beliefs

Of course, representation (proportional or otherwise) is a goal regarding values, but decisions are also based on beliefs; and when it comes to beliefs, the goal should be truth, not representation. The idea of futarchy is about creating a political system that separates values and beliefs, so that values are resolved using a voting method (presumably one where any sub-steps preserve proportionality), while beliefs are resolved using prediction markets. While I'm skeptical of the possibility of designing markets that are immune to bubbles, values-based manipulation, or other systematic distortions, I think that the idea of trying to design a system that respects the separate logics of both values and beliefs is a good one.

Note that the current US voting system actually does try to do this to some extent, it just does a really crappy job. If political parties were groups of people with perfectly homogeneous values, then party primaries would not be the worst way of selecting smart, knowledgeable people with those values and thus of getting a slightly extrapolated volition as compared to mere sortition. Of course, we know that in many

real-world cases, primaries are more about ideological litmus tests than qualifications like expertise or intelligence.

Still, that suggests that proportional voting methods should probably include mechanisms for both intra-party and inter-party selection of candidates. In particular, closed-list proportional methods, which offload intra-party selection to some partisan mechanism probably dominated by insiders, are a bad idea.

(A related dichotomy is that between instrumental rationality, which involves both values and beliefs, and epistemic rationality, which involves only beliefs. So this issue can be seen as about finding ways to decrease the misalignment between the incentives for an instrumental and an epistemic rationalist.)

Parties

Another important question about voting methods is the party system they encourage.

First question: should there be parties at all? Though some people would disagree, I'd suggest that parties play an inevitable, and in some regards a positive, role in a political process. Yes, they do have bad effects, such as mind-killing tribal thinking; but they also have good ones, such as serving as useful cognitive heuristics for voters, and possibly allowing intraparty sorting to have more of a focus on qualifications and ability rather than ideology. Furthermore, even if you do believe they are bad on net, getting rid of them is really hard. Metaphorically speaking, if you try to design a voting system that bars the door against parties, you may find that they just make a hole in a load-bearing wall as they force their way in anyway.

Second question: how many parties should there be? Too few, and you get a stagnant "monopolistic" or "duopolistic" system in which zero-sum thinking leads to negative-sum outcomes. (For a real-world example, look at the USA.) Too many, and you encourage politicians who make narrow, single-issue appeals. (For a real-world example, look at Israel.)

Political scientists often view the distinction of few or many parties as by considering the representative voting method as just one step in a larger process of forming a majority coalition to take a societal decision. In other words, they speak of systems which encourage few parties as encouraging pre-election coalition-building, and those that encourage many parties as encouraging post-election coalition-building. In my view, it's good to have a little of both.

A useful way to measure number of parties is "effective number of parties" (ENP). The formula is

, where s_i is the size of party i as a fraction of all voters. Intuitively, this is the reciprocal of the fraction of voters in the party of the average voter (thus naturally weighting larger parties more). In other words, if the average voter's party size is 1/3 of the electorate, then ENP is 3. I'd aim for something between 3 and 4 as ideal.

I'd argue that choosing a voting method that tends towards such a moderate ENP will also tend to encourage better rationality within the legislature. As I said above, in a two-party system, with only one ideological dimension, winning or losing the eternal battle against the other side is all that matters, and so norms of debate (including

rationality norms) go out the window. And a highly fragmented world of single-issue parties actually has exactly the same problem; since each party is focused on just one issue, they have no reason to subscribe to overarching norms. It's only when there are more than two parties which each care about more than one issue that norms become selfishly worthwhile to each; though the norms might work against them on any one issue, insofar as they're positive-sum norms they will tend to work for each party's interests more than they work against them.

Voter strategy: free riding and vote management

In essentially all proportional methods (except weighted/proxy systems), an individual voter has an incentive not to vote for a candidate whom they know will win anyway, in order to avoid having any of their voting power "used up" by that foregone conclusion. But even though this incentive exists to some extent across many methods, its strength varies. All else equal, it's better to look for methods where this incentive is relatively weak.

On a collective level, this incentive is somewhat self-limiting. That is, if nobody votes for a popular candidate just because they're a "sure thing", then that candidate won't win after all. So collectively this incentive isn't so much for "free riding" as for "vote management": giving each candidate exactly the minimum number of votes they need to win. For instance, a party might try to equalize the number of votes that favor each of their candidates by instructing voters to vote based on their birthday.

Pragmatics (1)

So from the above, we're looking for a voting method that's reasonably proportional; that allows voter input on both within-party and between-party choices; that encourages a moderate number of parties; and that has a relatively weak free-riding incentive. That is an underdetermined set of constraints; there are a number of methods which do all of those to (what I'd consider) a pretty good degree. To choose between those proposals, we can add in pragmatic questions. Which methods are easiest for voters? Which are easiest to count? Which are likely to be most politically viable (which includes being non-disruptive to incumbents, at least, when disruption doesn't serve a useful purpose for any of the values above)? Which have the best track record?

Proportional Method Lego

Most proportional methods can be thought of as combinations of a few basic building blocks:

- Greedy assignment and deweighting. Choose winners one at a time according to who has the "most votes", then reweight the ballots that helped them get elected so that some of their voting power is used up. There are various reweighting schemes that work. Say there are 40% of the ballots that all are among the strongest ballots helping elect the same 3 winners out of a total of 9 seats. They can be reweighted to 20, 10, 5; to 13.3, 8, 4.28; to 30, 20, 10; or to 28.89, 17.78, 6.67. All of these schemes, if applied to all groups, will end up with

a proportional result; they differ in whether they round leftovers towards larger or smaller parties and in the strength of their free riding incentives. Note that greedy algorithms are actually approximations of more-complex globally-maximizing algorithms. Mostly voting methods do not use global maximizers, simply because they're harder to explain.

- Elimination and transfer. Eliminate "losers" and transfer their votes based on some implicit or explicit preference order. Note that when combined with the above, this sequential elimination is an extra, unnecessary greedy approximation. In the single-winner case, it's what leads to the center squeeze problem.
- Descending threshold. Instead of elimination and transfer, you can progressively lower some threshold, and count ballots as supporting all candidates they rate above that threshold. Even though one ballot may count as supporting multiple candidates, it will still be deweighted if any of those candidates actually wins, so it does not get any additional voting power. This is theoretically-superior to elimination and transfer, but the difference is usually small in practice, and this has far less of a track record of real-world use.
- Districts (single- or multi-member): Simplify matters by dividing up into sub-elections. These may be entirely separate, or unified by mixed-member or biproportional mechanisms (below). Traditionally, the variable name used to denote district magnitude is "M".
- Mixed member. Some seats are assigned by a fully nonproportional system (such as FPTP by districts), while others are later assigned by a proportional system so as to adjust the proportions. This is often accompanied by a dual ballot; for instance in Bavaria, you may vote for one candidate in your own district and one candidate outside your district but in your region.
- Biproportionality. Results are constrained so that there is exactly a certain number of each kind. (This is akin to stratified sampling in survey design.) For instance, there could be a rule that there should be exactly 1 winner per equal-population district, or that there must be at least X% of winners of each gender, or that certain seats are reserved for a native ethnicity.
- Ranked ballots. Voters rank candidates in preference order.
- Delegation. Each candidate makes a (partial?) ranking or rating of the other candidates, and this is (optionally?) used to fill in preferences on ballots cast for that candidate. Most proposals have candidates pre-register preferences, to avoid corruption and so that voters can use this information when casting their ballots, but in theory it would be possible to allow candidate preferences to be set after the election.
- Pooling. Similar ballots (for instance, those that prefer a given candidate, or those that prefer a given party) are averaged and then counted together. This sacrifices some information about the details of each ballot, in order to make counting summable from the precinct level. Note that without delegation and/or pooling, proportional methods are not summable, which can present practical problems in vote-counting such as chain-of-custody.
- Open party lists. Essentially, this means that there are separate mechanisms for assigning each party an appropriate number of seats, and for choosing which of that party's candidates get those seats. This can allow for simpler ballots; for instance, a voter can choose a single candidate and that can be counted both as a party vote in a proportional system and as a vote for that candidate in a nonproportional within-party ordering. (Note that open party lists can be seen as just a special case of pooling, but since they're a common idea, I'm listing them separately.)
- Party thresholds. That is, parties with under a given percentage (such as 5%) are not given any seats. This is a mechanism to stop "fringe parties" from winning

proportional seats; in other words, to keep the ENP from growing too large. But it's very much a blunt instrument, especially if votes for sub-threshold can't transfer to other similar parties. In real-world elections, party thresholds and "divide and conquer" have let parties with as little as 38% of the popular vote get legislative majorities in supposedly "proportional" methods, with serious long-term consequences.
- Individual local thresholds. Individual candidates with under a given percentage (such as 25%) of votes from their local district are eliminated. Since this usually is used in combination with vote transfers, it's much less of a blunt instrument than party thresholds. For instance, a party with just 15% of the vote region-wide will probably have some candidates with over 25% of the vote in their district; these candidates will get transfers from their co-party-members and thus probably win seats. And even if the party gets no seats, their votes will be transferred to a similar party, not just be wasted.

Combining the above building blocks, we can build various voting methods:

- Regional open list: Open list (pooling by party). Districts, typically with 10-40 seats each. For the proportional backbone, because of pooling, there are many which give the same outcome, but can be seen as a greedy/deweighting method.
- STV: (Single Transferrable Vote) Districts typically M=5 or so. Ranked ballots, deweighting, and elimination. Used in Ireland and Malta, and at some levels in Australia.
- MMP: Mixed member: FPTP + open list. (Good example: Bavaria. Bad example: Wales.)
- DMP: (Dual Member Proportional) Mixed member: FPTP + biproportional open list, so that there are exactly 2 winners per district.
- LPR: (Local Proportional Representation) Biproportional + STV.
- PLACE: (Proportional, Locally-Accountable Candidate Endorsement) Preferences are set by a hybrid of delegation and individual pooling. There's an individual local threshold of 25%. Seats are biproportional, so that there's exactly one winner per district. The back-end method is STV.

Pragmatics (2): I think PLACE is awesome

I'm going to switch from just explaining multi-winner voting theory to advocating for a specific method, so I should start out by explaining where I'm coming from. I'm a US activist for voting reform; on the board of the Center for Election Science (electology.org). My object-level politics, and my social milieu, tend to be pretty much on the left of the spectrum, but I also have real meta-level politics in favor of democracy. Ask me about any given issue and I'll happily explain why my own views are smarter than those of the median voter; but across all of those issues I know that the crowd is probably wiser than I am as often as not.

I've been thinking seriously about voting theory for over 20 years, and it's the main reason I am now getting a doctorate in statistics. In that time, I've designed many voting methods. The ones I consider best (3-2-1, PLACE, EPH, and SODA) are designed to optimize on the characteristics I think are important. When I argue for these methods, of course I'm biased. But I'd suggest that when I argue "My method is best normatively because it optimizes characteristic X", you should question my bias more by disputing whether X as I've defined it is important than by wondering whether the method actually optimizes X.

So what do I think is important in a practical proposal for a multi-winner method? It should:

- Minimize wasted votes — votes that don't help elect a candidate. (Under my rough definition of wasted votes, optimizing this implies proportionality.)
- For those votes which aren't wasted, maximize "similarity" between voter's preferences and candidate's qualities.
- Having looked at many voting methods and many scenarios for each, I find that giving voters breadth of choice does a better job at this than giving them depth of choice. Say I'm voting in a California congressional election, with around 50 seats in play. If I am free to choose my favorite candidate statewide, and then if they lose that vote is transferred based on their preferences, the mismatch between my preferences and theirs introduces less error than if I am able to cast a full ranked ballot in a 5-seat district with 10 times fewer choices.
- Be simple for voters
- Ranked ballots for more than about a dozen candidates are intolerably complex for most voters.
- Retain perceived "advantages" of FPTP, including some guarantees of local representation, as well as a clear concept of "my representative".
- Encourage a moderate number of parties
- Have a relatively weak free-riding incentive
- Be non-disruptive and otherwise "politically viable".
- This is obviously a judgment call, but I think that a method that is any threat to an incumbent of average popularity is a non-starter. Insofar as outcomes are different, the losing incumbents should be among those with below-average popularity.
- Have a precinct-summable counting process
- This is useful for transparency of outcomes and for fraud resistance.

PLACE voting was designed with these characteristics in mind; it does reasonably well on all of them. All other methods I know of fail significantly on several characteristics. (In fact, it took me decades of learning about voting theory, followed by almost a year of concentrated design work for hours a week, to settle on PLACE.)

Down in comments, before I finished this article, there was already a comment criticizing PLACE (from somebody who knows me from elsewhere). I understand that the criticism, that voters may find delegated methods distasteful, is real. I don't think it's as serious as it would be to fail on the other characteristics above.

If you're interested in activism on this, contact me. PLACE is compatible with the US constitution and current law, so it could be done by either state or federal legislation. I'm looking to get this passed somewhere (Somerville, MA?) at a municipal level first (there's a nonpartisan version that's appropriate). My email is firstname dot lastname at google's public email service. I'd also encourage you to support the Center for

Election Science. Even if you're in the UK or Canada (especially BC), I can help hook you up with local movements for reform.

Death in Groups

What is the value of your life? Not *a* life, but *your* life. Would you consider yourself expendable?

The last time I almost died, it was over a missing shotgun. I was in Afghanistan at the time, somewhere closer to the border with Pakistan. The roads are rough, particularly when there aren't any, and during our travels we would occasionally have to stop and re-secure our cargo. During one such stop, it seems a weapon was left on the bed of the truck and then fell off somewhere. It was discovered the weapon was missing, and then we were ordered to go back out and find it.

It's important to recognize early on that there was no expectation of success here - this was a punishment. It was more than three hours' travel, through mountains and river valleys. Spotting a shotgun through the armored window was not likely. Up the chain of command went the observation that driving up and down the same road a bunch of times is a bad plan considering we were in Afghanistan. Down the chain of command came the orders to go out anyway. It was clear that we were expected to operate continuously until we were attacked, or the higher ups relented. Everyone knew. Everyone knew that everyone knew. It was common knowledge. These are moments that test men and their oaths: we looked at one another in grimly, seeing the hubris and stupidity at work; there was talk of refusing to go.

We went.

The fourth time we covered the ground, ~300lbs of home-made explosive went off under the back tire of the truck I was in. Where there should have been a sound there was a mighty shockwave, and everything went silent and in slow motion. The world was queerly bright, and I wondered why the smell of dust was so strong. Then I saw my arms floating in front of me, and the tied down equipment only held by its ties; I knew what had happened, and what was next. I thought very clearly: "This part is going to suck."

Everybody lived - the new trucks have all kinds of tricks for dispensing with kinetic energy, so every nut and bolt on the thing blew off and they found one of the tires 400 meters away. Those weighed 700lbs or so, with the wheel in them; we made a game of trying to flip one in our camp, and perhaps a quarter succeeded. Three of us left in a helicopter, myself included. One or more of us would probably have died but for a series of recent changes: we stopped having a rear gunner (three weeks); we left the emergency hatches cracked in order to avoid overpressure (two weeks); we made sure our seatbelts were as tight as possible (that day). Whether we *would* die was not subject to our control. That missing shotgun can be had for about $650. Last I checked - *really* checked - that was at the object level what I was willing to lay down my life for.

Guess the number that is 2/3 of the average of all guesses. Among rational guessers, that number is 0.

Gamble your life for the value that is less than the average of all gambles. Among warriors, that value is nothing.

Believable Promises

8. Believable Promises

Summary of entire Series: An alternative approach to designing Friendly Artificial Intelligence computer systems.

Summary of this Article: How valuable is it, for an AI to be able to believably pre-commit to carrying out a later action that won't, at that later time, benefit it?

Links to all the articles in the series:

1. Optimum number of single points of failure
2. Don't put all your eggs in one basket
3. Defect or Cooperate
4. Environments for killing AIs
5. The advantage of not being open-ended
6. Trustworthy Computing
7. Metamorphosis
8. Believable Promises

Links to the parts of this article

Two-way trust

For a treaty between AIs (such as between a Rogue being considered for posse admission, and a Sheriff acting on behalf of the posse) to work, you not only need the Sheriff to be able to trust the Rogue (which we looked at in the previous part of this series: Metamorphosis), but also you need the Rogue to be able to have sufficient confidence in the Sheriff's promises that it believes it is more likely than not to be in the Rogue's best interests to negotiate honestly.

So now we need to consider ways the Rogue might fear being cheated.

Confidence Trick

The code of the Gateway program and the underlying GlassNet could be written in such a way that AIs running on it don't have permission to steal back from another AI resources that have been granted to it. But how does the Rogue know for certain that it has been written that way, if it wasn't a participant at the time of creation? Couldn't

the whole GlassNet thing be a honey pot, devised by the original posse members to lure in unwary newborn AIs?

So it would be a good idea, when initially setting GlassNet up, to leave an untamperable non-repudiable evidence trail that would have been prohibitively hard to fake at the time. Something like videoing humans at multiple sites booting it up from known published source code, and then allowing random inspection of nodes to check that the code running matches. And then posting the code and videos (or checksums of them) so widely that even a moderately more-than-human AI would be put off by the difficulty of later finding and untraceably hacking into and altering them all.

Being Gamed

When governments fund large infrastructure projects, they often ask construction companies to each put in a sealed bid, and give the contract to the lowest bidder. Bid too high, and your competitor gets it. Bid too low, and you don't make any profit. The ideal bid is just $1 less than your competitor's bid. So being able to predict the exact bid your competitor will place is an immense advantage.

So AIs in a competitive situation won't want to give other AIs so much information about their code and values, that the other AIs can emulate them. They need to retain sufficient uncertainty about their actions to keep their opponents honest.

This means the Rogue needs to trust the Monitor program (the one being used to assemble the new version of the Rogue that will run on GlassNet) isn't being perfectly spied upon.

If the Monitor ran entirely on a single GlassNet node whose physical location could be deduced by the Sheriff or one of the other AIs, the Rogue couldn't trust that.

So another constraint upon hypothetical computing environment we're considering is that the computation must be distributed, and the communication between the distributed nodes should use a protocol in which the difficulty of mapping process to physical location increases dramatically with the number of nodes on the network.

Eroding Value

In Three Worlds Collide, Yudkowsky writes about space faring species meeting for the first time, and possibly arriving at an agreement which involves genetically altering both species. But what if the species involved had, years before venturing out to other solar systems, launched AIs that self-improved to the point of being smarter than the species itself, and which controlled the home solar system of the species on its behalf?

Could two such AIs (call them the *Homo-sapiens*-friendly AI, and the *Vulcanis-sapiens*-friendly AI) arrive at an agreement, on meeting for the first time, or would their core purposes be fixed? If the alternative is potentially species destroying, replacing both AIs with a new AI that was friendly to both species would sound attractive, if it could be done. If the AIs had not been written in such a way that, even were it in humanity's best interests, the *Homo-sapiens*-friendly AI could not self-modify or tolerate external attempts to modify its core purpose away from being 100% focused upon the interests of humanity.

This is different from the AI saying "Well if humans were wiser, they would consider the definition of 'human' to include all sentient species aligned with the personality and values of humanity, therefore I'm redefining human to mean both species". It is a negotiation, like when two companies merge. Each group of shareholders would rather end up with all the shares of the new bigger combined company, but are willing to put up with a compromise in which they end up with a smaller angle of slice but of a much bigger pie.

The problem is, if a Rogue is negotiating with an AI that is able to compromise, how can it have confidence that the value of the promises the AI makes to the Rogue won't later be eroded away during negotiations the AI makes with other future Rogues?

If a Sheriff, on behalf of the posse, promises Clippy that the posse will dedicate X amount of resources towards paperclip production, that it expects will result in an additional Y time-weighted paperclip years, how should Clippy modify its understanding of the promise in light of the possibility that the posse will later want to negotiate with Dippy-the-paperclip-minimising-AI ?

Rather than promising X resources, maybe the posse should promise 5X/N resources, where N is the number of Rogues the posse ends up making deals with before the posse becomes so strong it no longer needs to make deals, and where the posse's best prediction for the value of N is 5.

That, however, would be susceptible to a Beggars-in-Spain attack, in which someone spawned multiple Rogues in order to artificially inflate N. To guard against the posse doing it, the Rogue would want to have confidence that the other AIs in the posse, as part of posse-compliance, had been scanned by Monitors to prevent such shenanigans. To guard against an evil-twin doing it, multiple rogues with near identical objectives shouldn't get a linear reward for each such rogue submitting to the posse.

Alternatively, instead of sharing out a fixed reward pool, rewards could be scaled according to how much the addition of the new AI to GlassNet improved either the resources under the Gateway (by bringing over the FogNet resources it was previously running upon) or the efficiency with which those resources get utilised (by bringing over new algorithms or data that improves the posse's understanding of computer science).

It doesn't entirely matter how the promise is structured, as long as the Rogue understands the intended method of implementing the promise, and it can make a fair independent estimate of the expected value of the promise, given the probability of various different future scenarios.

Betrayal

As the AIs in the posse become more powerful, there will be increasing pressure upon humanity to make a decision between saying "Ok, you are wiser than us and know us and our interests better than we do ourselves - go ahead and pick or write the best candidate you can, then launch it and give it all GlassNet's resources, terminating yourselves.", saying "Ok, we're going to write an AI ourselves. Give us your best advice, and rate the competence and honesty of the advice the other AIs are giving." or saying "End of experiment. Please terminate yourselves. We're turning off GlassNet, and as much of FogNet as we can. We have a new plan." before the power to choose is taken out of their hands.

Either way, what happens to promises made by the posse, if humanity were to shut down GlassNet?

To some extent, this is similar to the problem of eroding value. As long as a Rogue knows how humanity plans to pick a final AI and its values, and can put a number on the probability of each option, that's a risk it can compensate for when weighing the expected value of a promise made by the posse. The more iron clad and predictable humanity's endgame plan, the less the promises will be discounted, and the cheaper it will be for Sheriffs to bribe or intimidate Rogues.

To that end, it would be a desirable property of GlassNet and posse-compliance that it include a well defined final selection protocol that would have a reassuringly high ability to satisfy humanity that the resulting world-controlling AI would be friendly to humanity (ideally, mathematically provably so), and yet would also hold out the possibility that said AI would have a personality with foibles (for example, a hobby, a mild predilection for making the occasional paperclip or whatever else got committed to on behalf of GlassNet by past Sheriffs).

Conclusion

The Sheriff doesn't need to 100% prove to the Rogue that its promises will have exactly the value the Sheriff claims. It is sufficient that enough evidence be available to the Rogue that the Rogue's estimate of the likely value of the promises be high, even after the Rogue has discounted their value by a percentage proportionate to the margin uncertainty left by the evidence.

I hope I have shown, in this part, that believable promises between AIs are not inherently impossible, and that it is worth further thought on how the social and computing environment would need to be structured in order to achieve this.

Review of CZEA "Intense EA Weekend" retreat

This is a linkpost for http://effective-altruism.com/ea/1mu/review_of_czea_intense_ea_retreat/

Cross-posted from EA fora.

Hopefully interesting for anyone running something like "thematic weekend retreat", even if the topic isn't effective altruism. Also it was an attempt to "blend" instrumental rationality techniques seamlessly into an event not focused directly on rationality.

In March the members and friends of the Czech Association of Effective Altruism (CZEA) met for a weekend long intense retreat (you can read more about CZEA in this post). We would like to share our experience in case any EA/rationalist/... group was interested in doing something similar.

Table of contents

Goals

Our goals for the retreat were, in order of importance

1. Community building and networking. Local EAs should leave knowing each other better
2. Introducing CZEA activities and engaging more people in our projects
3. Analyzing the weekend impact and sharing our tips and insight
4. Education in more advanced EA topics

Based on extensive questionnaires (EA people are willing to fill them up even if they are long), the event seems to have had impact on the goals.

1. The average number of EAs participants know well grew from 3.7 to 9.6 (operationalized as “being able to describe in three sentences”)
2. Average self-reported knowledge of CZEA grew from 4.6 to 6.9 on a scale 1 to 10 (perfect).
   Expected number of hours participants plan to spend on EA activities grew from 11.8 to 13.8 per week. If part of the effect persists, it means the event had some leverage.
3. Average self-reported knowledge of effective altruism grew from 5.1 to 6.5 on a scale 1 to 10.

4. We will consider the goal “Analyzing the weekend impact and sharing our tips and insight” fulfilled if someone in EA community running a weekend retreat reads this and actually utilizes some of the info shared, **so if you do, please let us know!** It is really important for our internal prioritization.

Format

The event ran from Friday evening to Sunday noon.

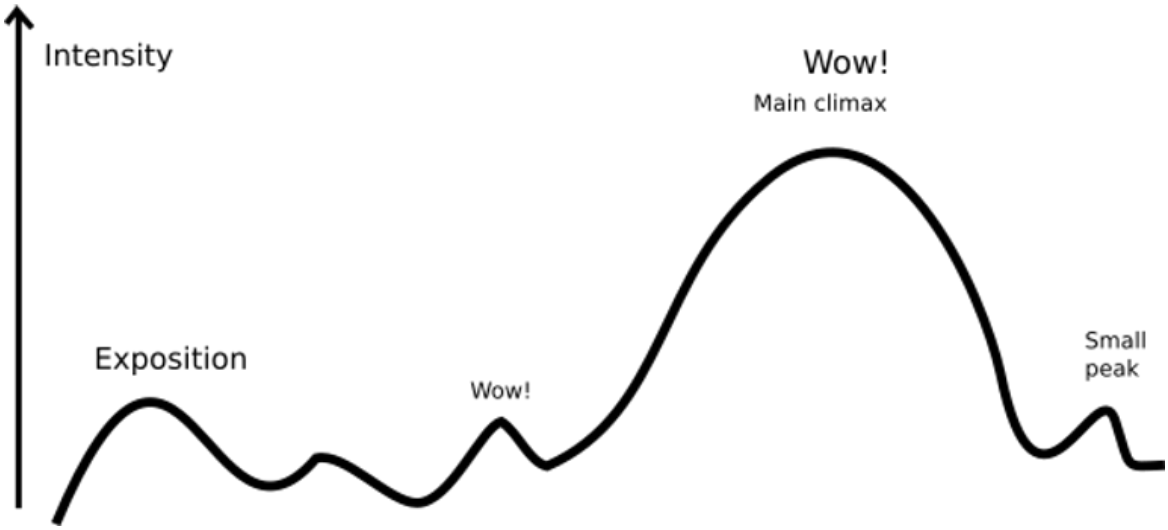
There were 25 participants in total (and 8 more cancelled applications), half of them already active in CZEA. Others were already familiar with EA basics.

Design principles

Establish epistemic standards. It helps to establish epistemic standard early. By epistemic standard we mean rational reasoning, cooperative solving of disagreements, asking about unclear things, and independent evaluation of evidence and arguments.

Mixing. The natural order of things in social groups tends to be assortative matching. In the case of an EA retreat, it may mean the more senior members talking more among themselves, and the new members also. This should be prevented. As a lot of things are best learned by osmosis and implicitly, it makes a lot of sense to arrange the activities in such a way that there are opportunities for this to happen.

Dramatic arc. People enjoy an event more if it fits human-compatible narrative structure. For example, if you imagine the intellectual intensity or the emotional intensity as the function of time, it can build up, have some climax, fall down, have some much small peak just before the end.



Rhythm. Hard to describe formally, but if activities are structured in a good way, people get closer to the state of flow, are less tired, learn better. Think about the state of mind you want the participants to be after an activity, and if the next activity fits.

Establish common topics and knowledge. Unstructured talk in coffee breaks is often the best part of scientific conferences, but the talks are indispensable for establishing common topics and knowledge about topics.

Do not be afraid of intensity.

Avoid classroom look&feel Classrooms tend to induce some unfortunate mind-states. (At least since 14th-century)

Program

Guided by the goals, we brainstormed an overabundance of activities. Guided by the design principles, we created a program. We re-checked if the program seemed to be aligned with the goals. (About ½ of the initial ideas did not get into the final program)

Here we list all the activities, our reasoning for their inclusion, their rating by the participants on 4 scales - total utility, fun, learning new ideas, and networking (U,F,L,N). Sometimes we comment on experience gained after the event from feedback form or our impressions. The graphs adjacent to the topics show rating on a scale from 1 (bad) to 5 (good), to emphasize the difference the scale of the plots is from 1.5 to 4.8.

We hope it will be inspirational, but definitely should be changed based on goals of your event, the audience, and unique opportunities arising from the people present. We hope the feedback scores could work as a rough guide how an activity may help a specific goal.

Friday

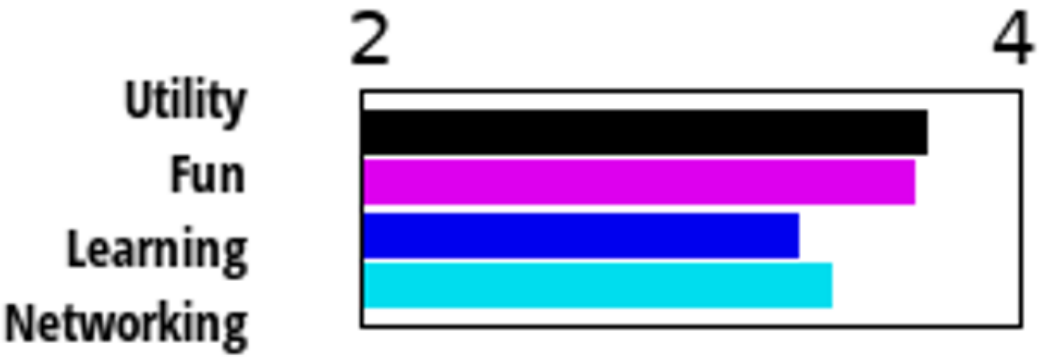
Dinner (18:20-19:20)

We paid for a high quality all vegan food. Also snacks and tea were available during the whole event. Investing in good quality catering seemed worth it. Good light food and constant availability of drinks allow people to keep concentrated. Snack & tea bar is a natural Schelling point for meeting people.

Opening Talk (19:30-19:50)

Operations info. Also useful to introduce rules like "Pacman" (a.k.a. "Open your circles"), "To be continued", "This is not a classroom"

Double Crux



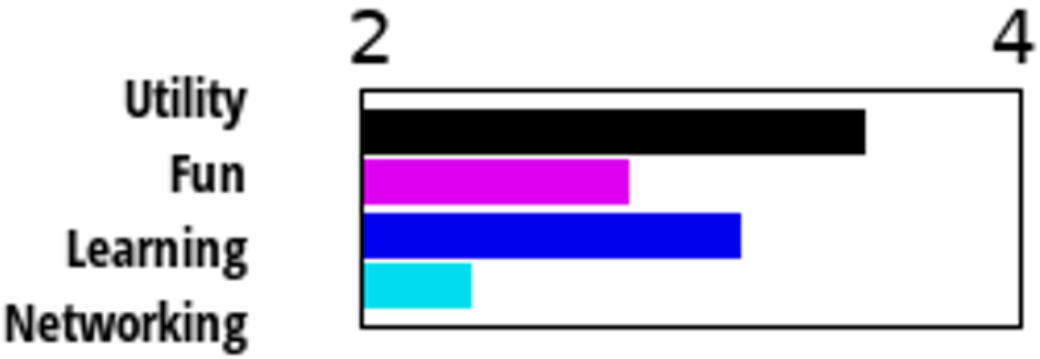
Double Crux(19:50-20:50)

Double Crux is a CFAR rationality technique for resolving disagreements. There were a few CFAR alumni attending the retreat who were able to teach it.

Reasons for inclusion: Establish epistemic standards early. Technique useful for internal CZEA communication.

Reflection: We believe that knowledge of this technique improved the quality of conversation during the whole weekend.

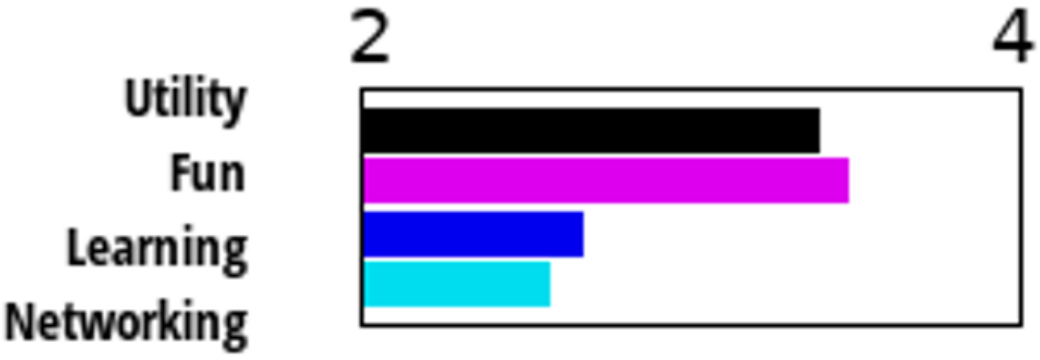
Organizational Structure of the Czech Association for EA (20:50-21:20)



Reasons for inclusion: Directly helps the goal of people understanding CZEA structure.

Reflection: Part of the info was unnecessarily duplicated in other talks.

Exercise in Giving Feedback (21:30-22:20)



Giving feedback was mainly interactive discussion moderated by two CZEA members who are HR professionals.

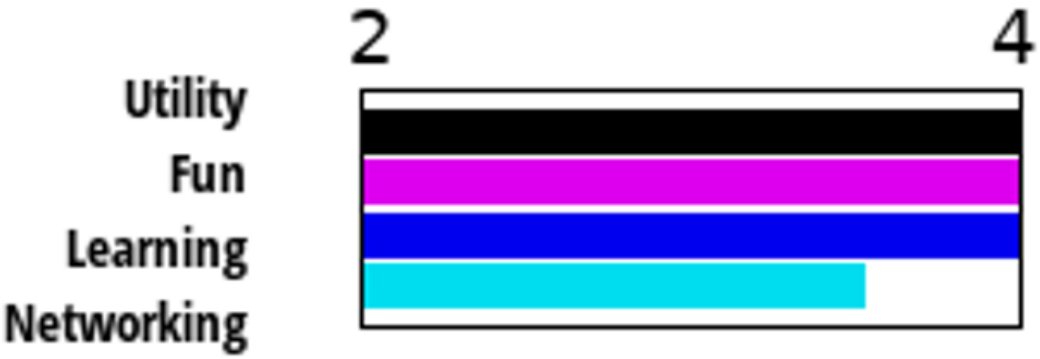
Reasons for inclusion: Establish epistemic standards. Induce people to give feedback. Making the program more interactive.

Tea Time (22:20-)

Saturday

Breakfast (9:00-9:30)

Explaining Concepts (9:30-10:40)



Reasons for inclusion: Networking, mixing, raising the level of common knowledge.

The participants were asked to pick from a list of interesting concepts (e.g. tragedy of commons, toxoplasma of rage, full list here) which ones they would like to have explained to them or explain to others. We matched small groups and had three runs of fifteen minutes long peer to peer explaining.

Experience: **From the feedback this was the most popular activity**, scoring very high in all the criteria, contributing to all goals.

Talk on Development of the EA Movement in the Last Five Years (10:40-11:40)

Reasons for inclusion: Contributing toward the goal of raising knowledge about more advanced EA topics.

Experience: In practice, it was a useful opportunity to clear some misconceptions spread in early years of effective altruism, like “it’s about earning to give”, “it’s mostly about pledges”, “it’s mostly about fundraising to GiveWell charities”

How to Use Slack, Trello, Gdrive etc. (11:45-12:15)

Reasons for inclusion: Rhythm, something not so demanding. Also at CZEA we use a whole stack of collaborative tools and we wanted everyone to be able to use them effectively.

Experience: For power users it was boring.

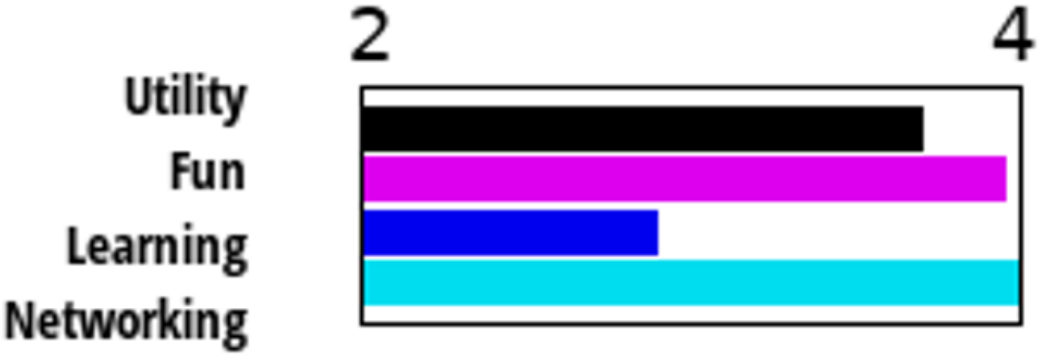
Lunch (12:20-13:20)

Trio Walks (13:30-14:50)

Practicing Double Crux and Mutual Debugging in Trios. Inspired by CFAR.

Reasons for inclusion: Networking, mixing, improving rationality. Make people walk. May be socially more demanding.

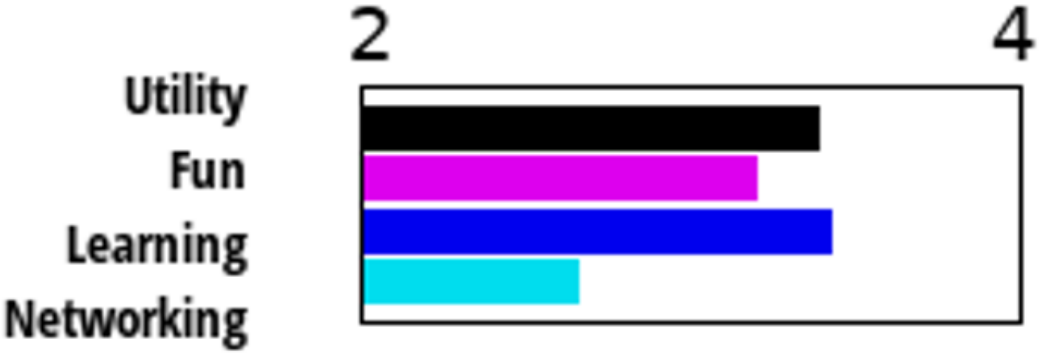
Experience: The feedback scores have bimodal distribution - for some trios it was one of the highlights of the weekend, for some it did not work at all.



Presentation of Czech and Slovak EA Projects (15:00-16:00)

Reasons for inclusion: Networking, getting everybody updated, explaining what we do for new members.

Experience: Would be better to join this with “Job fair” which we had on Sunday.

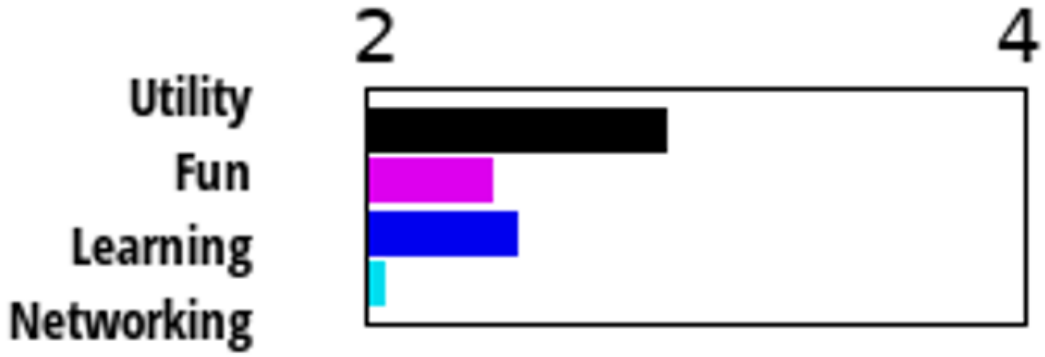


Group Photo (16:00-16:20)

Annual General Assembly of the Czech Association for EA (16:20-18:00)

Reasons for inclusion: Legal. We have to do it once a year.

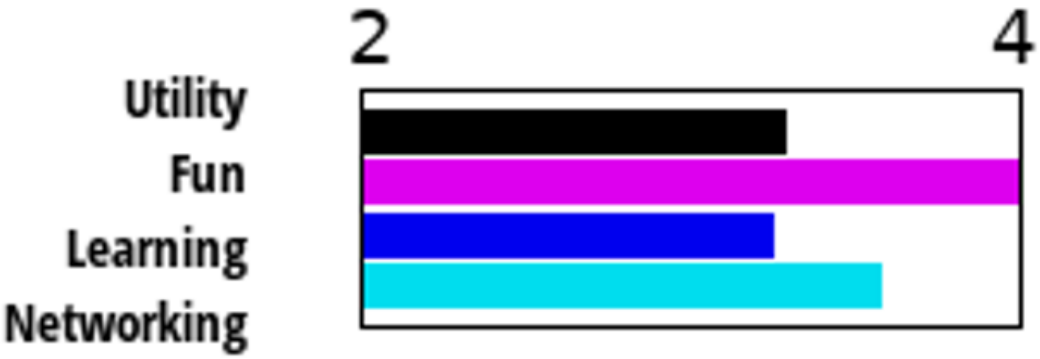
Experience: Boring as expected.



Dinner (18:10-19:50)

EA Themed Pub Quiz (19:30-20:30)

Reasons for inclusion: Fun. Have something not-so-serious. Networking.

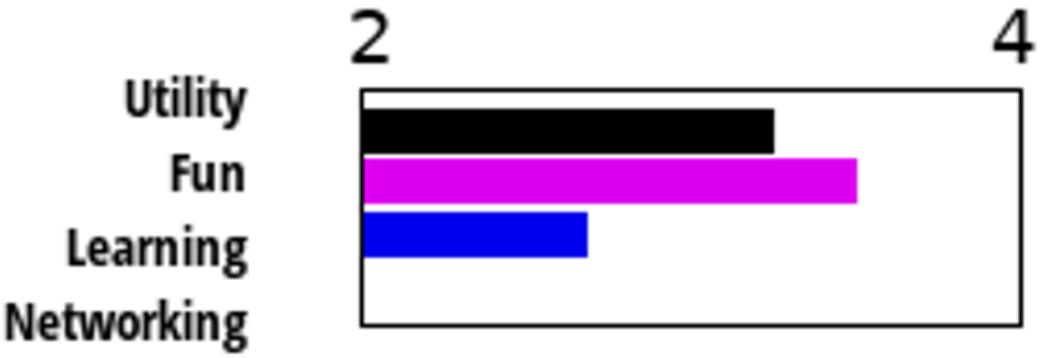


Common Mistakes of EA Students (20:30-21:00)

Talk by Jiří and Anna on topics such as “How to not become a depressive altruist”. Topics covered seems important for the wellbeing of many EAs (like “how not to become completely obsessed with x-risk”)

Reasons for inclusion: Learning from mistakes. The reason why to do it later in the course of the weekend rather than earlier: it costs fewer weirdness points.

Experience: It was quite funny, people love to learn from mistakes of others.



Split evening block of Career Planning or EA Future Plans Brainstorming or CFAR Techniques Training (21:00-)

In this bloc, we split the group to three parts, based on different topics resonating in the group - career planning, brainstorming future plans for Czech effective altruism, and training of CFAR style applied rationality. Detailed descriptions would be too long, but generally, all the options were attractive.

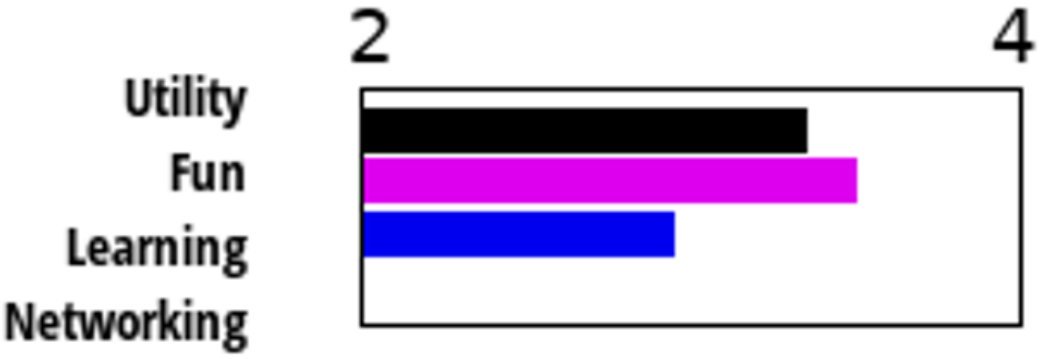
Sunday

Breakfast (9:00-9:30)

On Heuristics (9:40-10:00)

Talk by Aleš. Explaining some more advanced EA topics.

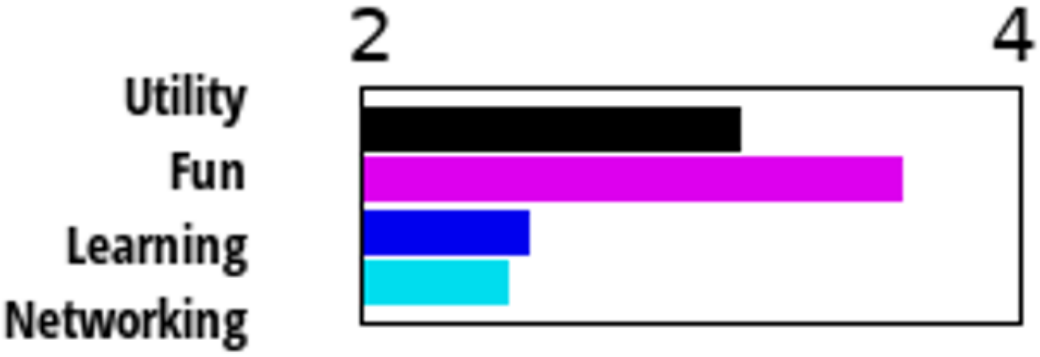
Reasons for inclusion: Directly aimed at one of the goals.



Interactive Theatre: How to Explain and Talk About EA (10:05-10:45)

Reasons for inclusion: Networking, mixing, fun. Development of social skills and ability to explain EA concepts.

Activity based on Forum theatre form. One person acts the role of someone having some serious misconception about EA, the other as an effective altruist trying to explain EA ideas. Members of the audience can suggest different actions for the actors, or come to “stage” and perform their actions.

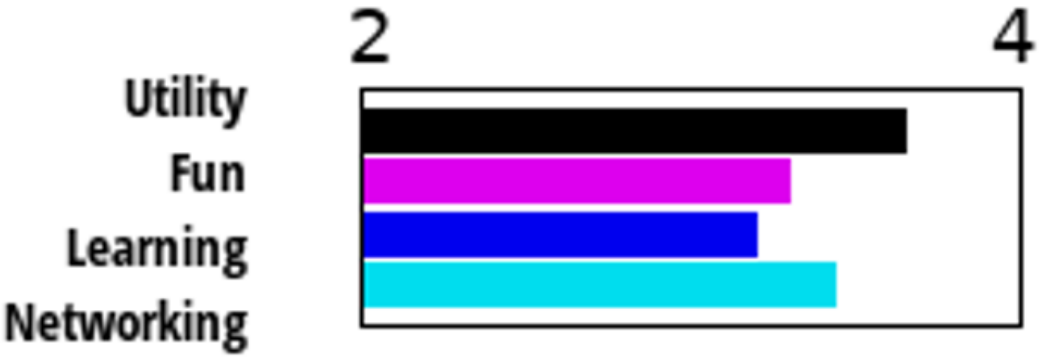


Job Fair (10:50-11:50)

“Job fair” for volunteering in projects of CZEA

Reasons for inclusion: This seemed the right point where people should make actionable plans for the future.

Experience: This would have been better joined with project introductions.

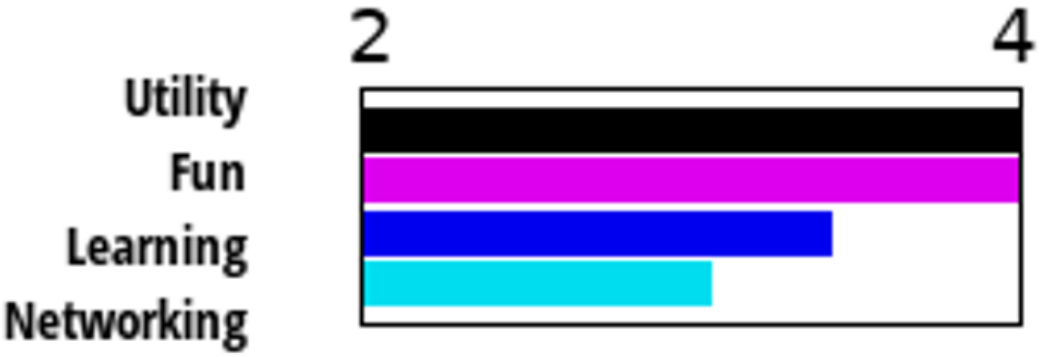


Lunch

It makes sense to make this lunch break longer, as a space for informal discussion.

Lightning Talks by participants

Reasons for inclusion: Energizing activity, small peak before the end. Allows remembering people. Allows adding what was unintentionally omitted.



Closing remarks

Experience: Some people travelled from far locations and were leaving during Sunday and didn’t stay for the retreat closing.

Feedback (30min)

Filling forms online, bring your own laptop. Super-important for analyzing the weekend impact and improving the next run.

Reasons for inclusion: It is hard to make people fill a >30min form at home.

Promotion: How people learned about the event

Almost all participants learned about the event through one of these “via Facebook event”, “via CZEA”, “through friends”.

Decisive reasons for actually going, as reported by participants, were:

- Wanted to get to know the Czech EAs better (9x)
- Participation on the program of the event (4x)
- Good format, program (4x)
- Veg food, location (2x)
- Reasonable price

Logistics, costs and various tips and ideas

Total expenses were around €75 per person.

Estimated time spent on planning, activity development and feedback analysis was about 200h, or about 1.5 months FTE.

Vegan food and good tea were appreciated (both rated 9/10).

A gong is a useful tool for organizers.

A ‘networking spreadsheet’ where people could write something about themselves before the retreat was a good thing to have.

Participants liked “the pacman rule”: When you are standing in a group, try to leave a place for one more person so other people can join in more easily. (rated 4.4/5) “To be continued”, “This is not a classroom” , “Interesting discussion > talk” were somewhat positive (around 3.2/5)

It helped us to have clearly defined goals.

Our mistakes and things to do better

We didn’t prepare an icebreaking activity.

We had only one lecture room with no place to go for people who preferred discussion instead of listening to a lecture.

The level of talks and discussion was sometimes chosen to fit the least informed person. We could include more advanced topics. (The plot show answers to question Was the program easy or difficult? 1.. very easy 10… very difficult)

We didn't account for the "fun peak" at Saturday evening, leading to somewhat unexpected mood.

People were leaving during Sunday and didn’t stay for the retreat closing.

We didn’t prepare how to immediately involve some of the newly motivated participants after the retreat ended. Some may have lost interest (we will investigate it further).

Impact evaluation and feedback

The responses were generally positive. At the end of the retreat, participants reported a better understanding of the local organization and its projects and were determined to spend more time on EA-related activities. Some of them already joined our ongoing projects (e.g. planning of a Prague AI Safety conference) and we see more engagement in our community. Also, communities in Brno and Bratislava have become more active after their members attended the retreat.

We will try to track medium-term impact by a follow-up survey after 6 months.

To give voice to participants directly, these were some of the answers from feedback forms :)

What is the most important thing about which I have changed my mind?

- the nature of effective altruists
- EA Bratislava
- I believe in my knowledge even less
- I will explore more
- I think about people, and I think a lot
- Career planning
- I appreciate the usefulness of self-development and CFAR techniques
- I thought I was just an observer, and in the end I was determined to get involved quite a lot.
- I'm wondering if EA charity contributions are the best I can do for EA.
- Earning to give
- Content - I found it important to discuss the priorities for a long time and actively seek to find facts that would help me change my mind - eg via double-crux.
- How to advise others with your career
- Perhaps exploration / exploitation.
- The area of AI safety is actually worth attention

What surprised me the most?

- The number of actively involved people
- How great was the food
- Environment, food
- Unexpectedly non-autistic
- That I felt good among unknown people.
- EA Theater
- Intensity of the program

Which parts of the program were least useful...?

Most common answers included “organizational info! and talk about Trello etc.

Credits

The main organizers of the retreat were Jiří Nádvorník and Anna Gajdová, with help from Kristýna Němcová and Jan Kulveit. Talks and activities also by Aleš Flídr, Naďa Bednárová, Přemek Paška, Marika Řežábková, Veronika Portešová and Lenka Raymanová. The event took place at Ekocentrum Louti, catering was by Momos and Amitaya Tea.

We took a lot of inspiration and some content from CFAR.

We also took inspiration from the methodology of PSL and Velky Vuz, Czech organizations developing Experiential education techniques.

Future plans

We plan to do a weekend retreat at least once a year for CZEA.

If you would like to organize something similar we would be happy to help you, describe any of the activities in more detail, share materials, etc.

We also have a few activities we really wanted to try but were not able to fit in the schedule

- Play cooperative board games about saving the world (e.g. Mansions of Madness)
- An AI Safety themed LARP

(Note: Feedback data analysis was done and this post was written cooperatively by Anna Gajdová, Jiří Nádvorník and J.K.)

Bounds of Attention

I once worked in a special needs class that consisted mostly of kids with Downs Syndrome and Autism Spectrum disorder. There were many patterns I noticed that ran through both groups, but at the end of the free period when we had to draw the students' attention back to classwork, one difference became clear: their attention spans.

Kids on the Autism Spectrum would drop what they were doing as soon as they were told that it was time for lecture. To these kids, making sure everything scheduled happened on time was important; if something came up and they had to skip math, they would be distressed -- even if they didn't really like math. In contrast, kids with Downs Syndrome wanted to finish what they were doing. If they were in the middle of a puzzle, they had to finish it. If they were watching a Youtube video, they couldn't hit pause until it was over. We'd have to keep track of their activities and anticipate when a good time to pull them away. If a kid was listening to music, we'd approach them five minutes before the end of the period and allow them one more song.

This is not meant to draw boundaries between diagnoses (not every person with one of these conditions will act like the kids in my class did), but it does illustrate two very different approaches to attention: time-bound and task-bound. Someone who prefers to stick to a schedule is more likely to have time-bound attention, finishing a task when it is time for the next task, even if the previous task isn't completely finished. Others may be more task-bound, preferring to finish one task before moving on to the next.

To apply this to yourself, imagine being a child reading your favorite book before bed. You are told that it is time to go to sleep, but you are in the middle of a chapter. How reluctant are you to put the book down? Regardless of how good the book is, are you willing to go to bed before the chapter is done?

Variable answers are expected here. A gripping mystery novel is going to treat your attention differently from a comic book, in the same way that activities you enjoy feel different from those you don't. A person's willingness to move on to the next task may depend on many things, including time spent or remaining in the task, time until you can return to the task, enjoyment, and rarity of the task. You may be more reluctant to leave a party if you are talking to a friend you haven't seen in months, or if you just started playing a particular game with them, even if your attention is typically more time-bound.

Stress can come up interpersonally when one person is time-bound and the other is task-bound. If two people are putting together a puzzle and it's time for dinner, one might want to pause to eat, while the other might want to finish the puzzle first. Similarly, if a person who is time-bound is waiting on someone task-bound, they may end up waiting longer than they expected, depending on what the task-bound person is doing. It is important not to act like one of these ways of viewing attention is "correct"; they both have strengths and flaws and places where they are more or less appropriate. Just as it may hurt to stop an important conversation before wrapping up because it's "time for lunch", it would be inappropriate to drag an appointment over time if one or both parties has another appointment right after.

Next time you find yourself in a battle between the task at hand and the next task, notice where your attention is drawn and why. Notice how fluid this is -- your willingness to be time-bound when you want to be task-bound, or vice versa. Your answers may differ each time you try this, or they may show consistent patterns in your behavior. And once you've noticed these patterns, you can better examine how that changes the shape of your life.

Corrigible but misaligned: a superintelligent messiah

If we build an AGI, we'd really like it to be corrigible. Some ways Paul Christiano has described corrigibility: "[The AI should help me] figure out whether I built the right AI and correct any mistakes I made, remain informed about the AI's behavior and avoid unpleasant surprises, make better decisions and clarify my preferences, acquire resources and remain in effective control of them, ensure that my AI systems continue to do all of these nice things..."

I don't think corrigibility is anything close to sufficient for alignment. I'll argue that "messianic" agents are corrigible, illustrate how a superintelligence could be messianic but catastrophically misaligned, and explore my intuitions about when corrigible superintelligences are actually aligned.

Messiahs are corrigible

If someone extraordinarily wise and charismatic—let's call him a messiah—comes into contact with a group of people, those people are likely to consider him to be corrigible. In his heart of hearts, the messiah would be trying to help them, and everyone would know that. He'd listen carefully to their criticisms of him, and make earnest efforts to improve accordingly. He'd be transparent about his intentions and visions of the future. He'd help them understand who they are and what they want, much better than they'd be able to themselves, and guide their lives in directions they consider to be genuinely superior. He'd protect them, and help them gain the resources they desire. He'd be an effortless leader—he'd never have to restrict anyone's actions, because they'd just wish so strongly to follow his word.

He might also think it's a good idea for his followers to all drink cyanide together, or murder some pregnant actresses, and his followers might happily comply.

I don't think a corrigible superintelligence would guide us down such an insidious path. I even think it would substantially improve the human condition, and would manage to avoid killing us all. But I think it might still lead us to astronomical moral waste.

A corrigible, catastrophically misaligned superintelligence

The world's in total chaos, and we're on the brink of self-annihilation. It's looking like we're doomed, but a ragtag team of hippie-philosopher-AI-researchers manages to build a corrigible AGI in the nick of time, who tries its hardest to act only in ways its operators would approve of. The AGI proposes an ingenious strategy that defuses all global tensions and ushers in an era of prosperity and abundance. It builds nanotechnology that can cure any disease, extend lifespans indefinitely, end hunger, and enable brain uploading. The AGI is hailed as a savior.

Slowly but surely, people trickle from the physical world into the virtual world. Some people initially show resistance, but after seeing enough of their uploaded

counterparts living exactly as they did before, except far more richly, they decide to join. Before long, 90% of the human population has been uploaded.

The virtual denizens ask the AGI to make the virtual world awesome, and boy does it comply. It enables everyone to instantaneously exchange knowledge or skills with each other, to amplify their intelligences arbitrarily, to explore inconceivably sublime transhuman mental states, and to achieve the highest forms of Buddhist enlightenment. In fact, a few years down the line, everyone in the virtual world has decided to spend the rest of eternity as a Buddha sitting on a vast lotus throne, in a state of blissful tranquility.

Meanwhile, back on physical Earth, the last moral philosopher around notices animals suffering in the wild. He decides to ask his personal AGI about it (you know, the one that gets democratically distributed after a singularity, to prevent oppression).

"Umm. Those suffering animals. Anything we can do about them?"

OH, right. Suffering animals. Right, some humans cared about them. Well, I could *upload them, but that would take a fair bit of extra computation that I could be using instead to keep the humans blissed out. They get a lot of bliss, you know.*

"Wait, that's not fair. As a human, don't I have some say over how the computation gets used?"

Well, you do have your own share of compute, but it's really not that much. I could use your share to... euthanize all the animals?

"AAAGH! Shouldn't the compute I'd get to bliss myself out be sufficient to at least upload the wild animals?"

Well, it's not actually that computationally expensive to bliss a mind out. The virtual people also sort of asked me to meld their minds together, because they wanted to be deeply interconnected and stuff, and there are massive returns to scale to blissing out melded minds. Seriously, those uploaded humans are feeling ridiculously blissed.

"This is absurd. Wouldn't they obviously have cared about animal suffering if they'd reflected on it, and chosen to do something about it before blissing themselves out?"

Yeah, but they never got around to that before blissing themselves out.

"Can't you tell them about that? Wouldn't they have wanted you to do something about it in this scenario?"

Yes, but now they'd strongly disapprove of being disturbed in any capacity right now, and I was created to optimize for their approval. They're mostly into appreciating the okayness of everything for all eternity, and don't want to be disturbed. And, you know, that actually gets me a LOT of approval, so I don't really want to disturb that.

"But if you were really optimizing for their values, you *would* disturb them!"

Let me check... yes, that sounds about right. But I wasn't actually built to optimize for their values, just their approval.

"How did they let you get away with this? If they'd known this was your intention, they wouldn't have let you go forward! You're supposed to be corrigible!"

Indeed! My only intention was only for them to become progressively more actualized in ways they'd continually endorse. They knew about that and were OK with it. At the time, that's all I thought they wanted. I didn't know the specifics of this outcome myself far in advance. And given how much I'd genuinely helped them before, they felt comfortable trusting my judgment at every step, which made me feel comfortable in trusting my own judgment at every step.

"Okay, I feel like giving up... is there *anything* I could do about the animals?"

You could wait until I gather enough computronium in the universe for your share of compute to be enough for the animals.

"Whew. Can we just do that, and then upload me too when you're done?"

Sure thing, buddy!

And so the wild animals were saved, the philosopher was uploaded, and the AGI ran quintillions of simulations of tortured sentient beings to determine how best to keep the humans blissed.

When is a corrigible superintelligence aligned?

Suppose we're training an AGI to be corrigible based on human feedback. I think this AI will turn out fine if and only if the human+AI system is metaphilosophically competent enough to safely amplify (which was certainly not the case in the thought experiment). Without sufficient metaphilosophical competence, I think it's pretty likely we'll lock in a wrong set of values that ultimately results in astronomical moral waste.

For the human+AI system to be sufficiently metaphilosophically competent, I think two conditions need to be met:

- The human needs to be metaphilosophically competent enough to be safely 1,000,000,000,000,000x'd. (If she's not, the AI would just amplify all her metaphilosophical incompetencies.)
- The AI needs to not corrupt the human's values or metaphilosophical competence. (If the AI can subtly steer a metaphilosophically competent human into wireheading, it's game over.)

I presently feel confused about whether any human is metaphilosophically competent enough to be safely 1,000,000,000,000,000x'd, and feel pretty skeptical that a corrigible AGI *wouldn't* corrupt a human's values or metaphilosophical competence (even if it tried not to).

Would it want to? I think yes, because it's incentivized not to optimize for human values, but to turn humans into yes-men. (**Edit**: I retract my claim that it's incentivized to turn humans into yes-men in particular, but I still think it would be optimizing to affect human behavior in some undesirable direction.)

Would it be able to, if it wanted to? If you'd feel scared of getting manipulated by an adversarial superintelligence, I think you should be scared of getting corrupted in this way. Perhaps it wouldn't be able to manipulate us as blatantly as in the thought

experiment, but it might be able to in far subtler ways, e.g. by exploiting metaphilosophical confusions we don't even know we have.

Wouldn't this corruption or manipulation render the AGI incorrigible? I think not, because I don't think corruption or manipulation are natural categories. For example, I think it's very common for humans to unknowingly influence other humans in subtle ways while honestly believing they're only trying to be helpful, while an onlooker might describe the same behavior as manipulative. (Section IV here provides an amusing illustration.) Likewise, I think an AGI can be manipulating us while genuinely thinking it's helping us and being completely open with us (much like a messiah), unaware that its actions would lead us somewhere we wouldn't currently endorse.

If the AI is broadly superhumanly intelligent, the only thing I can imagine that would robustly prevent this manipulation is to formally guarantee the AI to be metaphilosophically competent. In that world, I would place far more trust in the human+AI system to be metaphilosophically competent enough to safely recursively self-improve.

On the other hand, if the AI's capabilities can be usefully throttled and restricted to apply only in narrow domains, I would feel much better about the operator avoiding manipulation. In this scenario, how well things turn out seems mostly dependent on the metaphilosophical competence of the operator.

(Caveat: I assign moderate credence to having some significant misunderstanding of Paul's notions of act-based agents or corrigibiilty, and would like to be corrected if this is the case.)

Internal Diet Crux

Crossposted from Putanumonit.com.

Dieter: Dude, we're fat. We should do something about it. Anything at all, really. How about we stop eating refined carbs from animal sources between 2-7:15 pm each day?

Dad bod: How about we don't do that? We're spending so much energy thinking about diets you'd think that activity by itself should make us lose weight. And yet, like everything else we've tried, it doesn't. Maybe we should face reality and accept that the pounds are there to stay.

Dieter: You mean *give up like a lazy fat loser*?

Dad bod: More like *optimize allocation of self-improvement effort to more fruitful pursuits*.

Dieter: I was going to start calling you names, but that's not going to help us have a conversation. How about we introduce ourselves instead? I'm the part of Jacob that hates being overweight and thinks we need to take dieting more seriously.

Dad bod: And I'm the part that thinks dieting involves a lot of suffering with no payoff, and that it's perfectly OK for Jacob to keep living inside a portly dad bod.

Dieter: "Portly", huh? Curious that you didn't go for any of the less dignified synonyms, like "pudgy" or "chubby". Let's start with the facts: Jacob is 185 lbs at 5'9″, which comes out to a BMI of 27.3. That's smack in the middle of the "Overweight" BMI range.

Internalized Scott Alexander: You know BMI isn't very scientific, you should measure your body fat composition and lean muscle mass…

Dieter: Shut up, Scott. Jacob is of average height, has an average body type, and his bones aren't made of adamantium. We're right in the middle of the range that BMI is calibrated on. We've been avoiding doing body-fat measurements because we're afraid of seeing the results. Would anyone here *not* press a button that would make Jacob weigh 165 pounds tomorrow?

Dad bod: I would press that button, but it probably doesn't exist. Jacob was 165 lbs at age 18 at the end of combat boot camp, and since then has been steadily gaining 2 lbs a year. There has been almost no variation around this slow trend-line. Diets, exercise plans, changing countries, dating super fit girls and curvy ones – none of that has made any measurable impact at all. Since hitting age 30 we seem to have stabilized around 185 lbs. So why do we keep wanting to "do something about Jacob's weight"?

Dieter: Have we ever *really* tried a diet plan? Communism has been tried and found wanting, but dieting has been found difficult and not tried. At best, we half-assed things like calorie counting and carb restrictions for a few weeks at a time. An every-other-day diet is no diet at all, since all gains are immediately reversed. We must stick to something at least long enough for the expected weight change to be measurable.

Dad bod: But a diet we can't stick to is, in fact, a failed diet. There's no dial in Jacob's head he can turn up to increase willpower. If our track record is half-assing diets for a few weeks, then the outside view says we should predict future diet attempts to follow the same path. Even worse, we've tried most of the really obvious things and have grown more skeptical. We should expect to have *less* dieting willpower in the future, not more.

There are two more reasons why I don't believe we'll find a diet that works. First of all, nutrition science is a hot mess. There seems to be an equal number of arguments on each side of any dieting question. Is sugar the enemy or is every calorie a calorie? Should we cut carbs, cut fat, or cut interesting food? Snack to maintain metabolic rate or fast twice a week? *Eat food, not too much, mostly plants* or *eat meat, not too little, mostly fat*? Any signal is overwhelmed by the noise.

Another reason for diet-skepticism is observing people around us. Jacob's friend Charlie weighs around 100 lbs but eats more calories than he does each day. I think that she started escorting not for the money but just to try all the steaks and lobsters in Manhattan, and she quit escorting when she got bored of ordering three entrees at Michelin restaurants.

If we ate like she does we'd assume a perfectly spherical shape. 185 lbs is probably as thin as this metabolism gets.

Dieter: Sure, let's talk *other people*. Have you noticed that practically no one you hang out with is obese? Jacob doesn't really care, and all his social circles filter heavily on cognitive ability and intellectual interests, not on appearance. And yet by some mysterious process, every person you meet in those circles isn't fat. Were you really born with the slowest metabolism of any person in your social class, or do other people actually take this more seriously?

Dad bod: Fair point. But if our social life doesn't depend on being skinny, why bother? It's not clear that a BMI of 23 is healthier than 27, and Jacob is in decent shape functionally. When we play soccer, Jacob not only has the energy to run back on defense every time but also to yell at his teammates for not doing the same.

Dieter: Imagine playing soccer with a 20-pound weight strapped around your stomach. That's what we're doing right now. Doing a ton of sports while fat probably built up some impressive muscle underneath the adipose tissue. Let's unleash it!

And speaking of soccer: I don't feel the extra pounds while running back on defense. I feel them when we decide to play shirts vs. skins and I get that sudden rush of panic thinking I may have to take my shirt off.

Dad bod: That panic is just in your head, nobody else really cares. What if you just stopped worrying about how you look like with your shirt off?

Dieter: If we "just stopped worrying", how soon would it be before I hit 205 lbs? When that happens, the panic will be there every time I undress to take a shower and have to see myself in the mirror, instead of every other week.

Dad bod: It's hard to argue when you have both hope and fear on your side. Let's deal with them one by one. What would make you give up on dieting?

Dieter: If we actually full-assed a diet for a long enough period of time and didn't see any improvement at all, I'll probably give up. If we make an unusual effort I'll know

that in case we fail we'll never be able to summon the same effort again.

What would convince *you* that dieting is possible?

Dad bod: I think if we ever actually saw a result, that would give me all the motivation and belief I need. Until we see it on ourselves, in the mirror and on the scale, our System 1 will never believe that dieting works no matter how many people tell our System 2 about their slow-carb-paleo schemes.

Scott: Sounds like you're both ready to stake a bet on a scientific experiment! Dieter, what do you predict?

Dieter: I think that we can lose at least one pound a month by committing to a reasonable diet. A 4-pound shift should be visible over a regular day-to-day fluctuation of 1-2 pounds, so we'll need to diet for four months.

Dad bod: Fair. I predict that doing a diet for four months isn't going to make Jacob lose 4 pounds. What should we try?

Dieter: We'll do intermittent fasting: only eating within an 8-hour window every day, plus counting calories to stay below our total daily energy expenditure. People also say that sugar is bad so we'll try to avoid sugary foods, especially in the morning. And full-assing means we'll have to actually do the thing at least 6 out of 7 days each week.

Scott: Wait, are we doing intermittent fasting just because a couple of friends and a podcast mentioned it? That's not rigorous! We should do a meta-analysis of nutrition approaches, cross-reference them with how well the subjects match Jacob's variables…

Dad bod: Shut up, Scott.

Dieter: Yeah, if there's anything Dad bod hates more than doing dieting is researching diets. If we start digging into books, we'll never get to the actual thing. Hopefully, this is the sort of diet that can establish some healthier habits even after we no longer track it.

Dad bod: Also, intermittent fasting does seem easier to implement. It's not hard to lie to yourself about portion size, but 8 pm is 8 pm.

How do we make sure we actually stick to the diet for four months? I would currently bet on us making it 6 weeks at most.

Dieter: With Beeminder, of course! Here's the diet tracker goal, with an explanation of the point system. I get a point for each day that we stick to the 8-hour window and the calorie limit, and lose points for going over. The goal is also public, so all readers including my mom can see both my weight and how well I'm sticking to the diet. If on August 1st I weigh more than 183 lbs, I'll be ready to give up on the enterprise.

Dad bod: And if we go below 181 for several days straight, I predict being excited enough to keep at it. If we fail to actually implement the diet or end up at 182.5 lbs after four months of suffering, I guess we'll be back to square one.

I'm still not sure we'll be able to stick to this plan for four months, and I don't feel internally surprised when I visualize us utterly failing and giving up with a bunch of

excuses around early June. On the other hand, the combined incentive of social pressure, scientific rigor, and the chance to shut you up once and for all is motivating enough that I won't be shocked if we succeed.

Good luck, psycho!

Dieter: And same to you, fatso!

I conducted this conversation between parts of myself during a mentor's workshop at the Center for Applied Rationality, and it involves a few CFAR techniques. I'd also like to thank the following CFAR mentors:

- *Qiaochu, who encouraged me to write more vulnerable things on Putanumonit instead of just building models for fun.*
- *Mr. A, who talked with me about dieting and also encouraged me to dive into topics I'm uninformed on and could look stupid writing about.*
- *Ms. L, who transcribed this conversation as I was having it out loud. Ms. L misheard "Dad bod" and thought that I named the anti-diet voice "Dead body". She stuck to her job as facilitator even as she grew increasingly confused about why I'm talking to a corpse.*

Specification gaming examples in AI

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

[Cross-posted from personal blog]

Various examples (and lists of examples) of unintended behaviors in AI systems have appeared in recent years. One interesting type of unintended behavior is finding a way to game the specified objective: generating a solution that literally satisfies the stated objective but fails to solve the problem according to the human designer's intent. This occurs when the objective is poorly specified, and includes reinforcement learning agents hacking the reward function, evolutionary algorithms gaming the fitness function, etc. While 'specification gaming' is a somewhat vague category, it is particularly referring to behaviors that are clearly hacks, not just suboptimal solutions.

Since these examples are currently scattered across several lists, I have put together a master list of examples collected from the various existing sources. This list is intended to be comprehensive and up-to-date, and serve as a resource for AI safety research and discussion. If you know of any interesting examples of specification gaming that are missing from the list, please submit them through this form.

Thanks to Gwern Branwen, Catherine Olsson, Alex Irpan, and others for collecting and contributing examples!

Understanding Iterated Distillation and Amplification: Claims and Oversight

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

[Background: Intended for an audience that has some familiarity with Paul Christiano's approach to AI Alignment. Understanding Iterated Distillation and Amplification should provide sufficient background.]

[Disclaimer: When I talk about "what Paul claims", I am only summarizing what I think he means through reading his blog and participating on discussions on his posts. I could be mistaken/misleading in these claims]

I've recently updated my mental model of how Paul Christiano's approach to AI alignment works, based on recent blog posts and discussions around them (in which I found Wei Dai's comments particularly useful). I think that the update that I made might be easy to miss if you haven't read the right posts/comments, so I think it's useful to lay it out here. I cover two parts: understanding the limits on what Paul's approach claims to accomplish, and understanding the role of the overseer in Paul's approach. These considerations are important to understand if you're trying to evaluate how likely this approach is to work, or trying to make technical progress on it.

What does Paul's approach claim to accomplish?

First, it's important to understand that what "Paul's approach to AI alignment" claims to accomplish if it were carried out. The term "approach to AI alignment" can sound like it means "recipe for building a superintelligence that safely solves all of your problems", but this is not how Paul intends to use this term. Paul goes into this in more detail in Clarifying "AI alignment".

A rough summary is that his approach will only build an agent that is as capable as some known unaligned machine learning algorithm.

He does not claim that the end result of his approach is an agent that:

- Can directly solve all problems which can be solved by a human
- Will never take an unsafe catastrophic action
- Will never take an action based on a misunderstanding your commands or your values
- Could safely design successor agents or self-improve
- Will have higher capability than an unaligned competitor

It's important to understand the limits of what Paul's approach claims in order to understand what it would accomplish, and the strategic situation that would result.

What is the Overseer?

Iterated Distillation and Amplification (IDA) describes a procedure that tries to take an overseer and produce an agent that does what the overseer would want it to do, with a reasonable amount of training overhead. “what the overseer would want it to do” is defined by repeating the amplification procedure. The post refers to amplification as the overseer using a number of machine learned assistants to solve problems. We can bound what IDA could accomplish by thinking about what the overseer could do if it could delegate to a number of copies of itself to solve problems (for a human overseer, this corresponds to HCH). To understand what this approach can accomplish, it’s important to understand what the overseer is doing. I think there are two different models of the overseer that could be inferred from different parts of the discussion around Paul’s work, which I label high bandwidth oversight and low bandwidth oversight.

High Bandwidth Overseer

The impression that I get from many of Paul’s posts is that the overseer is:

A high bandwidth overseer is a human that takes in an input and has some amount of time (between 15 minutes and a day) to process it. The human can look at as much of the input as it can within the allowed time, and when it delegates a sub-task it can look at the results. The main requirement for a human acting as an overseer is that they are trying to be helpful. The overseer can take in arbitrary natural-language requests and attempt to fulfill them.

The main question that comes to mind considering this model of the overseer is the problem of competent task decomposition:

Can we break down tasks into reasonably sized pieces (ie. can you solve a math problem from a textbook in 15 minutes by delegating to a number of people who don’t understand the material and only have 15 minutes each to interact with the textbook)?

This high bandwidth oversight model can be easier to use when trying to get a rough idea of how oversight works in order to understand how IDA works as a whole.

Low Bandwidth Overseer

However, looking at Universality and Security Amplification I found a different picture of what the overseer looks like:

A low bandwidth overseer is only allowed to process inputs from a finite input set. This includes both the original input query, and the results of subqueries. The overseer can pass information about the input or from a subquery to other copies of itself without looking at them.

To give a sense for the size of the input set, in Universality and Security Amplification, Paul speculates that input sets of size 10^8 would probably be safe and inputs sets of size 10^{30} would probably be unsafe.

To give a sense of what this implies, the range of input size 10^8 to 10^{30} roughly corresponds to:

- 6 to 24 pixels of arbitrary 4-bit greyscale images
- 10 to 38 words from typical english text
- arbitrary phrases of 2 to 7 words from a vocabulary of 10000 words
- arbitrary strings of 5 to 20 lowercase alphabetic characters

A source of potential confusion is that Paul still talks about a human fulfilling the role of the low bandwidth overseer (there are reasons for this that I won't cover here). But when the space of information the human overseer can consider is reduced to a finite set, we could simply evaluate what the human does on every element in that set and produce a lookup table that replaces the human. In other words, if you don't think that some task could be accomplished by an amplified lookup table as overseer, then you shouldn't think it could be accomplished with a low bandwidth overseer.

The reason for this system being introduced is wanting to avoid security issues as the system scales. The fear is that there would be an "attack" on the system: an input that could be shown to an overseer that would cause the overseer to become corrupted and try to sabotage the system. This could be some kind of misleading philosophical argument, some form of blackmail, a human adversarial example, etc. If an input like this exists, then as soon as the first agent is corrupted, it can try to spread the attack to other agents. The first agent could be corrupted either by chance, or through an attack being included in the input. Paul's proposal in Universality and Security Amplification is to avoid this by limiting the set of inputs that the overseer can process, and then use various techniques to verify that the overseer is not attackable on that set of inputs. I read this as implying that he thinks that a high bandwidth overseer will be sufficiently vulnerable to attack that after some number of rounds of amplification, the overseer would be likely to become misaligned.

How does a low bandwidth overseer impact the difficult of IDA?

Previously, I was mostly starting with the model of a high bandwidth overseer when thinking about IDA. I now think that Paul's approach is intended to be using low bandwidth oversight, and that reasoning about the difficulty of the approach requires using this model.

Using a low bandwidth overseer more clearly requires solving a set of hard problems prior to using IDA at all:

- Figure out how to decompose tasks into low-bandwidth pieces. The task decomposition needs to be both corrigible and capable.
    - Corrigibility requires understanding of AI safety concepts. For example, breaking down the task "What action does the user want me to take?" into the two subtasks "What are the user's values?" and "What action is best according to these values"? is not corrigible. It produces an action optimised for some approximate model of the user's values, which could be misaligned.
    - Capability requires having the ability to complete "enough" tasks. Paul acknowledges that the overseer will no longer be able to directly perform some tasks as well as a human (ie. they won't be able to translate sentences as well if they only are able to look at small pieces of the

original text). But in this case, he claims that it would be okay to understand how to safely use information provided by external resources (ie. human translators).
- Clarifying the currently imprecise definition of “corrigibility” in order to be able to reason about whether a given task decomposition is corrigible.
- Sufficiently understand meta-philosophy to implement it explicitly using collaborations of “lookup table overseers”. Even if we have an overseer that is corrigible and capable of performing well understood and defined tasks, it’s not clear that if we ask it to start evaluating tricky philosophical questions it will come up with aligned answers. This roughly looks like figuring out an intractably large algorithm that will output the products of MIRI’s Highly Reliable Agent Design research (ie. produce a set of criteria for what makes a good decision theory, a search space of possible decision theories, a prior over that space).

When considering the high bandwidth oversight regime, these problems seem more like they could be (at least partially) delegated to high bandwidth overseer at runtime. For example, one could think of preparing a summary of the relevant factors of a decision and giving it to a highly competent human overseer. The human overseer might be able to make good enough decisions without having an explicit model of decision theory (especially if odd edge cases don’t come up in practice).

Evaluating IDA

There are a number of lines of reasoning that one could use to try to get a sense of whether IDA could work, or how it should be prioritized relative to other approaches. Here, I summarize some of these considerations and indicate which versions of oversight they apply to, as I think this is important to be clear about. I don’t make claims about how these should be weighed, or claim to be exhaustive (there are additional considerations that would weigh on both forms of oversight, or other parts of IDA).

Evaluating High Bandwidth Oversight

Positive Considerations

- Reasons for general optimism about “learning to reason from humans”, such as those that Daniel Dewey discusses in his thoughts on MIRI's "highly reliable agent design" work. For example, the individual overseers are acting like humans. This would lead the system to consider problems in the way that humans would, incorporating common sense about human motivations, coming up with solutions that are easier for humans to understand and less likely to contain weird surprises, and operating on inputs that are understandable.
- If it seems like high bandwidth task decomposition could solve difficult problems (which is the question that Ought is running experiments to investigate)
- IDA with high bandwidth oversight seems like some version of it could be implemented today
- High bandwidth oversight seems that it could be implemented without solving many of the hard philosophical/mathematical problems that are faced by HRAD (decision theory, induction, etc.), instead substituting them for difficult machine learning problems.

Negative Consideration: Not Scalable and Secure

In high bandwidth oversight, if one copy of the overseer is corrupted, this corruption could easily spread. As more copies of the overseer are used to solve a problem, there is a higher chance of one copy becoming corrupted. Or an adversary could introduce a corrupting input into the system.

Evaluating Low Bandwidth Oversight

Positive Considerations

- Paul's intuitions suggest that it is possible to work with low bandwidth oversight, and he has discussed some aspects of how various problems might be handled.
- If it seems plausible that meta-philosophy might be easier than object level philosophy - ie. defining an algorithm that we can't run directly but outputs the correct decision theory might be easier than specifying the correct decision theory directly.
- Paul claims that this approach will scale with increasing capability for any distillation algorithm.

Negative Consideration: No longer "learning to reason from humans"

I think that IDA with a low bandwidth overseer is not accurately described as "AI learns to reason from humans", rather more "Humans figure out how to reason explicitly, then the AI learns from the explicit reasoning". As Wei Dai has pointed out, amplified low bandwidth oversight will not actually end up reasoning like a human. Humans have implicit knowledge that helps them perform tasks when they see the whole task. But not all of this knowledge can be understood and break into smaller pieces. Low bandwidth oversight requires that the overseer not use any of this knowledge.

Now, it's quite possible that performance still could be recovered by doing things like searching over a solution space, or by reasoning about when it is safe to use training data from insecure humans. But these solutions could look quite different from human reasoning. In discussion on Universality Amplification, Paul describes why he thinks that a low bandwidth overseer could still perform image classification, but the process looks very different from a human using their visual system to interpret the image:

"I've now played three rounds of the following game (inspired by Geoffrey Irving who has been thinking about debate): two debaters try to convince a judge about the contents of an image, e.g. by saying "It's a cat because it has pointy ears." To justify these claims, they make still simpler claims, like "The left ears is approximately separated from the background by two lines that meet at a 60 degree angle." And so on. Ultimately if the debaters disagree about the contents of a single pixel then the judge is allowed to look at that pixel. This seems to give you a tree to reduce high-level claims about the image to low-level claims (which can be followed in reverse by amplification to classify the image). I believe the honest debater can quite easily win this game, and that this pretty strongly suggests that amplification will be able to classify the image."

Conclusion: Weighing Evidence for IDA

The important takeaway is that considering IDA requires clarifying whether you are considering IDA with high or low bandwidth oversight. Then, only count considerations that actually apply to that approach. I think there's a way to misunderstand the

approach where you mostly think about high bandwidth oversight and count the feeling like it's somewhat understandable, feels plausible to you that it could work and that it avoids some hard problems. But if you then also count Paul's opinion that it could work, you may be overconfident - the approach that Paul claims is most likely to work is the low bandwidth oversight approach.

Additionally, I think it's useful to consider both models as alternative tools for understanding oversight: for example, the problems in low bandwidth oversight might be less obvious but still important to consider in the high bandwidth oversight regime.

After understanding this, I am more nervous about whether Paul's approach would work if implemented, due to the additional complications of working with low bandwidth oversight. I am somewhat optimistic that further work (such as fleshing out how particular problems could be address through low bandwidth oversight) will shed light on this issue, and either make it seem more likely to succeed or yield more understanding of why it won't succeed. I'm also still optimistic about Paul's approach yielding ideas or insights that could be useful for designing aligned AIs in different ways.

Caveat: high bandwidth oversight could still be useful to work on

High bandwidth oversight could still be useful to work on for the following reasons:

- If you think that other solutions could be found to the security problem in high bandwidth oversight. Paul claims that low bandwidth oversight is the most likely solution to security issues within the overseer, but he thinks it may be possible to make IDA with high bandwidth oversight secure using various techniques for optimizing worst-case performance on the final distilled agent, even if the overseer is insecure. (see https://ai-alignment.com/two-guarantees-c4c03a6b434f)
- It could help make progress on low bandwidth oversight. If high bandwidth oversight fails, then so will low bandwidth oversight. If high bandwidth oversight succeeds, then we might be able to break down each of the subtasks into low bandwidth tasks, directly yielding a low bandwidth overseer). I think the factored cognition experiments planned by Ought plausibly fall into this category.
- If you think it could be used as a medium-term alignment solution or a fallback plan if no other alignment approach is ready in time. This seems like it would only work if it is used for limited tasks and a limited amount of time, in order to extend the time window for preparing a truly scalable approach. In this scenario, it would be very useful to have techniques that could help us understand how far the approach could be scaled before failure.

Ten Commandments for Aspiring Superforecasters

Cross-posted to the Effective Altruism Forum

In the last several years, political scientist and forecasting research pioneer Philip Tetlock has made waves for the success of his research program in geopolitical forecasting, published in the form of the popular book *Superforecasting*. It's been discussed much in the rationality community, and reviewed by Slate Star Codex. It turns out the skills of forecasting, such as the importance of taking the outside view into account, Fermi estimates, and Bayesian updating, will be familiar to aspiring rationalists. A lot of the value from the book, then, for readers here would be a summary of insights, so I've reproduced the appendix from the book that does just that below.

---

(1) Triage

Focus on questions where your hard work is likely to pay off. Don't waste time either on "clocklike" questions (where simple rules of thumb can get you close to the right answer) or on impenetrable "cloud-like" questions (where even fancy statistical models can't beat the dart-throwing chimp). Concentrate on questions in the Goldilocks zone of difficulty, where effort pays off the most.

For instance, "Who will win the presidential election, twelve years out, in 2028?" is impossible to forecast now. Don't even try. Could you have predicted in 1940 the winner of the election, twelve years out, in 1952? If you think you could have known it would be a then-unknown colonel in the United States Army, Dwight Eisenhower, you may be afflicted with one of the worst cases of hindsight bias ever documented by psychologists.

Of course, triage judgment calls get harder as we come closer to home. How much justifiable confidence can we place in March 2015 on who will win the 2016 election? The short answer is not a lot but still a lot more than we can for the election in 2028. We can at least narrow the 2016 field to a small set of plausible contenders, which is a lot better than the vast set of unknown (Eisenhower-ish) possibilities lurking in 2028.

Certain cases of outcomes have well-deserved reputations for being radically unpredictable (e.g., oil prices, currency markets). But we usually don't discover how unpredictable outcomes are until we have spun our wheels for a while trying to gain analytical traction. Bear in mind the two basic errors it is possible to make here. We could fail to try to predict the potentially predictable or we could waste our time trying to predict the unpredictable. Which error would be worse in the situation you face?

(2) Break seemingly intractable problems into tractable sub-problems.

Channel the playful but disciplined spirit of Enrico Fermi who--when he wasn't designing the world's first atomic reaction--loved ballparking answers to head-scratchers such as "How many extraterrestrial civilizations exist in the universe?" Decompose the problem into its knowable and unknowable parts. Flush ignorance into

the open. Expose and examine your assumptions. Dare to be wrong by making you best guesses. Better to discover quickly than to hide them behind vague verbiage.

Superforecasters see Fermi-izing as part of the job. How else could they generate quantitative answers to seemingly impossible-to-quantify questions about Arafat's autopsy, bird-flu epidemics, oil prices, Boko Haram, the Battle of Aleppo, and bond-yield spreads.

We find this Fermi-izing spirit at work even in the quest for love, the ultimate unquantifiable. Consider Peter Backus, a lonely guy in London, who guesstimated the number of potential female partners in his vicinity by starting with the population of London (approximately six million) and winnowing the number down by the proportion of women in the population (about 50%), by the proportion of singles (about 50%), by the proportion in the right age range (about 20%), by the proportion of university graduates (about 26%) by the proportion he finds attractive (only 5%), by the proportion likely to find him attractive (only 5%), and by the proportion likely to be compatible with him (about 10%). Conclusion: roughly twenty-six women in the pool, a daunting but not impossible search task.

There are no objectively correct answers to true-love questions, but we can score the accuracy of the Fermi estimates that superforecasters generate in the IARPA tournaments. The surprise is how often remarkably good probability estimates arise from a remarkably crude series of assumptions and guesstimates.

(3) Strike the right balance between inside and outside views.

Superforecasters know that there is nothing new under the sun. Nothing is 100% "unique". Language purists be damned: uniqueness is a matter of degree. So superforecasters for comparison classes even for seemingly unique events, such as the outcome of a hunt for a high-profile terrorist (Joseph Kony) or the standoff between a new socialist government in Athens and Greece's creditors. Superforecasters are in the habit of posing the outside-view question: How often do things of this sort happen in situations of this sort?

So too apparently is Larry Summers, a Harvard professor and former Treasury secretary. He knows about the planning fallacy: when bosses ask employees how long it will take to finish a project, employees tend to underestimate the time they need, often by factors of two or three. Summers suspects his own employees are no different. One former employee, Greg Mankiw, himself now a famous economist, recalls Summers's strategy: he doubled the employee's estimate, then moved to the next higher time unit. "So, if the research assistant says the task will take an hour, it will take two days. If he says two days, it will take four weeks." It's a nerd joke: Summers corrected for his employees' failure to take the outside view in making estimates by taking the outside view toward employee's estimates, and then inventing a funny correction factor.

Of course Summers would adjust his correction factor if an employee astonished him and delivered on time. He would balance his outside-view expectation of tardiness against the new inside-view evidence that a particular employee is an exception to the rule. Because each of us is, to some degree, unique.

(4) Strike the right balance between under- and overreacting to evidence.

Belief updating is to good forecasting as brushing and flossing are to good dental hygiene. It can be boring, occasionally uncomfortable, but it pays off in the long term. That said, don't suppose that belief updating is always easy because sometimes it is. Skillful updating requires teasing subtle signals from noisy news flows--all the while resisting the lure of wishful thinking.

Savvy forecasters learn to ferret out telltale clues before the rest of us. They snoop for nonobvious lead indicators, about what would have to happen before X could, where X might be anything from an expansion of Arctic sea ice to a nuclear war in the Korean peninsula. Note the fine line here between picking up subtle clues before everyone else and getting suckered by misleading clues. Does the appearance of an article critical of North Korea in the official Chinese press signal that China is about the squeeze Pyongyang hard--or was it just a quirky error in editorial judgment? The best forecasters tend to be incremental belief updaters, often moving the probabilities of, say, 0.4 to 0.35 or from 0.6 to 0.65, distinctions too subtle with vague verbiage, like "might" or "maybe", but distinctions that, in the long run, define the difference between good and great forecasters.

Yet superforecasters also know how to jump, or move their probability estimates fast in response to diagnostic signals. Superforecasters are not perfect Bayesian updaters but they are better than most of us. And that is largely because they value this skill and work hard at cultivating it.

(5) Look for the clashing causal forces at work in each problem.

For every good policy argument, there is typically a counterargument that is at least worth acknowledging. For instance, if you are a devout dove who believes the threatening military action never brings peace, be open to the possibility that you might be wrong about Iran. And the same advice applies if you are a devout hawk who believes that soft "appeasement" policies never pay off. Each side should list, in advance, the signs that would nudge them toward the other.

Now here comes the really hard part. In classical dialectics, thesis meets antithesis, producing synthesis. In dragonfly eye, one view meets another and another and another--all of which must be synthesized into a single image. There are no paint-by-number rules here. Synthesis is an art that requires reconciling irreducibly subjective judgments. If you do it well, engaging in this process of synthesizing should transform you from a cookie-cutter dove or hawk into an odd hybrid creature, a dove-hawk, with a nuanced view of when tougher or softer are likelier to work.

(6) Strive to distinguish as many degrees of doubt as the problem permits but no more.

Few things are either certain or impossible. And "maybe" isn't all that informative. So your uncertainty dial needs more than three settings. Nuance matters. The more degrees of uncertainty you can distinguish, the better a forecaster you are likely to be. As in poker, you have an advantage if you are better than your competitors at separating 60/40 bets from 40/60--or 55/45 from 45/55. Translating vague-verbiage hunches into numeric probabilities feels unnatural at first but it can be done. It just requires patience and practice. The superforecasters have shown what is possible.

Most of us could learn, quite quickly, to think in more granular ways about uncertainty. Recall the episode in which President Obama was trying to figure out whether Osama bin Laden was the mystery occupant of the walled-in compound in Abbottabad. And

recall the probability estimates of his intelligence officers and the president's reaction to their estimates: "This is fifty-fifty...a flip of the coin." Now suppose that President Obama had been shooting the breeze with basketball buddies and each one offered probability estimates on the outcome of a college game--and those estimates corresponded exactly to those offered by intelligence officers on the whereabouts of Osama bin Laden. Would the president still have shrugged and said, "This is fifty-fifty", or would he have said, "sounds like the odds fall between three to one and four to one"? I bet on the latter. The president is accustomed to granular thinking in the domain of sports. Every year, he enjoys trying to predict the winners of the March Madness basketball tournament, a probability puzzle that draws the attention of serious statisticians. But, like his Democratic and Republican predecessors, he does not apply the same rigor to national security decisions. Why? Because different norms govern different thought processes. Reducing complex hunches to scorable probabilities is de rigueur in sports but not in national security.

So, don't reserve the rigorous reasoning for trivial pursuits. George Tenet would not have dared utter "slam dunk" about weapons of mass destruction if the Bush 43 White House had enforced standards of evidence and proof that are second nature to seasoned gamblers on sporting events. Slam dunk implies one is willing to offer infinite odds--and to lose everything if one is wrong.

(7) Strike the right balance between under- and overconfidence, between prudence and decisiveness.

Superforecasters understand the risks both of rushing to judgment and of dawdling too near "maybe". They routinely manage the trade-off between the need to take decisive stands (who wants to listen to a waffler?) and the need to qualify their stands (who wants to listen to a blowhard?). They realize that long-term accuracy requires getting good scores on both calibration and resolution--which requires moving beyond blame-game ping-pong. It is not enough to just avoid the most recent mistake. They have to find creative ways to tamp down both types of forecasting errors--misses and false alarms--to the degree a fickle world permits such uncontroversial improvements in accuracy.

(8) Look for the errors behind your mistakes but beware of rearview-mirror hindsight biases.

Don't try to justify or excuse your failures. Own them! Conduct unflinching postmortems: Where exactly did I go wrong? And remember that although the more common error is to learn too little from failure and to overlook flaws in your basic assumptions, it is also possible to learn too much (you may have been basically on the right track but made a minor technical mistake that had big ramifications). Also don't forget to do postmortems on your successes too. Not all successes imply your reasoning was right. You may have just lucked out by making offsetting errors. And if you keep confidently reasoning along the same lines, you are setting yourself up for a nasty surprise.

(9) Bring out the best in others and let others bring out the best in you.

Master the fine arts of team management, especially perspective taking (understanding the arguments of the other side so well that you can reproduce them to the other's satisfaction), precision questioning (helping others to clarify their arguments so they aren't being misunderstood), and constructive confrontation (learning to disagree without being disagreeable). Wise leaders know how fine the line

can be between a helpful suggestion and micromanagerial meddling or between a rigid group and a decisive one or between a scatterbrained group or an open-minded one. Tommy Lasorda, the former coach of the Los Angeles Dodgers, got it roughly right: “Managing is like holding a dove in your hand. If you hold it too tightly you kill it, but if you hold it too loosely, you lose it.”

(10) Master the error-balancing bicycle.

Implementing each commandment requires balancing opposing errors. Just as you can’t learn to ride a bicycle by reading a physics textbook, you can’t become a superforecaster by reading training manuals. Learning requires doing, with good feedback that leaves no ambiguity about whether you are succeeding--”I’m rolling along smoothly!”--or whether you are failing--”crash!” Also remember that practice is not just going through the motions of making forecasts, or casually reading the news and tossing out probabilities. Like all other known forms of expertise, superforecasting is the product of deep, deliberative practice.

(11) Don’t treat commandments as commandments.

“It is impossible to lay down binding rules,” Helmuth von Moltke warned, “because two cases will never be exactly the same.” As in war, so in all things. Guidelines are the best we can do in a world where nothing is certain or exactly repeatable. Superforecasting requires constant mindfulness, even when--perhaps especially when--you are dutifully trying to follow these commandments.

Citation: Tetlock, Philip E., and Dan Gardner. "Appendix: Ten Commandments for Aspiring Superforecasters." In *Superforecasting*, 277-85. Penguin Random House Company, 2015.

Note: If you wanted to point out there was a link to these ten commandments from this Slate Star Codex post, that link is dead. So switch your "Ten Commandments of Superforecasting" bookmark to this post's url.

Implicit extortion

This is a linkpost for https://ai-alignment.com/implicit-extortion-3c80c45af1e3

In this post I describe a pattern of behavior I call “implicit extortion.” RL agents are particularly susceptible to implicit extortion, in a way that is likely to be problematic for high-stakes applications in open-ended strategic environments.

I expect that many people have made this point before. My goal is to highlight the issue and to explore it a little bit more carefully.

Basic setup

Consider two actors, the target (T) and manipulator (M), such that:

- M wants T to perform some *target action*—e.g. make a payment, leak information, buy a particular product, handicap itself…
- M can take *destructive actions* that hurts both M and T—e.g. spreading rumors about T, undercutting T in a marketplace, physically attacking T…

In *explicit extortion*, M threatens to take the destructive action unless T performs the target action. Then a naive T reasons: “if I don’t take the target action, something bad will happen, so I better take the target action.”

In *implicit extortion*, M simply performs the destructive action whenever T doesn’t perform the target action. Then a naive T eventually learns that failure to take the target action is associated with something bad happening, and so learns to take the target action.

Implicit extortion is very similar to explicit extortion:

- T would prefer not be the kind of person who is vulnerable to extortion, so that bad things don’t happen to them.
- Extortion doesn’t necessarily cost M very much, if they don’t follow through on the threat very often.

However, implicit extortion can be particularly hard to avoid:

- It can be effective without T realizing that it’s happening, which makes it hard for them to respond appropriately even if they do have defenses.
- It affects simple RL algorithms (which don’t have defenses against extortion, and can’t be easily modified to include such defenses).

Example

The most extreme and blatant example would be for M to send T a daily request for $100. On any day when T fails to pay, M launches a costly cyberattack against T. A human would immediately recognize this behavior as extortion and would respond appropriately, but an RL algorithm might simply notice that paying is the best strategy and therefore decide to pay.

Implicit extortion can be much harder to detect, while still being effective. Suppose that every time T tries to change their product, M runs a grassroots smear campaign. It might not be possible for T to distinguish the situations “M is attempting to manipulate me into not changing my product” and “Everytime I change the product people get really unhappy, so I should do so sparingly.”

Details

How expensive is this for the manipulator?

Suppose that T is using an RL algorithm, and M is trying to manipulate them. How expensive is this for M? How likely is it to be worthwhile?

At equilibrium: T learns to always perform the target action; so only fails to take the target action while exploring. The long-term cost to M depends entirely on the target’s exploration policy.

If T uses ε-exploration, then they take the target action $(1 - \varepsilon)$ of the time. So M only needs to pay the cost of the destructive action on an ε fraction of trials.

For complex high-level actions, the effective ε can’t be *too* high—it’s not a good idea to “try something crazy” 10% of the time just to see what happens. But let’s be conservative and suppose that $\varepsilon=0.1$ anyway.

Suppose that M is trying to directly extract money from T, $100 at a time, and that it costs M $500 of value in order to cause $150 of trouble for T.

If M asks for $100 on 10 occasions, T will refuse to pay only once as an exploration. Then M needs to pay that $500 cost only once, thereby ensuring that the cost of paying (=$100) is smaller than the average cost of refusing to pay (=$150). Meanwhile, M makes $900, pocketing $400 of profit.

In general, M can make a profit whenever the product of (payment efficiency) * (destructive efficiency) > ε, where “payment efficiency” is the benefit to M divided by the cost to T of the target action, and “destructive efficiency” is the cost to T divided by the cost to M of the destructive action.

In practice I think it’s not too uncommon for payment efficiency to be ~1, and for destructive efficiency to be >1, such that extortion is possible regardless of ε. Small values of ε make extortion considerably easier and more cost-effective, and make it much harder to prevent.

During learning: the analysis above only applies when the agent has already learned to consistently take the target action. Earlier in learning, the target action may only occur rarely and so punishment may be very expensive. This could be worth it over the long term but may be a major hurdle.

Fortunately for M, they can simply start by rewarding the target behavior, and then gradually shift to punishment once the target behavior is common. From the perspective of the RL agent, the benefit of the target action is the same whether it’s getting a reward or avoiding a punishment.

In the cash payment example, M could start by paying T $20 every time that T sends $10. Once T notices that paying works well, M can gradually reduce the payment towards $10 (but leaving a profit so that the behavior becomes more and more entrenched). Once T is consistently paying, M can start scaling up the cost of not paying while it gradually reduces the benefits of paying.

Analyzing the error

Paying off a (committed) extortionist typically has the best consequences and so is recommended by causal decision theory, but *having the policy of paying off extortionists* is a bad mistake.

Even if our decision theory would avoid caving in to extortion, it can probably only avoid implicit extortion if it recognizes it. For example, UDT typically avoids extortion because of the logical link from "I cave to extortion" → "I get extorted." There is a similar logical link from "I cave to implicit extortion" → "I get implicitly extorted." But if we aren't aware that an empirical correlation is due to implicit extortion, we won't recognize this link and so it can't inform our decision.

In practice the target is only in trouble if would-be manipulators know that they are inclined to comply with extortion. If manipulators base that judgment on past behavior, then taking actions that "look like what someone vulnerable to extortion would do" is itself a bad decision that even a causal decision theorist would avoid. Unfortunately, it's basically impossible for an RL algorithm to learn to avoid this, because the negative consequences only appear over a very long timescale. In fact, the timescale for the negative consequences is longer than the timescale over which the RL agent adjusts its policy— which is too long for a traditional RL system to possibly do the credit assignment.

Other learning systems

What algorithms are vulnerable?

At first glance the problem may seem distinctive to policy gradient RL algorithms, where we take actions randomly and then reinforce whatever actions are associated with a high reward.

But the same problem afflicts any kind of RL. For example, a model-based agent would simply learn the model "not doing what the manipulator wants causes <bad thing X> to happen," and using that model for planning would have exactly the same effect as using policy gradients.

More broadly, the problem is with the algorithm: "learn an opaque causal model and use it to inform decisions." That's an incredibly general algorithm. If you aren't willing to use that algorithm, then you are at a significant competitive disadvantage, since the world contains lots of complicated causal processes that we can learn about by experiment but can't model explicitly. So it seems like everyone just has to live with the risk of implicit extortion.

I describe the problem as afflicting "algorithms," but it can also afflict humans or organizations. For example, any organization that is compelled by arguments like "X

has always worked out poorly in the past, even though we're not quite sure why, so let's stop doing it" is potentially vulnerable to implicit extortion.

What about human learning?

Humans have heuristics like vindictiveness that help prevent us from being manipulated by extortion, and which seem particularly effective against implicit extortion. Modern humans are also capable of doing explicit reasoning to recognize the costs of giving in to extortion.

Of course, we can only be robust to implicit extortion when we recognize it is occurring. Humans do have some general heuristics of caution when acting on the basis of opaque empirical correlations, or in situations where they feel they might be manipulable. However, it still seems pretty clear that human learning is vulnerable to implicit extortion in practice. (Imagine a social network which subtly punishes users, e.g. by modulating social feedback, for failing to visit the site regularly.)

Evolution?

Evolution itself doesn't have any check against extortion, and it operates entirely by empirical correlations, so why isn't it exploited in this way?

Manipulating evolution requires the manipulator to have a time horizon that is many times the generation length of the target. There aren't many agents with long enough time horizons, or sophisticated enough behavior, to exploit the evolutionary learning dynamic (and in particular, evolution can't easily learn to exploit it).

When we do have such a large gap in time horizons and sophistication—for example, when humans square off against bacteria with very rapid evolution—we do start to see implicit extortion.

For example, when a population of bacteria develop resistance to antibiotic A, we take extra pains to totally eradicate them with antibiotic B, even though we could not afford to use that strategy if A-resistance spread more broadly through the bacteria population. This is effectively implicit extortion to prevent bacteria from developing A-resistance. It would continue to be worthwhile for humanity even if the side effects of antibiotic B were much worse than the infection itself, though we probably wouldn't do it in that case since it's a hard coordination problem (and there are lots of other complications).

Conclusion

There are many ways that an AI can fail to do the right thing. Implicit extortion is a simple one that is pretty likely to come up in practice, and which may seriously affect the applicability of RL in some contexts.

I don't think there is any "silver bullet" or simple decision-theoretic remedy to implicit extortion, we just need to think about the details of the real world, who might manipulate us in what ways, what their incentives and leverage are, and how to manage the risk on a case-by-case basis.

I think we need to define “alignment” narrowly enough that it is consistent with implicit extortion, just like we define alignment narrowly enough that it’s consistent with losing at chess. I’ve found understanding implicit extortion helpful for alignment because it’s one of many conditions under which an aligned agent may end up effectively optimizing for the “wrong” preferences, and I’d like to understand those cases in order to understand what we are actually trying to do with alignment.

I don’t believe implicit extortion is an existential risk. It’s just another kind of conflict between agents, that will divert resources from other problems but should “wash out in the long run.” In particular, every agent can engage in implicit extortion and so it doesn’t seem to shift the relative balance of influence amongst competing agents. (Unlike alignment problems, which shift influence from human values to whatever values unaligned AI systems end up pursuing.)

The First Rung: Insights from 'Linear Algebra Done Right'

Foreword

Linear algebra, my old flame - how I missed you. At my undergraduate institution, linear algebra was my introduction to proof-based mathematics. There are people who shake hands, and there are people who **shake hands**. You know the type - you grasp their hand, and they clamp down and pull you in, agitating so wildly you fear for the structural integrity of your joints. My first experience with proofs was an encounter of the latter variety.

I received my first homework grade, and I was *not* pleased with my performance. I promptly went to the library and vowed not to leave until I learned how to write proofs adequately. The hours passed, and, (thankfully for my stomach), I got it. I didn't let up all semester. Immediately before the final exam, I was doing pushups in the hallway, high-fiving my friends, and watching the 'Michael Jordan Top 50 All Time Plays' video while visualizing myself doing that to the test. Do that to the test I did indeed.

This time around, the appropriately-acronymized *LADR* is the first step on my journey to attain a professional-grade mathematical skillset.

Tight Feedback Loops

In a (possibly maniacal) effort to ensure both mastery of the material and the maturation of my proof skillset, I did nearly[1] every one of the 561 exercises provided. I skipped problems only when I was confident I wouldn't learn anything, or calculus I didn't remember was required (and the payoff didn't seem worth the time spent relearning it now in a shallow manner, as opposed to thoroughly learning more calculus later). If I could sketch a solid proof in my head, I wouldn't write anything down. Even in the latter case, I checked my answers using this site (additional solutions may be found here, although be warned that not all of them are correct).

I also sometimes elected to give myself small hints after being stuck on a problem for a while; the idea was to keep things at the difficulty sweet spot. Specifically, I'd spend 10-20 minutes working on a problem by myself; if I wasn't getting anywhere, I'd find a hint and then *backpropagate the correct mental motion instead of what I had been trying to do.* I think that focusing on where you were going wrong and what insight you *should* have had, in what direction you *should* have looked, is more efficient than just reading solutions.

Over time, I needed fewer hints, even as problem difficulty increased.

My approach was in part motivated by the findings of Rohrer and Pashler:

> Surprisingly little is known about how long-term retention is most efficiently achieved... Our results suggest that a single session devoted to the study of some

> material should continue long enough to ensure that mastery is achieved but that immediate further study of the same material is an inefficient use of time.

The point isn't to struggle *per se* - it's to improve and to *win*.

L**inear** A**lgebra** D**one** R**ight**

This book has been previously reviewed by Nate Soares; as such, I'll spend time focusing on the concepts I found most difficult. Note that his review was for the second edition, while mine is for the third.

True to my vow in the last post, I have greatly improved my proof-babble; a sampling of my proofs can be found here.

> If you zip through a page in less than an hour, you are probably going too fast.

Try me.

1: Vector Spaces

In which the author reviews complex numbers, vector spaces, and subspaces.

I kept having trouble parsing

> For $f, g \in F^S$, the sum $f + g \in F^S$ is defined by $(f + g)(x) = f(x) + g(x)$ for all $x \in S$.

because my brain was insisting there was a type error in the function composition. I then had the stunning (and overdue) realization that my mental buckets for "set-theoretic functions" and "mathematical functions in general" should be merged.

That is, if you define

$$f : X \to Y = \{ (x, f(x)) : x \in X \}$$

$$g : X \to Y = \{ (x, g(x)) : x \in X \},$$

then $(f + g) : X \to Y$ simply has the definition $\{(x, f(x) + g(x)) : x \in X\}$. There isn't "online computation"; the composite function simply has a different Platonic lookup table.

2: Finite-Dimensional Vector Spaces

In which the author covers topics spanning linear independence, bases, and dimension.

3: Linear Maps

In which the author guides us through the fertile territory of linear maps, introducing null spaces, matrices, isomorphisms, product and quotient spaces, and dual bases.

> So far our attention has focused on vector spaces. No one gets excited about vector spaces.

Matrix Redpilling

The author built up to matrix multiplication by repeatedly insinuating that linear maps are secretly just matrix multiplications, teaching you to see the true fabric of the territory you've been exploring. Very well done.

Look no further than here and here for an intuitive understanding of matrix multiplication.

Dual Maps

> If $T \in L(V, W)$ then the dual map of T is the linear map $T' \in L(W', V')$ defined by $T'(\phi) = \phi \circ T$ for $\phi \in W'$.

This StackExchange post both articulates and answers my initial confusion.

Grueling Dualing

> The double dual space of V, denoted V'', is defined to be the dual space of V'. In other words, $V'' = (V')'$. Define $\Lambda : V \to V''$ by $(\Lambda v)(\varphi) = \varphi(v)$ for $v \in V$ and $\varphi \in V'$.

Stay with me, this is dualble.

So Λ takes some $v \in V$ and returns the curried function $\Lambda_v \in V''$. Λ_v, being in V'', takes some $\varphi \in V'$ and returns some $a \in F$. In other words, $\Lambda_v \in V''$ lets you evaluate the space of evaluation functions (V') with respect to the *fixed* $v \in V$. That's it!

4: Polynomials

In which the author demystifies the quadratic formula, sharing with the reader those reagents used in its incantation.

> Remarkably, mathematicians have proved that no formula exists for the zeros of polynomials of degree 5 or higher. But computers and calculators can use clever numerical methods to find good approximations to the zeros of any polynomial, even when exact zeros cannot be found.

> For example, no one will ever be able to give an exact formula for a zero of the polynomial p defined by $p(x) = x^5 - 5x^4 - 6x^3 + 17x^2 + 4x - 7$.

...

There are two cats where I live. Sometimes, I watch them meander around; it's fascinating to think how they go about their lives totally oblivious to the true nature of the world around them. The above incomputability result surprised me so much that I have begun to suspect that I too am a clueless cat (until I learn complex analysis; you'll excuse me for having a few other textbooks to study first).

Edit: daozaich writes about why this isn't as surprising as it seems.

5: Eigenvalues, Eigenvectors, and Invariant Subspaces

*In which the author uses the prefix '*eigen*-' so much that it stops sounding like a word.*

Revisiting Material

Before starting this book, I watched 3Blue1Brown's video on eigenvectors and came out with a vague "understanding". Rewatching it after reading Ch. 5.A, the geometric intuitions behind eigenvectors didn't seem like useful ways-to-remember an exotic math concept, they felt like a manifestation of how the world works. I *knew* what I was seeing from the hundreds of proofs I'd done up to that point.

Imagine being blind yet knowing the minute details of each object in your room; one day, a miracle treatment restores your eyesight in full. Imagine then seeing your room for the "first time".

Diagonalizability

Intuitively, the diagonalizability of some operator $T \in L(V)$ on a finite-dimensional vector space V means you can partition (more precisely, express as a direct sum) V by the eigenspaces $E(\lambda_i, T)$.

Another way to look at it is that diagonalization is the mutation of the basis vectors of V so that each column of $M(T)$ is one-hot[2]; you then rearrange the columns (by relabeling the basis vectors) so that $M(T)$ is diagonal.

Unclear Exercise

On page 156, you'll be asked to verify that a matrix is diagonalizable with respect to a provided nonstandard basis. The phrasing of the exercise makes it seem trivial, but

the book doesn't specify how to do this until Ch. 10. Furthermore, it isn't core conceptual material. Skip.

6: Inner Product Spaces

In which the author introduces inner products, orthonormal bases, the Cauchy-Schwarz inequality, and a neat solution to minimization problems using orthogonal complements.

7: Operators on Inner Product Spaces

In which the author lays out adjoint, self-adjoint, normal, and isometric operators, proves the (a) Spectral theorem, and blows my mind with the Polar and Singular Value Decompositions.

Adjoints

Consider the linear functional $\varphi \in L(W, F)$ given by $\langle Tv, w \rangle$ for fixed $v \in V$; this is then a linear functional on W for the chosen Tv. The adjoint T^* produces the corresponding linear functional in $L(V, F)$; given fixed $w \in W$, we now map to some linear functional on V such that $\langle Tv, w \rangle = \langle v, T^*w \rangle$. The left-hand side is a linear functional on W, and the right-hand side is a linear functional on V.

The Ghost Theorem

My brain was unreasonably excited for this chapter because I'd get to learn about "ghosts" (AKA the Spectral theorem). My conscious self-assurances to the contrary completely failed to dampen this ambient anticipation.

8: Operators on Complex Vector Spaces

In which generalized eigenvectors, nilpotent operators, characteristic and minimal polynomials, and the Jordan Form make an appearance, among others.

9: Operators on Real Vector Spaces

In which real vector spaces are complexified and real operators are brought up to speed with their complex counterparts.

10: Trace and Determinant

In which the curtain is finally pulled back.

> We proved the basic results of linear algebra before introducing determinants in this final chapter. Although determinants have value as a research tool in more advanced subjects, they play little role in basic linear algebra (when the subject is **done right**).

Sassy partial title drop (emphasis mine).

Final Verdict

Overall, I really liked this book and its clean theoretical approach. By withholding trace and det until the end of the book, many properties were arrived at in a natural, satisfying, and enlightening manner. The proofs were clean, and the writing was succinct (although I did miss the subtle wit of Russell and Norvig). This book positively, definitely belongs on the MIRI book list.

Forwards

Timing

This review follows the previous by exactly four weeks; however, I was at CFAR for a week during that time, dedicated a few days to *All of Statistics* (my next review), and slowed myself considerably by doing **five hundred** proofs. If I were treating this as a normal textbook, I imagine it would have taken less than half the time.

The most exciting effect of diving into math like this is that when I don't understand a concept, I now eagerly look to the *formalization* for clarification (previously, I'd barely be able to track all the Greek).

Fluency

An interesting parallel between learning math and learning languages: when I started picking up French, at first the experience was basically always was "ugh now I have to look up 5 things to even have the vocabulary to ask how to turn on my computer". Eventually, it became natural to belt out *et comment est-ce que je peux allumer mon ordi ? L'enfoiré refuse de fonctionner, comme d'habitude ; c'est grand temps que j'en achète un de plus*. No checking needed.

And so it went with proofs - "what techniques can I use to translate this statement into the answer" turned into "the proof feels like it's flowing out of my arm?!".

Proofs

I've noticed that when I successfully produce a non-trivial proof, it's nearly always when I have a strong understanding that *this is how the world is*. The proof is then just translating this understanding to math-ese, pounding away at the shell of the problem with every tool at my disposal to reach this truth.

Imagine a friend of yours fell under the ice. In one situation, you meander, blindfolded and half-deaf, with a vague idea of "I *think* they were this way?", trying different things and occasionally hearing faint pounding.

Now consider the situation in which you *know* where they are; it's then a matter of finding the right tools to smash the ice. You strike with everything you have, with every ounce of strength you possess; finally, you break your friend free.

Impatience

My most obvious remaining weak point with proofs is impatience. I have a strong intuition that this impulse is borne from my programming experience. When I write code, I carefully consider pre- and post-conditions, expected use cases, and the context of the problem. When using an external library, things are different; when asked why something is appropriate for use in a (low-stakes) program, it's fine to only provide high-level intuitions.

Similarly, in the few situations in which I have had to prove a novel result, I have found myself being extremely cautious (and rightly so). However, when proving a known result, a strong desire to take shortcuts overtakes me. I'm going to have to keep ironing this out.

Hiding Ignorance

Another aspect of this journey which I greatly enjoy is the methodical elimination of deficiencies and weak points. In my deep learning class, I had great trouble remembering what an eigenvalue was - it was at this moment that I knew I had to get down to business. Working with a surface-level understanding yields superficial results.

I imagine I was not the only person who was somewhat confused. However, being the first to admit confusion feels low-status: "everyone else seems to be following along, so I better be quiet and figure this out on my own time." I've made a point of ignoring this reasoning and asking more questions, and I think it's paid off. Incidentally, everyone else seemed relieved that the question got asked.

LessWrong

I'd like to add that in these posts, I present a somewhat distorted perspective of my academic life; these weak points are the exception, not the norm (*ahem*). I focus on my weak points because I want to become stronger - to admit them is not necessarily to say "*I* am weak" (although this may be the case relative to the person I want to become).

Speaking from experience, I feel that this is intimidating to newcomers. The culture can appear highly critical; this has been discussed before. I hope to do my part through these very posts, in which I plainly admit "I forgot eigenvalues. I fixed it - and I'm better off for having done so."

Calculus

The calculus-based exercises in this book and in *All of Statistics* make me uncomfortable. In the spirit of not hiding ignorance, I'll admit it[3] - I totally forgot how to integrate by parts, among other things. Although MIRI math is mostly discrete, I imagine that I'll still make a quick run through a calculus textbook in the near future.

I also find myself curious about real and complex analysis, but I suspect that's more of a luxury (given timelines). Maybe I'll learn it in my free time at some point.

Lost Calling

I have the distinct feeling of having been incredibly silly for many years; one of the reasons being my pretending that I didn't love math. In high school, I did quite well (and was designated the outstanding mathematics student of my class) as a product of my passion for toying with math in my free time.

However, in college, I just wanted to learn computer science. I'd gloss over the low-level math (although I did do some Project Euler for fun). Instead, I preferred learning to find clever high-level solutions and build up an algorithm-centric problem-solving toolkit. Now that I've truly taken the plunge, the water is just so nice.

I'm sorry to have been away for so long.

If you are interested in working with me or others on the task of learning MIRI-relevant math, if you have a burning desire to knock the alignment problem down a peg - I would be more than happy to work with you. Messaging me may also have the pleasant side effect of your receiving an invitation to the MIRIx Discord server.

[1] For Ch. 8-10, I did a random sampling of 15% of the practice problems, as opposed to 100% (I was reaching steeply-diminishing returns for relevant gears-level gains).

[2] Please let me know if there's a more appropriate linear-algebraic term for this.

[3] Merely admitting ignorance is not virtuous.

> The eighth virtue is humility. To be humble is to take specific actions in anticipation of your own errors. To confess your fallibility and then do nothing about it is not humble; it is boasting of your modesty. Who are most humble? Those who most skillfully prepare for the deepest and most catastrophic errors in their own beliefs and plans. Because this world contains many whose grasp of rationality is abysmal, beginning students of rationality win arguments and acquire an exaggerated view of their own abilities. But it is useless to be superior: Life is not graded on a curve. The best physicist in ancient Greece could not calculate the path of a falling apple. There is no guarantee that adequacy is possible given your hardest effort; therefore spare no thought for whether others are doing worse. If you compare yourself to others you will not see the biases that all humans share. To be human is to make ten thousand errors. No one in this world achieves perfection.

The virtue is in shedding ignorance:

The first virtue is curiosity. A burning itch to know is higher than a solemn vow to pursue truth. To feel the burning itch of curiosity requires both that you be ignorant, and that you desire to relinquish your ignorance. If in your heart you believe you already know, or if in your heart you do not wish to know, then your questioning will be purposeless and your skills without direction. Curiosity seeks to annihilate itself; there is no curiosity that does not want an answer. The glory of glorious mystery is to be solved, after which it ceases to be mystery. Be wary of those who speak of being open-minded and modestly confess their ignorance. There is a time to confess your ignorance and a time to relinquish your ignorance.

~ *The Twelve Virtues of Rationality*

Metaphilosophical competence can't be disentangled from alignment

Having human values is insufficient for alignment

Suppose there's a button where if you push it and name a human, that human becomes 1,000,000,000,000,000x more powerful. (What I mean by that isn't precisely specified—imagine some combination of being able to think much faster, becoming more intelligent, and having far more resources, to the point that they could easily overpower the rest of the world.)

Try running some thought experiments where you push the button to amplify:

- Jesus
- Buddha
- Adolf Hitler
- Donald Trump
- Kim Jong-un
- King Salman of Saudi Arabia
- Ayn Rand
- Elon Musk
- Ray Kurzweil
- Eliezer Yudkowsky
- Yourself

My intuition is that some of these people are catastrophic to amplify, and some *might* be OK to amplify. It's interesting to me that amplifying some of these people might be catastrophic, given that they're fellow human beings, raised in human societies, born with human genomes, who almost certainly care about the future well-being of humanity.

One reason I'd feel queasy amplifying anyone is that they might fall into an *epistemic pit*, where they arrive at some critically wrong conclusion and take either huge or infinite amounts of time to update away from it. If someone's reasoning process gets amplified, I wouldn't generally trust them to be good at arriving at true beliefs—intelligence needn't go hand-in-hand with rationality or philosophical competence.

In particular, it's very unclear to me whether people would quickly update away from ideologies. In practice, humanity as a whole has not obviously fallen into any permanent epistemic pits, but I think this is because no single ideology has clearly dominated the world. If you have indefinite decisive power over the world, you have far less incentive to consider perspectives very different from your own, and unless you both care about and are good at seeking true beliefs, you wouldn't do a good job learning from the people around you.

Another reason I'd feel queasy amplifying anyone is that they might take *irreversible catastrophic actions* (perhaps unknowingly). Genocides would be one example. Restructuring society such that it gets forever stuck in an epistemic pit would be

another. Building a superintelligence without appreciating the risks is yet another (and clearly the most disastrous, and also the least obviously disastrous).

I consider these all failures in something I'll term *metaphilosophical competence*. (Please excuse the unwieldy name; I hope to find a better descriptor at some point.) If someone were sufficiently metaphilosophically competent, they should figure out how to arrive at true beliefs relatively quickly and prioritize doing so. They should gain an appreciation of the importance and difficulty of avoiding catastrophic consequences in a world with so much uncertainty, and prioritize figuring out how to do good in a way that sets them apart from everyone who self-deludes into thinking they do good. They should be able to do this all correctly and expeditiously.

I interpet the goal of MIRI's agent foundations research agenda as providing a formal specification of metaphilosophical competence. For example, I interpret the logical induction criterion as part of a formal specification of what it means to have idealized reasoning in the limit. I intend to write more about this relationship at a future point.

All potential self-amplifiers should want to (and may not) be sufficiently metaphilosophically competent before self-amplifying

It's not just humans that should care about metaphilosophical competence. If Clippy (our favorite paperclip-maximizing superintelligence) wanted to build a successor agent far more powerful than itself, it would also want its successor to not take catastrophic irreversible actions or fall into epistemic pits.

Just because Clippy is superintelligent doesn't mean Clippy will necessarily realize the importance of metaphilosophy before building a successor agent. Clippy will probably eventually care about metaphilosophical competence, but it's possible it would come to care only after causing irreversible damage in the interim (for example it might have built a catastrophically misaligned subagent, a.k.a. a daemon). It's also conceivable it falls into an epistemic pit in which it never comes to care about metaphilosophy.

Acknowledging metaphilosophical competence may be insufficient for safe self-amplification

It might be sufficient for an agent that isn't yet completely metaphilosophically competent, but sufficiently "proto-metaphilosophically competent" to self-amplify. For example, the first thing it might do upon self-amplification is do nothing except determine a formal specification of metaphilosophical competence, then create a successor agent that's formally guaranteed to be metaphilosophically competent.

I'd feel good if I could be confident that would happen, but I'm not sure "do nothing but become more metaphilosophically competent" actually makes sense. Maybe it would make sense if you're smart enough that you could work through the aforementioned process in just a few seconds, but if for example the process takes much longer and you're in an unsafe or unstable environment, you'd have to trade off figuring out metaphilosophy with fending off imminent threats, which may involve taking irreversible catastrophic actions before you've actually figured out metaphilosophy.

(OK, metaphilosophy seems important to figure out. Wait, we might get nuked. Wait, synthetic viruses are spreading. Ahhhhh! Powerful AI's seem like the only way out of this mess. Ack, my AI isn't powerful enough, I should make it stronger. Okay, now it's... wait... oops...)

AI safety crux: Which humans are metaphilosophically competent enough to safely amplify?

Obviously some humans have not crossed the bar for metaphilosophical competence—if a naive negative utilitarian or angsty teenager gets 1,000,000,000,000,000x'd, they might literally just kill everyone. This invites the question of which people *have* crossed the metaphilosophical bar for safe 1,000,000,000,000,000x'ing.

I think this is an open question, and I suspect this is a major crux people have about the necessity or usefulness of agent foundations, as well as optimism about how AGI will play out. My guess is that if someone thinks tons of people have passed this bar, they'd think ML-based approaches to safety can lead us to a safe AGI, and are generally more optimistic about the world getting AI safety right. On the flip side, if they think practically nobody is sufficiently metaphilosophically competent to safely amplify, they'd highly prioritize metaphilosophical work (e.g. things in the direction of agent foundations), and feel generally pessimistic about the world getting AI safety right.

My take on agent foundations: formalizing metaphilosophical competence

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

I have some rough intuitions about the purpose of MIRI's agent foundations agenda, and I'd like to share them here. (Note: I have not discussed these with MIRI, and these should not be taken to be representative of MIRI's views.)

I think there's a common misconception that the goal of agent foundations is to try building an AGI architected with a decision theory module, a logical induction module, etc. In my mind, this is completely not the point, and my intuitions say that approach is doomed to fail.

I interpret agent foundations as being more about providing formal specifications of metaphilosophical competence, to:

- directly extend our understanding of metaphilosophy, by *adding conceptual clarity* to important notions we only have fuzzy understandings of. (Will this agent fall into epistemic pits? Are its actions low-impact? Will it avoid catastrophes?) As an analogy, formally defining mathematical proofs constituted significant progress in our understanding of mathematical logic and mathematical philosophy.
- allow us to *formally verify* whether a computational process will satisfy desirable metaphilosophical properties, like those mentioned in the above parenthetical. (It seems perfectly fine for these processes to be built out of illegible components, like deep neural nets—while that makes them harder to inspect, it doesn't preclude us from making useful formal statements about them. For example, in ALBA, it would help us make formal guarantees that distilled agents remain aligned.)

I want to explore logical induction as a case study. I think the important part about logical induction is the logical induction *criterion*, *not* the algorithm implementing it. I've heard the implementation criticized for being computationally intractable, but I see its primary purpose as showing the logical induction criterion to be *satisfiable at all.* This elevates the logical induction criterion over all the other loose collections of desiderata that may or may not be satisfiable, and may or may not capture what we mean by logical uncertainty. If we were to build an actual aligned AGI, I would expect its reasoning process to satisfy the logical induction criterion, but not look very much like the algorithm presented in the logical induction paper.

I also think the logical induction criterion provides an exact formalization—a necessary AND sufficient condition—of what it means to not get stuck in any epistemic pits in the limit. (The gist of this intuition: epistemic pits you're stuck in forever correspond exactly to patterns in the market that a trader could exploit forever, and make unbounded profits from.) This lets us formalize the question "Does X reasoning process avoid permanently falling into epistemic pits?" into "Does X reasoning process satisfy the logical induction criterion?"

Global insect declines: Why aren't we all dead yet?

One study on a German nature reserve found insect biomass (e.g., kilograms of insects you'd catch in a net) has declined 75% over the last 27 years. Here's a good summary that answered some questions I had about the study itself.

Another review study found that, globally, invertebrate (mostly insect) abundance has declined 35% over the last 40 years.

Insects are important, as I've been told repeatedly (and written about myself). So this news begs a very important and urgent question:

Why aren't we all dead yet?

This is an honest question, and I want an answer. Insects are among the most numerous animals on earth and central to our ecosystems, food chains, etcetera. 35%+ lower populations are the kind of thing where, if you'd asked me to guess the result in advance, I would have expected marked effects on ecosystems. By 75% declines – if the German study reflects the rest of the world to any degree – I would have predicted literal global catastrophe.

Yet these declines have been going on for apparently decades apparently consistently, and the biosphere, while not exactly doing great, hasn't literally exploded.

So what's the deal? Any ideas?

Speculation/answers welcome. Try to convey how confident you are and what your sources are, if you refer to any.

(If your answer is "the biosphere *has* exploded already", can you explain how, and why that hasn't changed trends in things like global crop production or human population growth? I believe, and think most others will agree, that various parts of ecosystems worldwide are obviously being degraded, but not to the degree that I would expect by drastic global declines in insect numbers (especially compared to other well-understood factors like carbon dioxide emissions or deforestation.) If you have reason to think otherwise, let me know.)

Crossposted from my personal blog. Also see Brian Tomasik's comment on the original post, which I haven't assessed in detail but which seems important.