Best of LessWrong: May 2018

1. Inadequate Equilibria vs. Governance of the Commons
2. Expressive Vocabulary
3. Critch on career advice for junior AI-x-risk-concerned researchers
4. Challenges to Christiano's capability amplification proposal
5. Terrorism, Tylenol, and dangerous information
6. Meta-Honesty: Firming Up Honesty Around Its Edge-Cases
7. Understanding is translation
8. Decoupling vs Contextualising Norms
9. The Alignment Newsletter #7: 05/21/18
10. Why Universal Comparability of Utility?
11. The Sheepskin Effect
12. Of Two Minds
13. [CKC] [May 2018] What subjects are most important for an AI safety researcher to know? (Open Call)
14. Moral frameworks and the Harris/Klein debate
15. Talents
16. The Alignment Newsletter #8: 05/28/18
17. Mental Illness Is Not Evidence Against Abuse Allegations
18. Gaining Approval: Insights From "How To Prove It"
19. Open question: are minimal circuits daemon-free?
20. Hypotheticals: The Direct Application Fallacy
21. The Alignment Newsletter #6: 05/14/18
22. Varieties Of Argumentative Experience
23. Predicting Future Morality
24. Bounded Rationality: Two Cultures
25. Biodiversity for heretics
26. Tech economics pattern: "Commoditize Your Complement"
27. Societal Growth Requires Rehabilitation
28. Please Take the 2018 Effective Altruism Survey!
29. Last chance to participate in the 2018 EA Survey
30. There is a war.
31. Co-Proofs
32. Problems integrating decision theory and inverse reinforcement learning
33. Against Not Reading Math Books Problems-First (If You've Found It Helpful Before)
34. Decision theory and zero-sum game theory, NP and PSPACE
35. Fundamentals of Formalisation level 2: Basic Set Theory
36. A Self-Respect Feedback Loop
37. Thoughts on AI Safety via Debate
38. Soviet-era Jokes, Common Knowledge, Irony
39. Advocating for factual advocacy
40. Affordance Widths
41. Bayes' Law is About Multiple Hypothesis Testing
42. Shared interests vs. collective interests
43. Fuzzy Boundaries, Real Concepts
44. On Better Mental Representations Part I: Adopting 'Thinking Tools'
45. Everything I ever needed to know, I learned from World of Warcraft: Goodhart's law
46. [LINK] How to write a dominant assurance contract on the Ethereum blockchain
47. April links
48. Mini-review: The Book of Why
49. Brief comment on frontpage/personal distinction
50. Trivial inconveniences as an antidote to akrasia

Best of LessWrong: May 2018

1. Inadequate Equilibria vs. Governance of the Commons
2. Expressive Vocabulary
3. Critch on career advice for junior AI-x-risk-concerned researchers
4. Challenges to Christiano's capability amplification proposal
5. Terrorism, Tylenol, and dangerous information
6. Meta-Honesty: Firming Up Honesty Around Its Edge-Cases
7. Understanding is translation
8. Decoupling vs Contextualising Norms
9. The Alignment Newsletter #7: 05/21/18
10. Why Universal Comparability of Utility?
11. The Sheepskin Effect
12. Of Two Minds
13. [CKC] [May 2018] What subjects are most important for an AI safety researcher to know? (Open Call)
14. Moral frameworks and the Harris/Klein debate
15. Talents
16. The Alignment Newsletter #8: 05/28/18
17. Mental Illness Is Not Evidence Against Abuse Allegations
18. Gaining Approval: Insights From "How To Prove It"
19. Open question: are minimal circuits daemon-free?
20. Hypotheticals: The Direct Application Fallacy
21. The Alignment Newsletter #6: 05/14/18
22. Varieties Of Argumentative Experience
23. Predicting Future Morality
24. Bounded Rationality: Two Cultures
25. Biodiversity for heretics
26. Tech economics pattern: "Commoditize Your Complement"
27. Societal Growth Requires Rehabilitation
28. Please Take the 2018 Effective Altruism Survey!
29. Last chance to participate in the 2018 EA Survey
30. There is a war.
31. Co-Proofs
32. Problems integrating decision theory and inverse reinforcement learning
33. Against Not Reading Math Books Problems-First (If You've Found It Helpful Before)
34. Decision theory and zero-sum game theory, NP and PSPACE
35. Fundamentals of Formalisation level 2: Basic Set Theory
36. A Self-Respect Feedback Loop
37. Thoughts on AI Safety via Debate
38. Soviet-era Jokes, Common Knowledge, Irony
39. Advocating for factual advocacy
40. Affordance Widths
41. Bayes' Law is About Multiple Hypothesis Testing
42. Shared interests vs. collective interests
43. Fuzzy Boundaries, Real Concepts
44. On Better Mental Representations Part I: Adopting 'Thinking Tools'
45. Everything I ever needed to know, I learned from World of Warcraft: Goodhart's law
46. [LINK] How to write a dominant assurance contract on the Ethereum blockchain

47. April links
48. Mini-review: The Book of Why
49. Brief comment on frontpage/personal distinction
50. Trivial inconveniences as an antidote to akrasia

Inadequate Equilibria vs. Governance of the Commons

This is a cross post from http://250bpm.com/blog:128.

Introduction

In the past I've reviewed Eliezer Yudkowsky's "Inadequate Equilibria" book. My main complaint was that while it explains the problem of suboptimal Nash equilibria very well, it doesn't propose any solutions. Instead, it says that we should be aware of such coordination failures and we should expect ourselves to fare better than the official institutions in such cases. What Yudkowsky is saying (if I understand him correctly) is that given that the treatment of short bowel syndrome in babies is stuck in an inadequate equilibrium, there's no way to fix the problem on the system level. However, you can spot the problem and acquire the medication needed to keep your kid alive from abroad yourself.

In my review I've sketched a couple of ideas how to approach the problem, but that was just me trying to be clever. Something backed by evidence would make me much more happy.

So I've decided to have a look at how people are getting out of suboptimal equilibria in the real world. And that's how I've got to the Elinor Ostrom's book "Governing the Commons (The Evolution of Institutions for Collective Action)".

The book is very explicitly covering the same problem as Yudkowsky. As Ostrom say about her research in her Nobel lecture: "One of the key questions that we've been addressing is 'Are rational individuals hopelessly trapped in dilemmas?'"

Ostrom is an economist with a game-theoretic bent. But she's not a theoretician. She's a field researcher. I've expected some good read and I wasn't disappointed.

The book focuses on special subset of coordination problems, problems of management of what Ostrom calls "common pool resources". Common pool resource is something that, unlike private property, is not exclusionary, something that one cannot easily restrict people from taking advantage of, but that, unlike public property, can be depleted by excessive usage. So, for example, it's hard to exclude others from fishing in the ocean, yet, the fish stock can be depleted by overfishing. Ocean fishery is a common pool resource. Compare that with a private fish farm where others are prevented from fishing by a barrier. And on the other side, contrast it with a public weather forecast which doesn't get depleted as more people turn on the radio.

On the game theoretic level, these scenarios can be though of as "tragedy of the commons" or "prisoner's dilemma" problems. At the heart of the problem is the fact that while cooperating and limiting everyone's rate of extraction helps to sustain the resource, individual players get better short-term reward if they defect and overuse the resource. For example, limiting the number of cows on a common grazing land keeps the pasture sustainable. However, each villager is incentivized to get as many

cows as possible which will in turn result in overgrazing and destruction of the grazing land.

Against economic absolutism

As Ostrom notes, people reading about this kind of problem tend to jump to one of the two conclusions. One group believes that the grazing land should be nationalized and the laws should be enacted to prevent the overgrazing. The other group thinks that the land should be split into small plots and privatized.

The problem with the latter solution is that it's often not feasible. For example, there's no reasonable way to privatize an ocean fishery. Fish stock is migratory and cannot be split among stakeholders. So, if you invest in the growth of the fish stock by temporarily fishing less, the fish will eventually migrate elsewhere and get caught by other fishers.

In other cases the resource can be split but doing so results in suboptimal performance. For example, if conditions on the grazing land vary, some parts may be dry this year, but lush the next year. Some parts may have been underwater the last year but are all right this year. In such cases splitting the grazing land gives suboptimal outcome for every participant. Each of them experiences a sequence of good and bad years. Cows die of hunger one year, but then the grazing land is underutilized the next year. Each herder's ability to plan ahead is badly compromised.

As for the nationalization approach, Ostrom demostrates the problem using a simple game. She starts with the classic tragedy of the commons scenario. Everybody is incentivized to overgraze and the pasture will be eventually destroyed.

But then the state steps in. Land is nationalized and overgrazing is prevented by law. Participants who do overgraze are fined. And when we take the fine into account in our game-theoretic calculation the outcome suddenly changes. Overgrazing in no longer the optimal strategy. The pasture will be sustained. Hooray for nationalization!

But not so fast! As Ostrom notes, law enforcement is not perfect. Sometimes offenders are not caught and sometimes people are fined unjustly. She assumes a certain error rate, she does the calculation again and lo and behold! The incentives to overgraze are back!

And we haven't yet considered the cost of policing. If we spent too little on policing, we'll either have less policemen, resulting in higher error rate, or the policemen are underpaid and tempted to accept bribes which will in turn compromise the entire system.

And, undoubtly, there are many more facets of the policy that play a role and can affect the result of the calculation.

It turns out that thinking about common pool resources in absolutist ways is not helpful. In reality, there's a broad continuous multi-dimensional range of policy options. The villagers can split the land into private and public parts. They can spend more on law enforcement. They can police each another. They can limit the usage of common resource based on effort spent maintaining it. They can use rotational allocation of land, or maybe a lottery. And so on and so forth.

But it's even more complex than that. It not just policies that determine the outcome. The nature and the particularities of the resource itself may determine the optimal policy.

Take, for example, the beach seine fishery in Mawelle, Sri Lanka. The village beach is divided into two launching sites, one on the harbor side and one on the rock side. There are different rules for those two sites. The rules also vary depending on the time in the day. The study in question notes:

> The Mawelle fishers provide a coherent explanation for why they use this complex set of authority rules, rather than a simple rotation system, to equalize the opportunity to make a big catch. Four environmental or technological considerations affect the problem of equalizing access: (1) The harbor side produces the really big catches, but the rock side is more consistently productive when there are fewer fish. (2) The first catch of the morning is most likely to be the biggest catch of the day, and the prices are highest in the morning. (3) The weather affects the number of hauls that can be made in the day, and any system assigning a set hour of the day would be inefficient. (4) Beach-seining involves high labor inputs to prepare a net for use and to restack it afterward, and simple rotation systems allowing all nets to be used only once per rotation would involve higher labor cost.

As can be seen, the geography, the shape of the seabed or maybe the sea currents, the weather patterns, the details of the fishing technology and so on, they all affect the optimal policy for governing the fishery. And it turns out that the optimal policy differs for two sites barely kilometer and a half apart.

There's no universal "correct" solution to the problem. A host of minor details matter. And, let's be frank, we have yet to see an economist who incorporates a map of sea currents at a particular place at the Sri Lankan shore into his economic model.

The absolutist position is also untenable because the optimal policy may change over time. Consider climate change. It may cause the weather patterns in Mawelle to change. And suddenly, the governance system that has worked for centuries stops working even though none of the rules were changed.

All in all, it seems that organically grown institutions are a lot like Hayek's free markets. They are information-processing machines. They aggregate countless details, too small and numerous for any central planner to take into account, and generate a set of efficient governance rules.

Sustainable institutions

In her discussion of the governance systems Ostrom picks several cases that proved to be sustainable over a long period. She analyzes communal tenure of high mountain meadows and forests in Switzerland and Japan, the "huerta" irrigation systems in Spain and the "zanjera" irrigation system in Philippines. All of them are practiced for hundreds of year, in some cases even for thousand years. I am not going to go into details, but do read the book. At the end she concludes that sustainable common pool resource governance systems tend to have the following features:

1. Clearly defined community boundaries, no strangers allowed.
2. Policies are fitted to local conditions.

3. People making policies have skin in the game.
4. System for checking whether participants are playing by the rules.
5. Graduated sanctions.
6. Conflict-resolution mechanisms.
7. No external meddling (e.g. by the government).
8. Nested enterprises (applies only for large-scale institutions).

I am not going to reproduce the entire argument, but I cannot resist quoting few interesting paragraphs from the book and expressing some thoughts of my own (in no particular order).

I always thought that social capital (or common knowledge, if you will) is expensive to create yet very easy to destroy. The following tidbit seems to indicate that it can, in fact, be rather resilient. It would be nice to see more research on this aspect of the common knowledge. Note how it was sustained via a two-sided consensus about the temporary nature of the violations of the system:

> [During the depression in 30's] almost all the villagers knew that almost all the other villagers were breaking the rules: sneaking around the commons at night, cutting trees that were larger than the allowed size, even using wood-cutting tools that were not permitted. This is precisely the behavior that could get a tragedy of the commons started, but it did not happen in Yamanaka. Instead of regarding the general breakdown of rules as an opportunity to become full-time free riders and cast caution to the winds, the violators themselves tried to exercise self-discipline out of deference to the preservation of the commons, and stole from the commons only out of desperation. Inspectors or other witnesses who saw violations maintained silence out of sympathy for the violators' desperation and out of confidence that the problem was temporary and could not really hurt the commons.

As for the "graduated sanctions" point it seems that fines for occasional breaches of the rules are very low. So low, in fact, that it still may be profitable for the player to break the rule. One study that comes to mind here is the one by Gneezy and Rustichini that shows that when parents were fined for picking their kids late from the day-care center they become more likely to be late. The fine have normalized otherwise socially unacceptable behaviour. One has to ask whether this effect plays a role in the systems described by Ostrom.

One observation that came as a surprise to me was that moderate fine (as opposed to large fine which may cause resentment) can in fact increase the trust in the system. The fined participant can experience fist-hand that breaking the rules doesn't go unnoticed and thus gains confidence that others are not breaking the rules. Think of it as a probe mechanism: Every participant has a cheap way to test the monitoring and enforcement system and make sure that it's still working as intended. If the fines were very high the ability to monitor the enforcement system would be compromised.

From the chapter about Zanjera irrigation system in Philippines:

> A few parcels, located at the tail end of the system, are assigned to the officials of the association as payment for their services. The system not only provides a positive reward for services rendered but also enhances the incentives for those in leadership positions to try to get water to the tail end of the system.

There's not much to say about it. We should use this kind of approach in politics much more often.

I am including the following quote because the overpopulation problem is one of the examples of suboptimal equilibria. In Hirano, Nagaike and Yamanaka villages in Japan:

> Rights of access to the communally held lands were accorded only to a household unit, not to individuals as such. Consequently, households with many members had no advantage, and considerable disadvantages, in their access to the commons. Population growth was extremely low (0.025% for the period 1721-1846), and ownership patterns within villages were stable.

If you are interested in the topic look also here.

Finally, I've enjoyed finding out that resentment-free fair-division algorithms are actually being used in the wild. When harvesting wood from communal forests in Törbel, Switzerland:

> The first step is that the village forester marks the trees ready to be harvested. The second step is that the households eligible to receive timber from work teams and equally divide the work of cutting the trees, hauling the logs, and piling the logs into approximately equal stacks. A lottery is then used to assign particular stacks to the eligible households.

It seems that a similar lottery system is used in Zanjera irrigation system in Philippines.

How are the institutions created?

All of that is nice and interesting, but hey, let's remember what we are after here! We are not looking for the best institution. We are trying to find out how to escape a local maximum. How to jump from an institution that sucks, but happens to be a Nash equilibrium, to a better institution. The question thus is not how such a better institution looks like but rather how it gets created.

When discussing the problem Ostrom shifts the focus from the long-lived institutions (we have very little, if any, data on how they were created centuries ago) to the governance of water basins in Southern California. These institutions were crated not that long time ago, in forties, fifties and sixties, and a lot of detailed information is available.

The problem is as follows. Under today's Los Angeles metropolitan area there are natural underground water reservoirs, a water-bearing strata made of sand and gravel. These reservoirs get replenished at a constant rate by the rains that fall in the foothills and upper valleys. They are also used by many parties as a source of water. This naturally leads to a tragedy-of-the-commons-style problem. If everyone pumps as much water as they can they are going to suck the reservoir dry. Even worse, low level of water in the reservoir means that sea water starts to seep in and will eventually destroy the reservoir.

The situation was made worse by the standing law. One was entitled to specific amount of water based on how much of it they were using. Thus, lawyers advised everyone to pump as much as they could.

The first change occurred in Raymond Basin. After the attempt to reach a voluntary settlement failed, city of Pasadena initiated legal proceedings against city of Alhambra

and 30 other producers.

That changed the dynamics of the system. It was found out that the water is being pumped out at a rate exceeding the replenishment rate by 38%. The ruling would, presumably, redistribute the water so that it was pumped out at at most the replenishment rate. However, due to complex legal and ownership arrangements it was not at all clear what the ruling is going to be and who's going to take the worst hit. Every participant had to consider the scenario where they would be the ultimate loser. That provided an incentive to try to reach a negotiated settlement.

Within six months the parties had drafted a stipulated agreement to share the cutback on proportional basis signed by all but 2 of the 32 participants of the litigation. Rather than imposing his own solution the judge issued a final judgement based on the stipulated agreement.

West Basin came next. Unlike in the case of small Raymond Basin, the litigation had almost 500 parties. Also, when the numbers came in it turned out that pumping rate was three times the natural replenishment rate of the basin. Thus, each participant had to face drastic cutbacks if the case was settled by the judge. A forum was created for negotiation of the settlement. Although everyone had an incentive to agree on anything better than two thirds cutback, it took two years of negotiations and a threat of court to achieve an interim agreement to cut back the pumping back to the levels from the year 1949. The interim agreement was used for seven years while the producers pursued other strategies to enhance the local water supplies, to replenish the basin, and to try to convince non-signatories to agree to the curtailment.

In 1961, after 16 years a trial was held and proposed judgement was passed to the court.

Later on similar process happened for Central Basin.

Now here's a quote that I find interesting:

> No one really knows the exact costs involved in the West Basin litigation, given the large number of parties and the length of time involved, but the best available estimate is $3 million. [...] Amortizing the costs of the litigation over a 50-year period (as one would expect to do for the construction of a major physical facility), [...] the adjudication costs in West Basin amounted to an annualized cost of $2.50 per acre-foot of water rights.

I like the idea of treating the creation of an institution (i.e. the rules governing the usage of the basin) as an infrastructure project, similar to a dam. If you read the piece by Scott Alexander you'll get an impression that you have to at least sacrifice a black goat and hire an exorcist to solve a coordination problem. After reading Ostrom though, it doesn't look any more exotic than hiring a couple of lawyers and an accountant.

It should be also taken into account that construction of an institution is an information good of a kind and thus the cost can be reduced by having some prior information. For example, the litigation process in Central Basin happened after the litigation in West Basin and the participants were thus able to learn from the existing experience, eliminate unnecessary steps (for example, it was immediately clear that they are shooting for a negotiated settlement, not a court ruling) and the cost of the project, despite Central Basin being bigger and having more stakeholders, were estimated at mere $450,000.

Similar project had, however, failed in San Bernardino county. Ostrom lists several reasons for the failure, including that San Bernardino county is much bigger that the other basins, that it may actually consists of several physical basins and so on. But listen to this:

> No voluntary water associations were created to facilitate discussion of these issues, and no consensus emerged over time about any of them. Conflicts emerged between the large and small water pumpers, between advocates for development and advocates for no-growth policies, between industry and agriculture, between locals and "external experts", and between appointed personnel and elected officials. The lack of fundamental agreement led to acrimonious political conflict, including several recall elections, front-page stories in the local papers that pushed aside stories on the Watergate scandal, and finally the suspension of the litigation in 1974. No action has since been taken to limit groundwater pumping.

It seems to me that the real cause of the failure was that the model from West Basin was imposed on the participants, in top-down manner, without much discussion. The fora that existed in West Basin weren't created in San Bernardino. The result was one ugly coordination failure.

Which makes me wonder whether the process of creation of an institution may be instrumental in its eventual success. What if the same set of rules in the same circumstances may succeed or fail depending on how they were conceived? What if the process of creation, the one where people discuss their options, argue about them and bounce ideas off one another serves as a mechanism to establish common knowledge among them? That's a really interesting idea and I would love to see some experimental results to either confirm it or disprove it.

Can this be replicated elsewhere?

Obviously, Californian water producers had many advantages on their side. They lived in a democratic country. They had functional legal system and law enforcement. Corruption was low. The level of social trust was high.

Would they be able to succeed elsewhere? It's hard to say, obviously, but let's have a look at what Ostrom writes about Sri Lankan irrigation system.

She discusses the topic in the chapter devoted to coordination failures. She describes the situation in Kirindi Oya. The population was heterogenous, composed of individuals coming from different regions, castes and kinship groups, all of whom are initially poor. The upstream farmers were using more water than they needed (because having rice fields flooded helps with weed control) and left little, if any, for downstream farmers. The central regime was unwilling to enforce rules impartially. Police treated water offences as trivial. Monitoring was non-existent and disputes were sometimes solved by violent means. The infrastructure was damaged:

> Gates are missing, structures damaged, channels tapped by encroachers and others. When asked why they don't prevent some of the most blatant offenses, two young technical assistants replied "that they were afraid to because of the fear of being assaulted".

In short, the situation was as bad as it gets.

But then she recounts a positive story from the left bank of Gal Oya irrigation project. The situation was similar to that in Kirindi Oya. The fact that upstream farmers were mostly Sinhalese while downstream cultivators were mostly Tamil haven't made things better.

> Cooperation among farmers was minimal. Social relations among settlers, who came from different areas of the country were strained... Relations between farmers and Irrigation Department (ID) officials were marked by mistrust and recriminations. Farmers had no confidence in the competence or the trustworthiness of the ID's staff... Many field-level officials ... were notorious for their corruption and thuggery. The main obstacle to efficient water management, from the farmers' point of view, was the local-level officials, who had political and bureaucratic power behind them. On the other hand, the ID officials, especially irrigation engineers, believed that farmers could not use water responsibly and carefully. Therefore, they argued that it was necessary to organize, educate, and discipline the farmers to do what ID asked them to do. Thus farmers were considered a part of the problem while the latter constitute the solution.

Then an experiment was made. The idea was to introduce "catalysts" into the situation of mutual mistrust and unpredictability. "Institutional organizers," (IO) mostly college graduates who also had farm backgrounds were dispatched to the area. They've received a six week training on how to approach and motivate farmers and on technical subjects related to irrigation. Each went to a small area served by one distributory canal. Their purpose was not to impose a particular policy but rather to organize the farmers to plan self-help strategies. At the same time they had status to deal effectively with Irrigation Department officials.

> Instead of establishing a predefined organization, the IO tried to form a working committee to solve particular problems, such as repairing a broken gate or desilting a field channel. Further, IOs identified problems beyond those that could be solved by local farmers working together, problems that had to be articulated to ID officials and others. Once farmers were used to working together and had achieved benefits from group action, the IO would then help form a local organization and select, through consensus, a farmer-representative. This representative could articulate the interests of the other farmers on his field channel at larger meetings and report back to the others what had happened in larger arenas.

When the farmers started working on rehabilitation of the field channel the attitude of irrigation officials toward them started to change.

In the areas where the new system was introduced farmers started to use water rotation procedures quite generally. There were even deliberate efforts to make water available to the farmers downstream.

On the Sinhalese-Tamil boundary, for example, maintenance haven't been done in years. Farmers talked about previous murders over water disputes. Within few months after introducing the new system, Sinhalese and Tamil farmers began to work on clearing out the channels.

That's not to say there were no problems in Gal Oya project. However, the example shows that there are ways to get out of suboptimal equilibria even in a highly damaged and non-functional environment. Sometimes, though, it seems to require a little nudge from the outside.

Conclusion

Common pool resource problems are a strict subset of inadequate equilibria problems. However, I don't see why they would be inherently easier to solve than other types of problems. Maybe there's a viable solution for each such problem. Maybe there's a way to escape any suboptimal Nash equilibrium. Maybe all we have to do is to try and when we fail to try again.

Expressive Vocabulary

"Thou shalt not strike terms from others' expressive vocabulary without suitable replacement." - me

---

Suppose your friend says: "I don't buy that brand of dip. It's full of chemicals."

Reasonable answer: "I'm skeptical that any of them are harmful in these quantities; we don't have much reason to believe that."

Reasonable answer: "Yellow 5? Are you allergic?"

Reasonable answer: "Okay, let's get the kind with four easily recognizable ingredients."

No: "Technically, *everything* is chemicals. Dihydrogen monoxide!"

Pedantry is seldom a way to make friends and influence people, but this example particularly gets my goat because there doesn't seem to actually exist a word in English for *the thing you know perfectly well people mean* when they say "chemicals". When I tried to find one on Twitter, the closest options were "toxins" and "additives". But neither is right. "Toxins" excludes yellow 5 - or, whether it does or not might be a point of contention; but it isn't the thing originally expressed with the word "chemicals". People may want to avoid - or otherwise discuss - "chemicals" for reasons other than thinking they're literally toxic; if I tell a maid I'm sensitive to chemical smells but vinegar is okay this is *useful information*. "Additives" includes, say, added sugar, which, while a plausible complaint, is a *separate* complaint.

Suppose your grandma says, "Okay, no technology at the dinner table."

Reasonable answer: "I'll put the laptop away to make room for the potatoes, but I need the phone because I get anxious without it."

Reasonable answer: "Sure, Grandma."

Reasonable answer: "We can try that until Uncle Bill starts making easily falsified claims about Flat Earth."

No: "*Technically*, the *dinner table* is a technology. And so are your glasses, Grandma."

In this case a more precise word exists - "electronics" ambiguously includes the chandelier but at least firmly sets aside the question of whether your grandma wants you to eat naked and with your bare hands. But *refusing to know what she meant* because she could have gotten closer to saying it, not even literally (she isn't being *metaphorical*), but technically, pedantically, definitionally? This is both a bad social move and a bad epistemic one; you're having the conversation on a level that is wholly about verbal wallpaper. Do *you* prefer to say "electronics" or dip into synechdoche with "screens" or spend nine syllables on "internet enabled devices"? Are you actually unsure if your grandmother wants you to set aside your smart watch, dumb phone, or electric blanket of intermediate intellect? Use your own words, ask

your own questions, but don't enforce an inadequate prescriptivism with feigned incomprehension while your interlocutor only wants you to pass the peas.

Thou shalt not strike terms from others' expressive vocabulary without suitable replacement. It's a pet issue of mine; it's my pinned tweet. "Suitable replacement" means suitable across the board, Pareto improvement as seen by the user along every axis a word can have. I think people are within their rights to reject a proposed replacement for not meaning the right thing, sounding ugly, being one syllable longer, being hard to spell, not rhyming in a poem they're trying to write, and vague gut feeling that you're just trying to control them. I extend this as far as "gypsy" and "Eskimo", at least (and with slightly less fervor to a slur beyond that if you really don't have another term for Brazil nuts).

Suitable replacement is a very high standard. It has to be. If you take someone's words away - and refusing to understand them when the problem is not in fact in your understanding does that, since words are tools to communicate - they are very direly crippled. Many people *think* communicatively; while you might not be their only outlet for working through their ideas, social shame for imprecise language can do your work for you across the board if you hit someone vulnerable hard enough. If you offer them worse words instead of expecting them to guess, they might only be crippled to the degree of wearing uncomfortable shoes, but that's still too much. Don't set up shop a block farther away than you had to and dress code folks for wearing Crocs. Communication is already difficult.

Some things I am *not* saying:

- you, yes you, have to talk to people who use words you can't stand or in ways you can't stand

Nah. Block people on every website you use over ship names and disown your sister for saying "moist" for all I care. You also have my blanket permission to use any sarcastic defense mechanism that works for you against your abusive parents if you have those, or whatever.

- you should not offer people better words for whatever value of "better"

By all means offer. "I think the preferred term is 'transgender' this week." But if they can't abide the difference in shade of meaning or mouthfeel, maybe even if they overtly announce it's just because they want to call it like they see it and they see it in some horrid way, don't try to correct them by pretending to be missing that section of your dictionary when you really aren't.

- Humpty Dumpty was right, words mean whatever the speaker wants them to mean, all is descriptivist chaos, "literally irregardless"

No. My examples have in common that they point at things and you can tell what things they are by being a speaker of the language in the conversational context. If someone starts calling cardboard boxes "pants" for no reason they're just wrong and you don't have to learn their stupid code.

Edit 12/2019:

sirjackholland wrote a comment, now slightly buried, including this paragraph:

> But I genuinely don't know what "natural" is supposed to (approximately) carve up, especially in the realm of foods. If you boil tea leaves, are the resulting compounds natural? If yes, then at what point do things become unnatural? If no, then is anything that's not raw and unprocessed unnatural, including e.g. cooked meat or boiled potatoes? There is clearly a spectrum between "raw and unprocessed" and "industrially engineered" but I don't see any reasonable place to draw the line. And this makes the word "natural" in the context of foods too vague to be useful - every time someone uses it, you have to ask a series of followup questions to figure out where they (arbitrarily) draw the line.

To which I replied:

> I want to point out that there are lots of situations where English speakers fluently use words that don't have clear dividing lines between their applicability and their inapplicability - it depends on context and details. "The music is loud." What if I'm deaf or far away or like to be able to feel the bass line in my bones? That doesn't make the sentence impermissible or even hard to understand and I don't need the speaker to produce a decibel value. "If you go to high altitudes, the air is thinner and you might get dizzy." How high? If I'm dizzy in Denver and the speaker thinks you shouldn't need to adjust your behavior until there are Sherpas about and meanwhile Batman can breathe in space, that doesn't make the sentence false, let alone useless. "It's cold, bring a jacket." Oh you sweet summer child, I'm good in short sleeves, thanks, I just *don't know what you meant by* "*cold*" -
>
> There are lots of conversational purposes for which you don't in fact have to know where someone draws the line. You don't even need to be able to agree on every point's ordering in the spectrum ("it's colder today" "that's just windchill"). The words *gesture in a direction*. I think "chemicals" does too, and you know what direction because you came up with "unprocessed" as a gloss on "low in chemicals". If someone doesn't buy that brand of dip because it's full of chemicals, in your innocent confusion I suggest you glance at the ingredients list for a guess at the threshold in question.

In the linguistic sense, a term's use can be "felicitous", without it having to be precise, literally accurate, etc. If you don't know what a word means but you know what spectrum it's on... that suggests that actually you know what the word means.

Critch on career advice for junior AI-x-risk-concerned researchers

In a recent e-mail thread, Andrew Critch sent me the following "subtle problem with sending junior AI-x-risk-concerned researchers into AI capabilities research". Here's the explanation he wrote of his view, shared with his permission:

I'm fairly concerned with the practice of telling people who "really care about AI safety" to go into AI capabilities research, unless they are very junior researchers who are using general AI research as a place to improve their skills until they're able to contribute to AI safety later. (See Leveraging Academia).

The reason is not a fear that they will contribute to AI capabilities advancement in some manner that will be marginally detrimental to the future. It's also not a fear that they'll fail to change the company's culture in the ways they'd hope, and end up feeling discouraged. What I'm afraid of is that **they'll feel pressure to start pretending to themselves, or to others, that their work is "relevant to safety"**. Then what we end up with are companies and departments filled with people who are "concerned about safety", creating a false sense of security that something relevant is being done, when all we have are a bunch of simmering concerns and concomitant rationalizations.

This fear of mine requires some context from my background as a researcher. I see this problem with environmentalists who "really care about climate change", who tell themselves they're "working on it" by studying the roots of a fairly arbitrary species of tree in a fairly arbitrary ecosystem that won't generalize to anything likely to help with climate change.

My assessment that their work won't generalize is mostly not from my own outside view; it comes from asking the researcher about how their work is likely to have an impact, and getting a response that either says nothing more than "I'm not sure, but it seems relevant somehow", or an argument with a lot of caveats like "X might help with Y, which might help with Z, which might help with climate change, but we really can't be sure, and it's not my job to defend the relevance of my work. It's intrinsically interesting to me, and you never know if something could turn out to be useful that seemed useless at first."

At the same time, I know other climate scientists who seem to have actually done an explicit or implicit Fermi estimate for the probability that they will personally soon discover a species of bacteria that could safely scrub the Earth's atmosphere of excess carbon. That's much better.

I've seen the same sort of problem with political scientists who are "really concerned about nuclear war" who tell themselves they're "working on it" by trying to produce a minor generalization of an edge case of a voting theorem that, when asked, they don't think will be used by anyone ever.

At the same time, I know other political scientists who seem to be trying really hard to work backward from a certain geopolitical outcome, and earnestly working out the

details of what the world would need to make that outcome happen. That's much better.

Having said this, I do think it's fine and good if society wants to sponsor a person to study obscure roots of obscure trees that probably won't help with climate change, or edge cases of theorems that no one will ever use or even take inspiration from, but I would like everyone to be on the same page that in such cases what we're sponsoring is intellectual freedom and development, and not climate change prevention or nuclear war prevention. If folks want to study fairly obscure phenomena because it feels like the next thing their mind needs to understand the world better, we shouldn't pressure them to have to think that the next thing they learn might "stop climate change" or "prevent nuclear war", or else we fuel the fire of false pretenses about which of the world's research gaps are being earnestly taken care of.

Unfortunately, the above pattern of "justifying" research by just reflecting on what you care about, rationalizing it, and not checking the rationalization for rationality, appears to me to be extremely prevalent among folks who care about climate change or nuclear war, and this is not something I want to see replicated elsewhere, especially not in the burgeoning fields of AI safety, AI ethics, or AI x-risk reduction. And I'm concerned that if we tell folks to go into AI research just to "be concerned", we'll be fueling a false sense of security by filling departments and companies with people who "seem to really care" but aren't doing correspondingly relevant research work, and creating a research culture where concerns about safety, ethics, or x-risk do not result in actually prioritizing research into safety, ethics, or x-risk.

When you're giving general-purpose career advice, the meme "do AI yourself, so you're around to help make it safe" is a really bad meme. It fuels a narrative that says "Being a good person standing next to the development of dangerous tech makes the tech less dangerous." Just standing nearby doesn't actually help unless you're doing technical safety research. Just standing nearby does create a false sense of security through the mere-exposure effect. And the "just stand nearby" attitude drives people to worsen race conditions by creating new competitors in different geographical locations, so they can exercise their Stand Nearby powers to ensure the tech is safe.

Important: the above paragraphs are advice *about what advice to give*, because of the social pressures and tendencies to rationalize that advice-giving often produces. By contrast, if you're a person who's worried about AI, and thinking about a career in AI research, I *do not* wish to discourage you from going into AI capabilities research. To you, what I want to say is something different....

Step 1: Learn by doing. Leverage Academia. Get into a good grad school for AI research, and focus first on learning things that feel like they will help you personally to understand AI safety better (or AI ethics, or AI x-risk; replace by your area of interest throughout). Don't worry about whether you're "contributing" to AI safety too early in your graduate career. Before you're actually ready to make real contributions to the field, try to avoid rationalizing doing things because "they might help with safety"; instead, do things because "they might help me personally to understand safety better, in ways that might be idiosyncratic to me and my own learning process."

Remember, what you need to learn to understand safety, and what the field needs to progress, might be pretty different, and you need to have the freedom to learn whatever gaps seem important to you personally. Early in your research career, you need to be in "consume" mode more than "produce" mode, and it's fine if your way of

"consuming" knowledge and skill is to "produce" things that aren't very externally valuable. So, try to avoid rationalizing the externally-usable safety-value of ideas or tools you produce on your way to understanding how to produce externally-usable safety research later.

The societal value of you producing your earliest research results will be that they help you personally to fill gaps in your mind that matter for your personal understanding of AI safety, and that's all the justification you need in my books. So, *do* focus on learning things that *you* need to understand safety better, but *don't* expect those things to be a "contribution" that will matter to others.

Step 2: Once you've learned enough that you're able to start contributing to research in AI safety (or ethics, or x-risk), then start focusing directly on making safety research contributions that others might find insightful. When you're ready enough to start actually producing advances in your field, *that's* when it's time to start thinking about the social impact of those advances would be, and start shifting your focus somewhat away from learning (consuming) and somewhat more toward contributing (producing).

(Content from Critch ends here.)

Challenges to Christiano's capability amplification proposal

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

The following is a basically unedited summary I wrote up on March 16 of my take on Paul Christiano's AGI alignment approach (described in "ALBA" and "Iterated Distillation and Amplification"). Where Paul had comments and replies, I've included them below.

I see a lot of free variables with respect to what exactly Paul might have in mind. I've sometimes tried presenting Paul with my objections and then he replies in a way that locally answers some of my question but I think would make other difficulties worse. My global objection is thus something like, "I don't see any concrete setup and *consistent simultaneous* setting of the variables where this whole scheme works." These difficulties are not minor or technical; they appear to me quite severe. I try to walk through the details below.

It should be understood at all times that I do not claim to be able to pass Paul's ITT for Paul's view and that this is me criticizing my own, potentially straw misunderstanding of what I imagine Paul might be advocating.

Paul Christiano
Overall take: I think that these are all legitimate difficulties faced by my proposal and to a large extent I agree with Eliezer's account of those problems (though not his account of my current beliefs).
I don't understand exactly how hard Eliezer expects these problems to be; my impression is "just about as hard as solving alignment from scratch," but I don't have a clear sense of why.
To some extent we are probably disagreeing about alternatives. From my perspective, the difficulties with my approach (e.g. better understanding the forms of optimization that cause trouble, or how to avoid optimization daemons in systems about as smart as you are, or how to address X-and-only-X) are also problems for alternative alignment approaches. I think it's a mistake to think that tiling agents, or decision theory, or naturalized induction, or logical uncertainty, are going to make the situation qualitatively better for these problems, so work on those problems looks to me like procrastinating on the key difficulties. I agree with the intuition that progress on the agent foundations agenda "ought to be possible," and I agree that it will help at least a *little bit* with the problems Eliezer describes in this document, but overall agent foundations seems way less promising than a direct attack on the problems (given that we haven't tried the direct attack nearly enough to give up). Working through philosophical issues in the context of a concrete alignment strategy generally seems more promising to me than trying to think about them in the abstract, and I think this is evidenced by the fact that most of the core difficulties in my approach would also afflict research based on agent foundations.
The main way I could see agent foundations research as helping to address these problems, rather than merely deferring them, is if we plan to eschew large-scale ML altogether. That seems to me like a very serious handicap, so I'd only go that direction once I was quite pessimistic about solving these problems. My subjective experience is of making continuous significant progress rather than being stuck. I agree there is clear evidence that the problems are "difficult" in the sense that we are going to have to make progress in

order to solve them, but not that they are "difficult" in the sense that P vs. NP or even your typical open problem in CS is probably difficult (and even then if your options were "prove P != NP" or "try to beat Google at building an AGI without using large-scale ML," I don't think it's obvious which option you should consider more promising).

First and foremost, I don't understand how "preserving alignment while amplifying capabilities" is supposed to work at all under this scenario, in a way consistent with other things that I've understood Paul to say.

I want to first go through an obvious point that I expect Paul and I agree upon: Not every system of locally aligned parts has globally aligned output, and some additional assumption beyond "the parts are aligned" is necessary to yield the conclusion "global behavior is aligned". The straw assertion "an aggregate of aligned parts is aligned" is the reverse of the argument that Searle uses to ask us to imagine that an (immortal) human being who speaks only English, who has been trained do things with many many pieces of paper that instantiate a Turing machine, can't be part of a whole system that understands Chinese, because the individual pieces and steps of the system aren't locally imbued with understanding Chinese. Here the compositionally non-preserved property is "lack of understanding of Chinese"; we can't expect "alignment" to be any more necessarily preserved than this, except by further assumptions.

The second-to-last time Paul and I conversed at length, I kept probing Paul for what in practice the non-compacted-by-training version of a big aggregate of small aligned agents would look like. He described people, living for a single day, routing around phone numbers of other agents with nobody having any concept of the global picture. I used the term "Chinese Room Bureaucracy" to describe this. Paul seemed to think that this was an amusing but perhaps not inappropriate term.

If no agent in the Chinese Room Bureaucracy has a full view of which actions have which consequences and why, this cuts off the most obvious route by which the alignment of any agent could apply to the alignment of the whole. The way I usually imagine things, the alignment of an agent applies to things that the agent understands. If you have a big aggregate of agents that understands something the little local agent doesn't understand, the big aggregate doesn't inherit alignment from the little agents. Searle's Chinese Room can understand Chinese even if the person inside it doesn't understand Chinese, and this correspondingly implies, by default, that the person inside the Chinese Room is powerless to express their own taste in restaurant orders.

I don't understand Paul's model of how a ton of little not-so-bright agents yield a big powerful understanding in aggregate, in a way that doesn't effectively consist of them running AGI code that they don't understand.

Paul Christiano
The argument for alignment isn't that "a system made of aligned neurons is aligned." Unalignment isn't a thing that magically happens; it's the result of specific optimization pressures in the system that create trouble. My goal is to (a) first construct weaker agents who aren't internally doing problematic optimization, (b) put them together in a way that improves capability without doing other problematic optimization, (c) iterate that process.

Paul has previously challenged me to name a bottleneck that I think a Christiano-style system can't pass. This is hard because (a) I'm not sure I understand Paul's system, and (b) it's clearest if I name a task for which we don't have a present crisp algorithm. But:

The bottleneck I named in my last discussion with Paul was, "We have copies of a starting agent, which run for at most one cumulative day before being terminated, and this agent hasn't previously learned much math but is smart and can get to understanding algebra by the end of the day even though the agent started out knowing just concrete arithmetic. How does a system of such agents, without just operating a Turing machine that operates an AGI, get to the point of inventing Hessian-free optimization in a neural net?"

This is a slightly obsolete example because nobody uses Hessian-free optimization anymore. But I wanted to find an example of an agent that needed to do something that didn't have a simple human metaphor. We can understand second derivatives using metaphors like acceleration. "Hessian-free optimization" is something that doesn't have an obvious metaphor that can explain it, well enough to use it in an engineering design, to somebody who doesn't have a mathy and not just metaphorical understanding of calculus. Even if it did have such a metaphor, that metaphor would still be very unlikely to be invented by someone who didn't understand calculus.

I don't see how Paul expects lots of little agents who can learn algebra in a day, being run in sequence, to aggregate into something that can build designs using Hessian-free optimization, *without* the little agents having effectively the role of an immortal dog that's been trained to operate a Turing machine. So I also don't see how Paul expects the putative alignment of the little agents to pass through this mysterious aggregation form of understanding, into alignment of the system that understands Hessian-free optimization.

I expect this is already understood, but I state as an obvious fact that alignment is not in general a compositionally preserved property of cognitive systems: If you train a bunch of good and moral people to operate the elements of a Turing machine and nobody has a global view of what's going on, their goodness and morality does not pass through to the Turing machine. Even if we let the good and moral people have discretion as to when to write a different symbol than the usual rules call for, they still can't be effective at aligning the global system, because they don't individually understand whether the Hessian-free optimization is being used for good or evil, because they don't understand Hessian-free optimization or the thoughts that incorporate it. So we would not like to rest the system on the false assumption "any system composed of aligned subagents is aligned", which we know to be generally false because of this counterexample. We would like there to instead be some narrower assumption, perhaps with additional premises, which is actually true, on which the system's alignment rests. I don't know what narrower assumption Paul wants to use.

Paul asks us to consider AlphaGo as a model of capability amplification.

My view of AlphaGo would be as follows: We understand Monte Carlo Tree Search. MCTS is an iterable algorithm whose intermediate outputs can be plugged into further iterations of the algorithm. So we can use supervised learning where our systems of

gradient descent can capture and foreshorten the computation of some but not all of the details of winning moves revealed by the short MCTS, plug in the learned outputs to MCTS, and get a pseudo-version of "running MCTS longer and wider" which is weaker than an MCTS actually that broad and deep, but more powerful than the raw MCTS run previously. The alignment of this system is provided by the crisp formal loss function at the end of the MCTS.

Here's an alternate case where, as far as I can tell, a naive straw version of capability amplification clearly wouldn't work. Suppose we have an RNN that plays Go. It's been constructed in such fashion that if we iterate the RNN for longer, the Go move gets somewhat better. "Aha," says the straw capability amplifier, "clearly we can just take this RNN, train another network to approximate its internal state after 100 iterations from the initial Go position; we feed that internal state into the RNN at the start, then train the amplifying network to approximate the internal state of that RNN after it runs for another 200 iterations. The result will clearly go on trying to 'win at Go' because the original RNN was trying to win at Go; the amplified system preserves the values of the original." This doesn't work because, let us say by hypothesis, the RNN can't get arbitrarily better at Go if you go on iterating it; and the nature of the capability amplification setup doesn't permit any outside loss function that could tell the amplified RNN whether it's doing better or worse at Go.

> **Paul Christiano**
> I definitely agree that amplification doesn't work better than "let the human think for arbitrarily long." I don't think that's a strong objection, because I think humans (even humans who only have a short period of time) will eventually converge to good enough answers to the questions we face.

The RNN has only whatever opinion it converges to, or whatever set of opinions it diverges to, to tell itself how well it's doing. This is exactly what it is for capability amplification to preserve alignment; but this in turn means that capability amplification only works to the extent that what we are amplifying has within itself the capability to be very smart in the limit.

If we're effectively constructing a civilization of long-lived Paul Christianos, then this difficulty is somewhat alleviated. There are still things that can go wrong with this civilization qua civilization (even aside from objections I name later as to whether we can actually safely and realistically do that). I do however believe that a civilization of Pauls could do nice things.

But other parts of Paul's story don't permit this, or at least that's what Paul was saying last time; Paul's supervised learning setup only lets the simulated component people operate for a day, because we can't get enough labeled cases if the people have to each run for a month.

Furthermore, as I understand it, the "realistic" version of this is supposed to start with agents dumber than Paul. According to my understanding of something Paul said in answer to a later objection, the agents in the system are supposed to be even dumber than an average human (but aligned). It is not at all obvious to me that an arbitrarily large system of agents with IQ 90, who each only live for one day, can implement a much smarter agent in a fashion analogous to the internal agents themselves achieving understandings to which they can apply their alignment in a globally

effective way, rather than them blindly implementing a larger algorithm they don't understand.

I'm not sure a system of one-day-living IQ-90 humans ever gets to the point of inventing fire or the wheel.

If Paul has an intuition saying "Well, of course they eventually start doing Hessian-free optimization in a way that makes their understanding effective upon it to create global alignment; I can’t figure out how to convince you otherwise if you don’t already see that," I'm not quite sure where to go from there, except onwards to my other challenges.

> **Paul Christiano**
> Well, I can see one obvious way to convince you otherwise: actually run the experiment. But before doing that I'd like to be more precise about what you expect to work and not work, since I'm not going to literally do the HF optimization example (developing new algorithms is way, way beyond the scope of existing ML). I think we can do stuff that looks (to me) even harder than inventing HF optimization. But I don't know if I have a good enough model of your model to know what you'd actually consider harder.

Unless of course you have so many agents in the (uncompressed) aggregate that the aggregate implements a smarter genetic algorithm that is maximizing the approval of the internal agents. If you take something much smarter than IQ 90 humans living for one day, and train it to get the IQ 90 humans to output large numbers signaling their approval, I would by default expect it to hack the IQ 90 one-day humans, who are not secure systems. We're back to the global system being smarter than the individual agents in a way which doesn't preserve alignment.

> **Paul Christiano**
> Definitely agree that even if the agents are aligned, they can implement unaligned optimization, and then we're back to square one. Amplification only works if we can improve capability without doing unaligned optimization. I think this is a disagreement about the decomposability of cognitive work. I hope we can resolve it by actually finding concrete, simple tasks where we have differing intuitions, and then doing empirical tests.

The central interesting-to-me idea in capability amplification is that by *exactly* imitating humans, we can bypass the usual dooms of reinforcement learning. If arguendo you can construct an exact imitation of a human, it possesses exactly the same alignment properties as the human; and this is true in a way that is not true if we take a reinforcement learner and ask it to maximize an approval signal originating from the human. (If the subject is Paul Christiano, or Carl Shulman, I for one am willing to say these humans are reasonably aligned; and I'm pretty much okay with somebody giving them the keys to the universe in expectation that the keys will later be handed back.)

It is not obvious to me how fast alignment-preservation degrades as the exactness of the imitation is weakened. This matters because of things Paul has said which sound

to me like he's not advocating for perfect imitation, in response to challenges I've given about how perfect imitation would be very expensive. That is, the answer he gave to a challenge about the expense of perfection makes the answer to "How fast do we lose alignment guarantees as we move away from perfection?" become very important.

One example of a doom I'd expect from standard reinforcement learning would be what I'd term the "X-and-only-X" problem. I unfortunately haven't written this up yet, so I'm going to try to summarize it briefly here.

X-and-only-X is what I call the issue where the property that's easy to verify and train is X, but the property you want is "this was optimized for X and only X and doesn't contain a whole bunch of possible subtle bad Ys that could be hard to detect formulaically from the final output of the system".

For example, imagine X is "give me a program which solves a Rubik's Cube". You can run the program and verify that it solves Rubik's Cubes, and use a loss function over its average performance which also takes into account how many steps the program's solutions require.

The property Y is that the program the AI gives you also modulates RAM to send GSM cellphone signals.

That is: It's much easier to verify "This is a program which at least solves the Rubik's Cube" than "This is a program which was optimized to solve the Rubik's Cube and only that and was not optimized for anything else on the side."

If I were going to talk about trying to do aligned AGI under the standard ML paradigms, I'd talk about how this creates a differential ease of development between "build a system that does X" and "build a system that does X and only X and not Y in some subtle way". If you just want X however unsafely, you can build the X-classifier and use that as a loss function and let reinforcement learning loose with whatever equivalent of gradient descent or other generic optimization method the future uses. If the safety property you want is optimized-for-X-and-just-X-and-not-any-possible-number-of-hidden-Ys, then you can't write a simple loss function for that the way you can for X.

> **Paul Christiano**
> According to my understanding of optimization / use of language: the agent produced by RL is optimized only for X. However, optimization for X is liable to produce a Y-optimizer. So the actions of the agent are both X-optimized and Y-optimized.

The team that's building a less safe AGI can plug in the X-evaluator and let rip, the team that wants to build a safe AGI can't do things the easy way and has to solve new basic problems in order to get a trustworthy system. It's not unsolvable, but it's an element of the class of added difficulties of alignment such that the whole class extremely plausibly adds up to an extra two years of development.

In Paul's capability-amplification scenario, if we can get exact imitation, we are genuinely completely bypassing the whole paradigm that creates the X-and-only-X

problem. If you can get exact imitation of a human, the outputs have only and exactly whatever properties the human already has. This kind of genuinely different viewpoint is why I continue to be excited about Paul's thinking.

> **Paul Christiano**
> I agree that perfect imitation would be a way to get around the X-and-only-X problem. However, I don't think that it's plausible and it's not how my approach hopes to get around the X-and-only-X problem.
> I would solve X-and-only-X in two steps:
> First, given an agent and an action which has been optimized for undesirable consequence Y, we'd like to be able to tell that the action has this undesirable side effect. I think we can do this by having a smarter agent act as an overseer, and giving the smarter agent suitable insight into the cognition of the weaker agent (e.g. by sharing weights between the weak agent and an explanation-generating agent). This is what I'm calling informed oversight.
> Second, given an agent, identify situations in which it is especially likely to produce bad outcomes, or proofs that it won't, or enough understanding of its internals that you can see why it won't. This is discussed in "Techniques for Optimizing Worst-Case Performance."
> (It also obviously requires a smarter agent, which you hope to get by induction + amplification).
> I think that both of those are hard problems, in addition to the assumption that amplification will work. But I don't yet see reason to be super pessimistic about either of them.

On the other hand, suppose we don't have exact imitation. How fast do we lose the defense against X-and-only-X? Well, that depends on the inexactness of the imitation; under what kind of distance metric is the imperfect imitation 'near' to the original? Like, if we're talking about Euclidean distance in the output, I expect you lose the X-and-only-X guarantee pretty damn fast against smart adversarial perturbations.

On the other other hand, suppose that the inexactness of the imitation is "This agent behaves exactly like Paul Christiano but 5 IQ points dumber." If this is only and precisely the form of inexactness produced, and we know that for sure, then I'd say we have a pretty good guarantee against slightly-dumber-Paul producing the likes of Rubik's Cube solvers containing hidden GSM signalers.

On the other other other hand, suppose the inexactness of the imitation is "This agent passes the Turing Test; a human can't tell it apart from a human." Then X-and-only-X is thrown completely out the window. We have no guarantee of non-Y for any Y a human can't detect, which covers an enormous amount of lethal territory, which is why we can't just sanitize the outputs of an untrusted superintelligence by having a human inspect the outputs to see if they have any humanly obvious bad consequences.

Speaking of inexact imitation: It seems to me that having an AI output a *high-fidelity* imitation of human behavior, sufficiently high-fidelity to preserve properties like "being smart" and "being a good person" and "still being a good person under some odd strains like being assembled into an enormous Chinese Room Bureaucracy", is a pretty huge ask.

It seems to me obvious, though this is the sort of point where I've been surprised about what other people don't consider obvious, that in general exact imitation is a bigger ask than superior capability. Building a Go player that imitates Shuusaku's Go

play so well that a scholar couldn't tell the difference, is a bigger ask than building a Go player that could defeat Shuusaku in a match. A human is much smarter than a pocket calculator but would still be unable to imitate one without using a paper and pencil; to imitate the pocket calculator you need all of the pocket calculator's abilities in addition to your own.

Correspondingly, a realistic AI we build that literally passes the strong version of the Turing Test would probably have to be much smarter than the other humans in the test, probably smarter than any human on Earth, because it would have to possess all the human capabilities in addition to its own. Or at least all the human capabilities that can be exhibited to another human over the course of however long the Turing Test lasts. (Note that on the version of capability amplification I heard, capabilities that can be exhibited over the course of a day are the only kinds of capabilities we're allowed to amplify.)

> **Paul Christiano**
> Totally agree, and for this reason I agree that you can't rely on perfect imitation to solve the X-and-only-X problem and hence need other solutions. If you convince me that either informed oversight or reliability is impossible, then I'll be largely convinced that I'm doomed.

An AI that learns to exactly imitate humans, not just passing the Turing Test to the limits of human discrimination on human inspection, but perfect imitation with all added bad subtle properties thereby excluded, must be so cognitively powerful that its learnable hypothesis space includes systems equivalent to entire human brains. I see no way that we're not talking about a superintelligence here.

So to postulate *perfect* imitation, we would first of all run into the problems that:

> (a) The AGI required to learn this imitation is *extremely* powerful, and this could imply a dangerous delay between when we can build any dangerous AGI at all, and when we can build AGIs that would work for alignment using perfect-imitation capability amplification.
>
> (b) Since we cannot invoke a perfect-imitation capability amplification setup to get this very powerful AGI in the first place (because it is already the least AGI that we can use to even get started on perfect-imitation capability amplification), we already have an extremely dangerous unaligned superintelligence sitting around that we are trying to use to implement our scheme for alignment.

Now, we may perhaps reply that the imitation is less than perfect and can be done with a dumber, less dangerous AI; perhaps even so dumb as to not be enormously superintelligent. But then we are tweaking the “perfection of imitation” setting, which could rapidly blow up our alignment guarantees against the standard dooms of standard machine learning paradigms.

I'm worried that you have to degrade the level of imitation *a lot* before it becomes less than an *enormous* ask, to the point that what's being imitated isn't very intelligent, isn't human, and/or isn't known to be aligned.

To be specific: I think that if you want to imitate IQ-90 humans thinking for one day, and imitate them so specifically that the imitations are generally intelligent and locally aligned even in the limit of being aggregated into weird bureaucracies, you're looking at an AGI powerful enough to think about whole systems loosely analogous to IQ-90 humans.

> **Paul Christiano**
> It's important that my argument for alignment-of-amplification goes through *not* doing problematic optimization. So if we combine that with a good enough solution to informed oversight and reliability (and amplification, and the induction working so far...), then we can continue to train imperfect imitations that definitely don't do problematic optimization. They'll mess up all over the place, and so might not be able to be competent (another problem amplification needs to handle), but the goal is to set things up so that being a lot dumber doesn't break alignment.

I think that is a very powerful AGI. I think this AGI is smart enough to slip all kinds of shenanigans past you, unless you are using a methodology that can produce faithful imitations from unaligned AGIs. I think this is an AGI that can do powerful feats of engineering, unless it is somehow able to simulate humans doing powerful feats of engineering without itself being capable of powerful feats of engineering.

And then furthermore the capability amplification schema requires the AGI to be powerful enough to learn to imitate *amplified* systems of humans as well!

If you don’t degrade the AI to well below the point of superintelligence, we require extremely strong Faraday-cage keep-God-in-a-box don't-peek-at-the-code AI boxing. So we cannot for example let an untrusted superintelligence originate queries that it can use to learn human behavior; it has to be strictly unsupervised example-based learning rather than a query model. This interacts with worries about sample complexity below.

> **Paul Christiano**
> I appreciate how often you had to argue against disjunctions here in order to give me the benefit of the doubt.

Or you need to postulate some prior alignment of the AGI that lets you robustly and safely direct it at the imitation problem. Then (a) your alignment schema depends on an unspecified prior alignment schema that already works to make very powerful AGIs do things. And (b) you probably aren't allowing unrestrained intelligence explosions, which probably introduces a big unavoidable delay between the first point where very powerful AGI is technologically possible, and the later point where you can build and safely align the extremely powerful AGI that you're using as the base unit in your imitation scheme.

The last I checked with Paul, his scheme bit the bullet on using powerful smart dangerous units to imitate humans. Paul's scheme for bypassing this, as I heard and

understood it back then, involved a scheme which:

- Trained a million smart predictors of sufficient diversity that at least one predictor would happen to end up with the true sole simple goal of predicting faithfully;
- Compared the predictors against each other, finding cases where their predictions diverged sufficiently;
- Tested them against reality;
- Had enough sampling power to perform a sufficient number of tests that, even if a superintelligence cunningly delays the time when it starts trying to slip in bad predictions, we can pause the system and run the experimental test and generate new samples and kick out the unfaithful predictor at that time.

That's a one-millionfold performance penalty, if I understood it correctly.

If that was just a concrete illustration to help me understand what's happening at all, and actually there's supposed to be some obvious less naive more efficient scheme for making use of mostly untrusted superintelligences which doesn't have the giant performance penalty, then I haven't seen it yet.

Paul Christiano
There are several possible techniques for cutting this performance penalty down, e.g., representing a large ensemble implicitly or using a tricky distillation procedure. I totally agree if you have a millionfold slowdown then you are out of luck.
When we talked about this, I agreed that "train a reliable system" was an important subproblem. Train an implicit ensemble is one possible approach to that problem. At the time of our discussion I listed 3 techniques. Since then I've become quite pessimistic about training an implicit ensemble, and have become more optimistic about verification and transparency. (My thinking about the problem is also generally much sharper.) The three approaches I currently consider most live are the ones in "Techniques for Optimizing Worst-Case Performance."
(As usual, all of it requires amplification + the induction.)
(In the original discussion, I listed "large ensemble" as a possible approach in response to your "how could you possibly solve this?" question, not as a necessary ingredient.)

I also have trouble seeing how kicking out unfaithful predictors is possible under an unsupervised-learning paradigm that doesn't actually allow hypercomputation.

The problem being: the exact output of a human is always going to be very improbable. Even if the agents can only send each other text messages and no illustrations (which further reduces their collective intelligence and ability to produce textbooks readable in a day), a text message is still an enormous possibility space. If we ask superintelligences to give us their predictions for what the human does next, their predictions cannot possibly actually include a giant lookup table next to all the probable outputs of the human.

I think I see how I'd "solve" this problem using hypercomputation and an enormous number of samples: I require the superintelligences to give me their giant lookup tables of probabilities for exact outputs, calculate total distances between probability tables (KL divergence or whatever), and if the distance is great enough, I sample a human and do a Bayesian update. I drop any superintelligences whose prior

probability goes low enough. I keep doing this until only agreeing superintelligences remain.

Under this system it looks to me like you need a *lot* of samples to kick out any superintelligences trying to sneak in bad suggestions. Let's say that an uncompressed Chinese Room bureaucracy is doing the equivalent of sending around a quadrillion messages, and that the output can be corrupted by an adversarial attack against one billion of those messages such that there's a $\frac{1}{1,000,000}$ chance of outputting an adversarially chosen alternative to each of those billion messages. Well, the only way a superintelligence can allocate a $\frac{1}{1,000,000}$ generation probability to those fake messages, is by stealing 0.0001% probability mass from all the other messages that might be output. So each time we check against a real human, we update by $\log_2 \frac{999,999}{1,000,000}$ bits, and after taking a million samples we've driven down the relative odds of this naughty superintelligence by $\frac{1}{e}$.

We had to take a million samples to do that, and each of the samples involves a human thinking for one day. And that just got us e:1 evidence, so we've got to take a million samples 7 times to get to the point of 1000:1 odds against the superintelligence being honest. And if we have a million superintelligences in the distribution, some of them can make skewed bets à la predicting coinflips whereby a thousand of those will be able to gain 1000:1 advantages over the others, so actually you've got to take 14 million samples to be sure of beating any bad predictor classes that collectively try that scheme.

I think last time I talked to Paul about this, Paul was like, "Sure, we're talking about a gigantic government project that can take millions of samples."

I'm not sure, I could be wrong about this, but I think you actually need 14 million samples per bad predictor that you want to kick out of the distribution.

Even if I'm wrong about that, we need millions of humans who are smart enough and aligned enough that random samples from their number can agglomerate into a smart aligned Chinese Room Bureaucracy.

I also don't see a way to do this without comparing the giant lookup tables using hypercomputation, but there I might just be missing a math trick.

> **Paul Christiano**
> My best guess is that this can be done under plausible assumptions with O(1) samples per bad predictor. It's tricky, but we are good at tricky math problems, so it's not very scary relative to the other problems we face. (Unfortunately, I think that a large implicit ensemble is probably stuck anyway, in part because a million predictors isn't enough. But I'm not confident about that.)

(If you're thinking of GANs, then so far as I can tell, the discriminator has to be at least as smart as the generator, and you have to trust the discriminator, and there isn't any simpler problem with respect to how you find a trusted superintelligence within a collective of untrusted ones to act as your discriminator.)

(EDIT: Actually, after thinking about this for another five minutes, maybe I do see how to do it with GANs and lower sample complexity.)

An *even larger* issue is that I don't see any obvious way to carry out a scheme like this one *at all* with respect to imperfect imitations. (And the above scheme I thought of with GANs would also just fail.)

> **Paul Christiano**
> I think we could probably get over this too, it's another tricky math problem. I think this kind of problem is reliably either impossible, or else *radically* easier than most of the other stuff we are dealing with in alignment.
> (Though I endorse the overall intuition that large implicit ensembles are doomed.)

I think these arguments are collectively something like a crux. That is, unless I've missed one of my own thought processes in the course of writing this up rapidly, or assumed a shared background assumption that isn't actually shared.

Let's say that D is the degree of imperfection allowed by some system of capability amplification, and call D-imperfect imitations D-imitations. Iterated D-imitations of amplified systems of D-imitations will be termed DD-imitations. Then I think I'd start to be pragmatically interested in capability amplification as I understood it, if I believed all of the following:

- We can, before the world is ended by other unaligned AIs, get AIs powerful enough to learn D-imitations and DD-imitations;
- D-imitations and DD-imitations robustly preserve the goodness of the people being imitated, despite the imperfection of the imitation;
- D-imitations agglomerate to sufficient cognitive power to perform a pivotal act in a way that causes the alignment of the components to be effective upon aligning the whole; and imperfect DD-imitation preserves this property;
- We can find any way of either:
    - Individually trusting one AI that powerful to faithfully perform the task of D-imitation (but then why can't we just use this scheme to align a powerful AGI in the first place?);
    - Find a scheme for agglomerating mostly untrustworthy powerful intelligences which:
        - Doesn't require giant lookup tables, doesn't require a GAN with a trusted discriminator unless you can say how to produce the trusted discriminator, and can use actual human samples as fuel to discriminate trustworthiness among untrusted generators *of D-imitations*.
        - Is extremely sample-efficient (let's say you can clear 100 people who are trustworthy to be part of an amplified-capability system, which already sounds to me like a huge damned ask); *or* you can exhibit to me a social schema which agglomerates mostly untrusted humans into a Chinese Room Bureaucracy that we trust to perform a pivotal task, and a political schema that you trust to do things involving millions of humans, in which case you can take millions of samples but not billions. Honestly, I just don't currently believe in AI scenarios in which good and trustworthy governments carry out complicated AI alignment schemas involving millions of people, so if you go down

this path we end up with different cruxes; but I would already be pretty impressed if you got all the other cruxes.
- Is not too computationally inefficient; more like 20-1 slowdown than 1,000,000-1. Because I don't think you can get the latter degree of advantage over other AGI projects elsewhere in the world. Unless you are postulating massive global perfect surveillance schemes that don't wreck humanity's future, carried out by hyper-competent, hyper-trustworthy great powers with a deep commitment to cosmopolitan value — very unlike the observed characteristics of present great powers, and going unopposed by any other major government. Again, if we go down this branch of the challenge then we are no longer at the original crux.

I worry that going down the last two branches of the challenge could create the illusion of a political disagreement, when I have what seem to me like strong technical objections at the previous branches. I would prefer that the more technical cruxes be considered first. If Paul answered all the other technical cruxes and presented a scheme for capability amplification that worked with a moderately utopian world government, I would already have been surprised. I wouldn't actually try it because you cannot get a moderately utopian world government, but Paul would have won many points and I would be interested in trying to refine the scheme further because it had already been refined further than I thought possible. On my present view, trying anything like this should either just plain not get started (if you wait to satisfy extreme computational demands and sampling power before proceeding), just plain fail (if you use weak AIs to try to imitate humans), or just plain kill you (if you use a superintelligence).

Paul Christiano
I think that the disagreement is almost entirely technical. I think if we really needed 1M people it wouldn't be a dealbreaker, but that's because of a technical rather than political disagreement (about what those people need to be doing). And I agree that 1,000,000x slowdown is unacceptable (I think even a 10x slowdown is almost totally doomed).

I restate that these objections seem *to me* to collectively sum up to “This is fundamentally just not a way you can get an aligned powerful AGI unless you already have an aligned superintelligence”, rather than “Some further insights are required for this to work in practice.” But who knows what further insights may really bring? Movement in thoughtspace consists of better understanding, not cleverer tools.

I continue to be excited by Paul’s thinking on this subject; I just don’t think it works in the present state.

Paul Christiano
On this point, we agree. I don’t think anyone is claiming to be done with the alignment problem, the main question is about what directions are most promising for making progress.

On my view, this is not an unusual state of mind to be in with respect to alignment research. I can't point to any MIRI paper that works to align an AGI. Other people seem to think that they ought to currently be in a state of having a pretty much workable scheme for aligning an AGI, which I would consider to be an odd expectation. I would think that a sane point of view consisted in having ideas for addressing some problems that created further difficulties that needed to be fixed and didn't address most other problems at all; a map with what you think are the big unsolved areas clearly marked. Being able to have a thought which *genuinely squarely attacks any alignment difficulty at all* despite any other difficulties it implies, is already in my view a large and unusual accomplishment. The insight "trustworthy imitation of human external behavior would avert many default dooms as they manifest in external behavior unlike human behavior" may prove vital at some point. I continue to recommend throwing as much money at Paul as he says he can use, and I wish he said he knew how to use larger amounts of money.

Terrorism, Tylenol, and dangerous information

Recently, there has been an alarming development in the field of terrorist attacks; more and more terrorists seem to be committing attacks via crashing vehicles, often large trucks, into crowds of people. This method has several advantages for an attacker - it is very easy to obtain a vehicle, it is very difficult for police to protect against this sort of attack, and it does not particularly require special training on the part of the attacker.

While these attacks are an unwelcome development, I would like to propose an even more worrisome question - *why didn't this happen sooner?*

I see no reason to believe that there has been any particular technological development that has caused this method to become prevalent recently; trucks have been in mass production for over a hundred years. Similarly, terrorism itself is not particularly new - just look to the anarchist attacks of the late 19th and early 20th century. Why, then, weren't truck attacks being made earlier?

The answer, I think, is both simple and frightening. *The types of people who make attacks hadn't thought of it yet.* The main obstacle to these attacks was psychological and intellectual, not physical, and once attackers realized these methods were effective the number of attacks of this sort began increasing. If the Galleanists had realized this attack method was available, they might well have done it back in '21 -- but they didn't, and indeed nobody motivated to carry out these attacks seemed to until much later.

Another instance - though one with less lasting harm - pertains to Tylenol. In 1982, a criminal with unknown motives tampered with several Tylenol bottles, poisoning the capsules with cyanide and then replacing them on store shelves. Seven people died in the original attack, which caused a mass panic to the point where police cars were sent to drive down the streets broadcasting warnings against Tylenol from their loudspeakers; more people still were killed in later "copycat" crimes.

In this case, there was a better solution than with the truck rammings - in the aftermath of these events, greatly increased packaging security was put into place for over-the-counter medications. Capsules (which are comparatively easy to adulterate) fell out of favor somewhat in favor of tablets; further, pharmaceutical companies began putting tamper-resistant seals on their products and the government made product tampering a federal offense. Such attacks are now much harder to commit.

However, the core question remains - why was it that it took until 1982 for there to be a public attack like this, and then there were many more (TIME claims hundreds!) in short succession? *The types of people who make attacks hadn't thought of it yet.* Once the first attack and the panic around it exposed this vulnerability, opportunistic attackers carried out their own plans, and swift action suddenly became necessary - swift action to close a security hole that had been open for years and years!

One practical implication of this phenomenon is quite worrisome - one must be very careful to avoid accidentally spreading dangerous information. If the main constraint on an attack vector can really just be that the types of people who make attacks haven't thought of it yet, it's very important to avoid spreading knowledge of potential

ways in which we're vulnerable to these attacks - you might wind up giving the wrong person dangerous ideas!

Many otherwise analytical or strategic thinkers that I have encountered seem to fall prey to the typical mind fallacy in these cases, assuming that others will also have put thought into these things and thus that there's no real risk in discussing them - after all, these methods are "obvious" or even "publicly known". Certainly I have made this mistake myself before!

However, what is "publicly known" in some book or white paper somewhere may only be practically known by a few people. Openly discussing such matters, especially online, risks many more people seeing it than otherwise would. Further, I would generally say that the types of people who make attacks are cunning but unimaginative. They are able to execute existing plans fairly effectively, but are comparatively unlikely to come up with novel methods. This means that there's extra reason to be wary that you might have come up with something they haven't.

Thus, when dealing with potentially dangerous information, care should be taken to prevent it from spreading. That doesn't, of course, mean that you can't talk these matters over with trusted colleagues or study to help prepare defenses and solve vulnerabilities - but it does mean that you should be careful when doing so.

As strange as it seems, it is very possible that the only reason things haven't gone wrong in just the way you're thinking of is that dangerous people haven't thought of it yet - and if so, you don't want to be the one giving them ideas!

Author's note: Sincere thanks to those who assisted me with this post; their assistance has made it safer and more compelling.

Meta-Honesty: Firming Up Honesty Around Its Edge-Cases

(Cross-posted from Facebook.)

0: Tl;dr.

- A problem with the obvious-seeming "wizard's code of honesty" aka "never say things that are false" is that it draws on high verbal intelligence and unusually permissive social embeddings. I.e., you can't always say "Fine" to "How are you?" This has always made me feel very uncomfortable about the privilege implicit in recommending that anyone else be more honest.
- Genuinely consistent Glomarization (i.e., consistently saying "I cannot confirm or deny" whether or not there's anything to conceal) does not work in principle because there are too many counterfactual selves who might want to conceal something.
- Glomarization also doesn't work in practice if the Nazis show up at your door asking if you have fugitive Jews in your attic.
- If you would lie to Nazis about fugitive Jews, then absolute truthsaying can't be the whole story, which makes "never say things that are false" feel to me like a shaky foundation in that it is literally false, and something less shaky would be nice.
- Robin Hanson's "automatic norms" problem suggests different people might have very different ideas about what constitutes a good person's normal honesty, without realizing that they have very different ideas. Perceived violations of an honesty norm can blow up and cause interpersonal conflict. It seems to me that this is something that doesn't always work well when people leave it alone.

A rule which seems to me more "normal" than the wizard's literal-truth rule, more like a version of standard human honesty reinforced around the edges, would be as follows:

"Don't lie when a normal highly honest person wouldn't, and furthermore, be honest when somebody asks you which hypothetical circumstances would cause you to lie or mislead—absolutely honest, if they ask under this code. However, questions about meta-honesty should be careful not to probe object-level information."

I've been tentatively calling this "meta-honesty", but better terminology is solicited.

1: Glomarization can't practically cover many cases.

Suppose that last night I helped hide a fugitive marijuana seller from the Feds. You ask me what I was doing last night, and I, preferring not to emit false statements, reply, "I can't confirm or deny what I was doing last night."

We now have two major problems here:

- Even on an ordinary day, if you casually ask me what I was doing last night, I theoretically ought to answer "I can't confirm or deny what I was doing last night" because some of my counterfactual selves were hiding fugitive marijuana sellers from the Feds. If I don't do this consistently, and I actually was hiding fugitives last night, I can't Glomarize without revealing information. But then the number of counterfactuals I have to worry about is too large for me to ever answer anything.
- If the Feds actually ask you this question, they will not be familiar with your previous practice of Glomarization and will probably not be very impressed with your answer.

This doesn't mean that Glomarization is never helpful. If you ask me whether my submarine is carrying nuclear weapons, or whether I'm secretly the author of "The Waves Arisen", I think most listeners would understand if I replied, "I have a consistent policy of not saying which submarines are carrying nuclear weapons, nor whether I wrote or helped write a document that doesn't have my name on it." An ordinary honest person does not need to lie on these occasions because Glomarization is both theoretically possible and pragmatically practical, so one should adopt a consistent Glomarization rather than lie.

But that doesn't work for hiding fugitives. Or any other occasion where an ordinary high-honesty person would consider it obligatory to lie, in answer to a question where the asker is not expecting evasion or Glomarization.

(I'm sure some people reading this think it's all very cute for me to be worried about the fact that I wouldn't tell the truth all the time. Feel free to state this in the comments so that we aren't confused about who's using which norms. Smirking about it, or laughing, especially conveys important info about you.)

2: The law of no literal falsehood.

One formulation of my automatic norm for honesty, the one that feels like the obvious default from which any departure requires a crushingly heavy justification, was given by Ursula K. LeGuin in *A Wizard of Earthsea*:

> He told his tale, and one man said, "But who saw this wonder of dragons slain and dragons baffled? What if he—"
>
> "Be still!" the Head Isle-Man said roughly, for he knew, as did most of them, that a wizard may have subtle ways of telling the truth, and may keep the truth to himself, but that if he says a thing the thing is as he says. For that is his mastery.

Or in simpler summary, this policy says:

Don't say things that are literally false.

Or with some of the unspoken finicky details added back in: "Don't say things that you believe to be literally false in a context where people will (with reasonably high probability) persistently believe that you believe them to be true." Jokes are still

allowed, even jokes that only get revealed as jokes ten seconds later. Or quotations, etcetera ad obviousum.

The no-literal-falsehood code of honesty has three huge advantages:

- To the extent people observe you to consistently practice it, it is easier for you to communicate believably when you *want* to say a thing. They may still not be able to trust you perfectly, but the hypothetical is "Did this person break their big-deal code of honesty?" rather than "Did this person tell an ordinary lie?" One would hope this would be good for coordination and other interpersonal issues, though this might only be a fond wish on my part.
- Most people, even most unusually honest people, wander about their lives in a fog of internal distortions of reality. Repeatedly asking yourself of every sentence you say aloud to another person, "Is this statement actually and literally true?", helps you build a skill for navigating out of your internal smog of not-quite-truths. For that is our mastery.
- It's good for your soul. At least, it's good for my soul for reasons I'd expect to generalize if I'm not just committing the typical-mind fallacy.

From Frank Hebert's *Dune Messiah*, writing about Truthsayers, people who had trained to extreme heights the ability to tell when others were lying and who also never lied themselves:

> "It requires that you have an inner agreement with truth which allows ready recognition."

This is probably not true in normal human practice for detecting other people's lies. I'd expect a lot of con artists are better than a lot of honest people at that.

But the phrase "It requires you have an inner agreement with truth which allows ready recognition" is something that resonates strongly with me. It feels like it points to the part that's good for your soul. Saying only true things is a kind of respect for the truth, a pact that you forge with it.

3: The privilege of truthtelling.

I've never suggested to anyone else that they adopt the wizard's code of honesty.

The code of literal truth only lets people navigate anything like ordinary social reality to the extent that they are very fast on their verbal feet, and can respond to the question "How are you?" by saying "Getting along" instead of "Horribly" or with an awkward silence while they try to think of something technically true. (Because often "I'm fine" is false, you see. If this has never bothered you then you are perhaps not in the target audience for this essay.)

So I haven't advocated any particular code of honesty before now. I was aware of the fact that I had an unusually high verbal SAT score, and also, that I spend little time interfacing with mundanes and am not dependent on them for my daily bread. I thought it wasn't my place for me to suggest to anyone else that they try their hand at saying only true things all the time, or for me to act like this conveys moral virtue. I'm only even describing the wizard's code publicly now that I can think of at least one alternative.

I once heard somebody claim that rationalists ought to practice lying, so that they could separate their internal honesty from any fears of needing to *say* what they believed. That is, if they became good at lying, they'd feel freer to consider geocentrism without worrying what the Church would think about it. I do not in fact think this would be good for the soul, or for a cooperative spirit between people. This is the sort of proposed solution of which I say, "That is a terrible solution and there has to be a better way."

But I do see the problem that person was trying to solve. One can also be privileged in stubbornness when it comes to overriding the fear of other people finding out what you believe. I can see how telling fewer routine lies than usual would make that fear even worse, exacerbating the pressure it can place on what you believe you believe; especially if you didn't have a lot of confidence in your verbal agility. It's one more reason not to pressure people (even a little) into adopting the wizard's code, but then it would be nice to have some other code instead.

4: Literal-truth as my automatic norm, maybe not shared.

This set of thoughts started, as so many things do, with a post by Robin Hanson.

In particular Robin tweeted the paper: "The surprising costs of silence: Asymmetric preferences for prosocial lies of commission and omission."

> Abstract: Across 7 experiments (N = 3883), we demonstrate that communicators and targets make egocentric moral judgments of deception. Specifically, communicators focus more on the costs of deception to them—for example, the guilt they feel when they break a moral rule—whereas targets focus more on whether deception helps or harms them. As a result, communicators and targets make asymmetric judgments of prosocial lies of commission and omission: Communicators often believe that omitting information is more ethical than telling a prosocial lie, whereas targets often believe the opposite.

This got me wondering whether my default norm of the wizard's code is something other people will even perceive as prosocial. Yes, indeed, I feel like not saying things is much more law-abiding than telling literal falsehoods. But if people feel just as wounded, or *more* wounded, then that policy isn't really benefiting anyone else. It's just letting me feel ethical and maybe being good for my own personal soul.

Robin commented, "Mention all relevant issues, even if you have to lie about them."

I don't think this is a bullet I can bite in daily practice. I think I still want to emit literal truths for most dilemmas short of hiding fugitives. But it's one more argument worth mentioning against trying to make an absolute wizard's code into a bedrock solution for interpersonal reliability.

Robin also published a blog post about "automatic norms" in general:

> We are to just know easily and surely which actions violate norms, without needing to reflect on or discuss the matter. We are to presume that framing

> effects are unimportant, and that everyone agrees on the relevant norms and how they are to be applied.
>
> In a relatively simple world with limited sets of actions and norms, and a small set of people who grew up together and later often enough observe and gossip about possible norm violations of others, such people might in fact learn from enough examples to mostly apply the same norms the same way. This was plausibly the case for most of our distant ancestors. They could in fact mostly be sure that, if they judged themselves as innocent, most everyone else would agree. And if they judged someone else as guilty, others should agree with that as well. Norm application could in fact usually be obvious and automatic.
>
> Today however, there are far more people, and more intermixed, who grow up in widely varying contexts and now face far larger spaces of possible actions and action contexts. Relative to this huge space, gossip about particular norm violations is small and fragmented...
>
> We must see ourselves as tolerating a *lot* of norm violation. We actually tell others about and attempt to punish socially only a tiny fraction of the violations that we could know of. When we look most anywhere at behavior details, it must seem to us like we are living in a Sodom and Gomorrah of sin. Compared to the ancient world, it must seem a lot easier to get away for a long time with a lot of norm violations...
>
> We must also see ourselves as tolerating a *lot* of overeager busybodies applying what they see as norms to what we see as our own private business where their social norms shouldn’t apply.

This made me realize that the wizard's code of honesty I grew up with is, indeed, an automatic norm for me. Which meant I was probably overestimating and eliezeromorphizing the degree to which other people even cared at all, or would think I was keeping any promises by doing it. Again, I don't see this as a good reason to give up on emitting literally true sentences almost all of the time, but it's one more reason I feel more open to alternatives than I would've ten years ago. That said, I do expect a lot of people reading this also have something like that same automatic norm, and I still feel like that makes us more like part of the same tribe.

5: Counterargument: The problem of non-absolute rules.

A proposal like this one ought to come with a lot of warning signs attached. Here's one of them:

There's a passage in John M. Ford's *Web of Angels*, when the protagonist has finally killed someone even after all the times his mentor taught him to never ever kill. His mentor says:

> "No words can prevent all killing. Words are not iron bands. But I taught you to hesitate, to stay your hands until the weight of duty crushed them down."

Surprise! Really the mentor just meant to try to get him to *wait* before killing people instead of jumping to that *right away*.

Humans are kind of insane, and there are all sorts of insane institutions that have evolved among us. A fairly large number of those institutions are twisted up in such a way that something explodes if people try to talk openly about how they work.

It's a human kind of thinking to verbally insist that "Don't kill" is an absolute rule, why, it's right up there in the Ten Commandments. Except that what soldiers do doesn't count, at least if they're on the right side of the war. And sure, it's also okay to kill a crazy person with a gun who's in the middle of shooting up a school, because that's just not what the absolute law "Don't kill" *means*, you know!

Why? Because any rule that's not labeled "absolute, no exceptions" lacks weight in people's minds. So you have to perform that the "Don't kill" commandment is absolute and exceptionless (even though it totally isn't), because that's what it takes to get people to even *hesitate*. To stay their hands *at least* until the weight of duty is crushing them down. A rule that isn't even absolute? People just disregard that whenever.

(I speculate this may have to do with how the human mind reuses physical ontology for moral ontology. I speculate that brains started with an ontology for material possibility and impossibility, and reused that ontology for morality; and it internally feels like only the moral reuse of "impossible" is a rigid moral law, while anything short of "moral-impossible" is more like a guideline. Kind of like how, if something isn't absolutely certain, people think that means it's okay to make up their own opinion about it, because if it's not absolutely certain it must not be the domain of Authority. But I digress, and it's just a hypothesis. We don't need to know exactly what is the buried cause of the surface craziness to observe that the craziness is in fact there.)

So you have to perform that the Law is absolute in order to make the actual flexible Law exist. That doesn't mean people lie about how the Law applies to the edge cases—that's not what I mean to convey by the notion of "performing" a statement. More like, proclaim the Law is absolute and then *just not talk about* anything that contradicts the absoluteness.

And when that happens, it's one more little chunk of insanity that nobody can talk about on the meta-level without it exploding.

Now, you will note that I am going ahead and writing this all down explicitly, because... well, because I expect that in the long run we have to find a way that doesn't require a little knot of madness that nobody is allowed to describe faithfully on the meta-level. So we might as well start today.

I trust that you, the reader, will be able to understand that "Don't kill" is the kind of rule where you give it enough force-as-though-of-absoluteness that it actually takes a deontology-breaking weight of duty to crush down your hands, as opposed to you cheerfully going "oh well I guess there's a crushing weight now! let's go!" at the first sign of inconvenience.

Actually, I don't trust that everyone reading this can do that. That's not even close to literally true. But most you won't ever be called on to kill, and society frowns upon that strongly enough to discourage you anyway. So I did feel it was worth the risk to write that example explicitly.

"Don't lie" is more dangerous to mess with. That's something that most people don't take as an exceptionless absolute to begin with, even in the sense of performing its absoluteless so that it will exist at all. Even extremely honest people will agree that you can lie to the Gestapo about whether you are hiding any Jews in the attic, and not bother to Glomarize your response either; and I think they will mostly agree that this is in fact a "lie" rather than trying to dance around the subject. People who are less than extremely honest think that "I'm fine" is an okay way to answer "How are you?" even if you're not fine.

So there's still a very obvious thing that could go wrong in people's heads, a very obvious way that the notion of "meta-honesty" could blow up, or *any other code*besides "don't say false things" could blow up. It's why the very first description in the opening paragraphs says "Don't lie when a normal highly honest person wouldn't, and furthermore..." and you should never omit that preamble if you post any discussion of this on your own blog. THIS IS NOT THE IDEA THAT IT'S OKAY TO LIE SO LONG AS YOU ARE HONEST ABOUT WHEN YOU WOULD LIE IF ANYONE ASKS. It's not an escape hatch.

If anything, meta-honesty is the idea that you should be careful enough about when you break the rule "Don't lie" that, if somebody else asked the hypothetical question, you would be willing to PUBLICLY DEFEND EVERY ONE OF THOSE EXTRAORDINARY EXCEPTIONS as times when even an unusually honest person should lie.

(Unless you were never claiming to be unusually honest, and your pattern of meta-honest responses to hypotheticals openly shows that you lie about as much as an average person. But even here, I'd worry that anyone who lets themselves be as wicked as they imagine the 'average' person to be, would be an unusually wicked person indeed. After all, if Robin Hanson speaks true, we are constantly surrounded by people violating what seem to us like automatic norms.)

6: Meta-honesty, the basics.

Okay, enough preamble, let's speak of the details of meta-honesty, which may or may not be a terrible idea to even talk about, we don't know at this point.

The basic formulation of meta-honesty would be:

"Be at least as honest as an unusually honest person. Furthermore, when somebody asks for it and especially when you believe they're asking for it under this code, try to convey to them a frank and accurate picture of the sort of circumstances under which you would lie. Literally never swear by your meta-honesty that you wouldn't lie about a hypothetical situation that you would in fact lie about."

My first horrible terminology for this was the "Bayesian code of honesty", on the theory that this code meant your sentences never provided Bayesian evidence in the wrong direction. Suppose you say "Hey, Eliezer, what were you doing last night?" and I reply "Staying at home doing the usual things I do before going to bed, why?" If you have a good mental picture of what I would lie about, you have now definitely learned that I was *not* out watching a movie, because that is not something I would lie about. A very large number of possibilities have been ruled out, and most of your remaining probability mass should now be on me having stayed home last night. You know that I

wasn't on a secret date with somebody who doesn't want it known we're dating, because you can ask me that hypothetical and I'll say, "Sure, I'd happily hide that fact, but that isn't enough to force me to *lie*. I would just say 'Sorry, I can't tell you where I was last night,' instead of lying."

You have *not* however gained any Bayesian evidence against my hiding a fugitive marijuana seller from the Feds, where somebody's life or freedom is at stake and it's vital to conceal that a secret even exists in the first place. Ideally we'd have common knowledge of that, and hopefully we'd agree that it was fine to lie in that case to a friend who asks a casual-seeming question.

Let's be clear, although this is a kind of softening of deception, it's still deception. Even if somebody has extensively discussed your code of honesty with you, they aren't logically omniscient and won't explicitly have the possibility in mind every time. That's why we should go on holding ourselves to the standard of, "Would I defend this lie even if the person I was defending it to had never heard of meta-honesty?"

"Eliezer," you say, "if you had a temporary schizophrenic breakdown and robbed a bank and this news hadn't become public, would you lie to keep it from becoming public?"

And this would cause me to stop and think and agonize for a bit (which itself tells you something about me, that my answer is not instantly No or Yes). I do have important work to do which should not be trashed without strong reason, and this hypothetical situation would not have involved a great deliberate betrayal on my part; but it is also the sort of thing that you could reasonably argue an unusually honest person ought *not* to lie about, where lies do not in general serve the social good.

I think in the end I might reply something like "I wouldn't lie freely and would probably try to use at least technical truth or Glomarize, but in the end I might conceal that event rather than letting my work be trashed for no reason. I think I'd understand if somebody else had done likewise, if I thought they were doing good work in the first place. Except that obviously I'd need to tell various people who are engaged in positive-sum trades with me, where it's a directly important issue to them whether I can be trusted never to have mental breakdowns, and remove myself from certain positions of trust. And if it happened twice I'd be more likely to give up. If it got to the point where people were openly asking questions I don't imagine myself as trying to continue a lie. I also want to caveat that I'm describing my ethical views, what I think is right in this situation, and obviously enough pressure can make people violate their own ethics and it's not always predictable how much pressure it takes, though I generally consider myself fairly strong in that regard. But if this had actually happened I would have spent a lot more time thinking about it than the two minutes I spent writing this paragraph." And this would help give you an accurate picture of the sort of person that I am in general, and what I take into account in considering exceptions.

Insofar as you are practicing a mental discipline in being meta-honest, the discipline is to be explicitly aware of every time you say something false, and to ask yourself, "Would I be okay publicly saying, if somebody asked me the hypothetical, that this is a situation where a person ought to lie?"

I still worry that this is *not* the thing that people need to do to establish their inner pact with truth. Maybe you could pick some friends to whom you just never tell any kind of literal falsehood, in the process of becoming initially aware of how many false

things you were just saying all the time… but I don't actually know if that works either. Maybe that's like trying to stop smoking cigarettes on odd-numbered days. It'd be something to notice if the experimental answer is "In reality, meta-honesty turns out not to work for practicing the respect of truth."

Meta-honesty should be for people who are comfortable, not with absolute honesty, but with not trying to appear any more honest than they are. This itself is not the ordinary equilibrium, and if you want to do things the standard human way and not forsake a well-tested and somewhat enforced social equilibrium in pursuit of a bright-eyed novel idealistic agenda, then you should not declare yourself meta-honest, or should let somebody else try it first.

7: Consistent object-level glomarization in meta-level honest responses.

Glomarization can be workable when restricted to special cases, such as only questions about nuclear weapons and submarines. Meta-honesty is such a special case and, if we're doing this, we should all Glomarize it accordingly. In particular meta-questions are not to be used to extract object-level data, and we should all respect that in our questions, and consistently Glomarize about it in our answers, including some random times when Glomarization seems silly.

Some key responses that need to be standard:

- "That question sounds too object-level."
- "I think you're doing meta-honesty wrong."
- "I think I'm supposed to Glomarize that sort of answer in general."
- "I should answer a more abstract version of that."
- "I worry that some of my counterfactual selves are not in a mutually beneficial situation in this discussion."

And if you clearly say that you "irrevocably worry" about any of these things, it means the meta-honest conversation has crashed; the other person is not supposed to keep pressing you, and if they do, you can lie. Ideally, this is something you should consistently do in any case where a substantial measure of your counterfactual selves as the other person might imagine them would be feeling pressured to the point of maybe meta-lying. That is, you should not only say "irrevocably worry" in cases where you actually have something to conceal, you should say it in cases where the discussion would be pressuring somebody who did have something to conceal and this seems high-enough-probability to you or to your model of the person talking to you.

For example: "Eliezer, would you lie about having robbed a bank?"

I consider whether this sounds like an attempt to extract object-level information from some of my counterfactual selves, and conclude that you probably place very little probability on my having actually robbed a bank. I reply, "Either it is the case that I did rob a bank and I think it is okay to lie about that, or alternatively, my reply is as follows: I wouldn't ordinarily rob a bank. It seems to me that you are postulating some extraordinary circumstance which has driven me to rob a bank, and you need to tell me more about this extraordinary circumstance before I tell you whether I'd lie about

it. Or you're postulating a counterfactual version of me that's fallen far enough off the ethical rails that he'd probably stop being honest too."

Some additional statements that ought to be taken as praiseworthy:

- "I only feel free to have a frank discussion about that if everyone in the room has agreed to abide by the meta-honesty code."
- "I notice that I'm feeling interrogated, and should not try to give a code-abiding answer to that right now."
- "It is either the case that this actually happened and I think it is okay to lie about it, or that my current quick guess is that I wouldn't lie in that case."
- "Hold on, let me either generate a random number or pretend to generate a random number, such that if I'm actually generating a random number and it comes up as 0, I will try to seem more evasive than usual in this conversation even if I have nothing to actually hide."

This is not *supposed* to be a clever way to extract information from people and you should shut down any attempt to use it that way.

"Harry," says HPMOR!Dumbledore, "I ask you under the code of meta-honesty (which we have just anachronistically acquired): Would you lie about having robbed the Gringotts Bank?"

Harry thinks, *Maybe this is about the Azkaban breakout*, and says, "Do you in fact suspect me of having robbed a bank?"

"I think that if I suspected you of having robbed a bank," says Dumbledore, "and I did not wish you to know that, I would not ask you if you had robbed a bank. Why do you ask?"

"Because the circumstances under which you're invoking meta-honesty have something to do with how I answer," says Harry (who has suddenly acquired a view on this subject that some might consider implausibly detailed). "In particular, I think I react differently depending on whether this is basically about you trying to construct a new mutually beneficial arrangement with the person you think I am, or if you're in an adversarial situation with respect to some of my counterfactual selves (where the term 'counterfactual' is standardly taken to include the actual world as one that is counterfactually conditioned on being like itself). Also I think it might be a good idea generally that the first time you try to have an important meta-honest conversation with someone, you first spend some time having a meta-meta-honest conversation to make sure you're on the same page about meta-honesty."

"I am not sure I understood all that," said Dumbledore. "Do you mean that if you think we have become enemies, you might meta-lie to me about when you would lie?"

Harry shook his head. "No," said Harry, "because then if we weren't enemies, you would still never really be able to trust what I say even assuming me to abide by my code of honesty. You would have to worry that maybe I secretly thought you were an enemy and didn't tell you. But the fact that I'm meta-honest shouldn't be something that you can use against me to figure out whether I... sneaked into the girl's dorm and wrote in somebody's diary, say. So if I'm in that situation I've got to protect my counterfactual selves and Glomarize harder. Whereas if this is more of a situation where you want to know if we can go to Mordor together, then I'd feel more open and try to give you a fuller picture of me with more detail and not worry as much about Glomarizing the specific questions you ask."

"I suspect," Dumbledore said gravely, "that those who try to be honest at all will always be at something of a disadvantage relative to the most ready liars, at least if they've robbed Gringotts. But yes, Harry, I am afraid that this is more of a situation where I am... concerned... about some of your counterfactual selves. But then why would you answer at all, in such a case?"

"Because sometimes people are honest and have good intentions," answered Harry, "and I think that if in general they can have an accurate picture of the other person's honesty, everybody is on net a bit better off. Even if I *had* robbed a bank, for example, you and I would both still not want anything bad to happen to Britain. And some of my counterfactual selves are innocent, and they're not better off if you think I'm more dishonest than I am."

"Then I ask again," said Dumbledore, "under the code of meta-honesty, whether you would lie about having robbed a bank."

"Then my answer is that I wouldn't ordinarily rob a bank," Harry said, "and I'd feel even worse about lying about having robbed a bank, than having robbed a bank. And I'd know that if I robbed a bank I'd also have to lie about it. So whatever weird reason made me rob the bank, it'd have to be weird enough that I was willing to rob the bank *and* willing to lie about it, which would take a pretty extreme situation. Where it should be clear that I'm not trying to answer about having specifically robbed a bank, I'm trying to give you a general picture of what sort of person I am."

"What if you had been blackmailed into robbing the bank?" inquired Dumbledore. "Or what if things crept up on you bit by bit, so that in the end you found yourself in an absurd situation you'd never intended to enter?"

Harry shrugged helplessly. "Either it's the case that I did end up in a weird situation and I don't want to let you know about that, or alternatively, I feel like you're describing a very broad range of possibilities that I'd have to think about more, because I haven't yet ended up in that kind of situation and I'm not quite sure how I'd behave... I think I'd have in mind that just telling the Headmaster the truth can prevent big problems from blowing up any further, but there'd be cases extreme enough that I wouldn't do that either... I mean, the basic answer is, yes, there's things that would make me lie right to your face, but, I wouldn't do that just for having stolen candy from the kitchen, I don't think. I'd just be like 'I consistently Glomarize when people ask me if I've stolen candy from the kitchen.'"

"Would you lie to protect a fugitive from the law who you believed was being unjustly pursued?" said Dumbledore. "Say that Sirius Black was innocent after all, as the Quibbler would have it; would you hide him and then lie to Minerva about what you were doing on the night you hid him?"

"Yes," Harry said unhesitatingly. "That's something I think is clearly right and the sort of thing I would ordinarily do."

"What if you thought that somebody was being unjustly pursued by the law?" said Dumbledore. "Is that worth lying for?"

Harry frowned. "I think I maybe shouldn't answer questions about what in particular I think ought not to be illegal," Harry said, "but if the basic question is whether I'd lie to my friends to keep somebody who'd only used marijuana out of Azkaban, then I would. I don't think my friends should expect otherwise; they should want me to do that."

"We are not being meta-honest any more, only honest," said Dumbledore. "Were you involved in the breakout from Azkaban?"

Harry stared at Dumbledore a moment, and said, "I think I have to consistently not confirm or deny that, but my counterfactual and possibly actual self who wasn't involved would really like to know at this point what the *heck* you think might've happened."

If somehow the Gestapo has heard of meta-honesty and they ask you if you would lie to hide Jews in your attic, they *should* be expecting a law-abiding meta-honest Nazi citizen to Glomarize, abstract the question in such a way as to not reveal the object-level information that might be sought, or say "I irrevocably worry that some of my counterfactual selves are not in a positive-sum situation." That, I now proclaim, ought to be done by Nazis and resistance fighters alike, when you are faced with a question that *might* be meant to reveal object-level information about what happened.

"Eliezer," says the hypothetical Gestapo officer who has somehow heard about my meta-honesty code, "it happens that I'm a person who's heard of meta-honesty. Now, are you the sort of person who would lie about having Jews hidden in your attic?"

This hypothetical Gestapo officer has a gun. Most people asking you meta-honest questions won't have a gun. In fact I bet this will literally never happen until the end of the world. Let's suppose he has a gun anyway.

"I am the following sort of person," I reply. "If I was hiding the Führer in my attic to protect him from Jewish assassins, I'd lie about that to the assassins. It's clear you know about my code of meta-honesty, so you should understand that is a very innocent thing to say. But these circumstances and the exact counterfactual you are asking make me nervous, so I'm afraid to utter the words I think you may be looking for, namely the admission that if I were the kind of person who'd hide Jews in his attic then I'd be the kind of person who would lie to protect them. Can I say that I believe that in respect to your question as you mean it, I think that is no more and no less true of me than it is true of you?"

"My, you are fast on your verbal feet," says the Gestapo officer. "If somebody were less fast on their verbal feet, would you tell them that it was acceptable for a meta-honest person to just meta-lie to the Jewish assassins in order to hide the Führer?"

"If they didn't feel that their counterfactual loyal Nazi self would think that their counterfactual disloyal self was being pressured and clearly state that fact irrevocably," I say, "I'd say that, just like their counterfactual loyal self, they should make some effort to reveal the general limits of their honesty without betraying any of their counterfactual selves, but say they irrevocably couldn't handle the conversation as soon as they thought their alternate loyal self would think their alternate's counterfactual disloyal self couldn't handle the conversation. It's not as if the Jewish assassins would be fooled if they said otherwise. If the Jewish assassins do continue past that point, which is blatantly forbidden and everyone should know that, they may lie."

"I see," says the Gestapo officer. "If you are telling me the truth, I think I have grasped the extent of what you claim to be honest about." He turns to his subordinates. "Go search his attic."

"Now I'm curious," I say. "What would you have done if I'd sworn to you that I was an absolutely loyal German citizen, and that my character was such that I would certainly

never lie about having Jews in my attic even if I were the sort of disloyal citizen who had Jews in his attic in the first place?"

"I would have detailed twice as many men to search your house," says the Gestapo officer, "and had you detained, for that is not the response I would expect from an honest Nazi who knew how meta-honesty was supposed to work. Now I ask you meta-meta-honestly, why haven't you said that you are irrevocably worried that I am abusing the code? Obviously I put substantial probability on you being a traitor, meaning I am deliberately pressuring you into a meta-conversation and trying to use your code of honesty against those counterfactual selves. Why didn't you just shut me down?"

"Because you do have a gun, sir," I say. "I agree that it's what the rules called for me to say, but I thought over the situation and decided that I was comfortable with saying that in general this was a sort of situation where that rule could be bent so as for me to not end up being shot—and I tell you meta-meta-honestly that I *do* believe the situation has to be that extreme in order for that rule to even be bent."

Really the principle is that it is not okay to meta-ask what the Gestapo officer is meta-asking here. This kind of detailed-edge-case-checking conversation might be appropriate for shoring up the edges of an interaction intended to be mutually beneficial, but absolutely not for storming in looking for Jews in the attic of a person who in your mind has a lot of measure on having something to hide.

But I do want to have trustworthy foundations somewhere.

And I think it's reasonable to expect that over the course of a human lifetime you will *literally never* end up in a situation where a Gestapo officer who has read this essay is pointing a gun at you and asking overly-object-level-probing meta-honesty questions, and will shoot you if you try to glomarize but will believe you if you lie outright, given that we *all know* that everyone, innocent or guilty, is supposed to glomarize in situations like that. Up until today I don't think I've ever seen any questions like this being asked in real life at all, even hanging out with a number of people who are heavily into recursion.

So if one is declaring the meta-honesty code at all, then one shouldn't meta-lie, period; I think the rules have been set up to allow that to be absolute. I don't want you to have to worry that maybe I think I'm being pressured, or maybe I thought you meta-asked the wrong thing, so now I think it's okay to meta-lie even though I haven't given any outward sign of that. To that end, I am willing to sacrifice the very tiny fraction of the measure of my future selves who will end up facing an extremely weird Gestapo officer. To me, for now, there doesn't seem to be *any* real-life circumstance where you should lie in response to a meta-honesty question—rather than consistently glomarize that kind of question, consistently abstract that kind of question, consistently answer in an analogy rather than the original question, or consistently say "I believe some counterfactual versions of me would say that cuts too close to the object level." (It being a standard convention that counterfactuals may include the actual.)

I also think we can reasonably expect that from now until the end of the world, honest people should literally absolutely never need to *evade or mislead at all* on the meta-meta-level, like if somebody asks if you feel like the meta-level conversation has abided by the rules. (And just like meta-honesty doesn't excuse object-level dishonesty, by saying that meta-meta-honesty seems like it could be everywhere

open and total, I don't mean to excuse meta-level lies. We should all still regard meta-lies as extremely bad and a Code Violation and You Cannot Be Trusted Anymore.)

If there's a meta-honest discussion about someone's code of honesty, *and* a discussion of what they think about the current meta-meta conditions of how the meta-honesty code is being used, and it sounds to you like they think things are fine… then things should be fine, period. If you ask, do they think that any pressure strong enough to potentially shake their meta-honesty is potentially around, do they think that the overall situation here would have treated any of their plausible counterfactual selves in a negative-sum way, and they say no it's all fine—then that is supposed to be absolute under the code. That ought to establish a foundation that's as reliable as the person's claim to be meta-honest at all.

If you go through all that and lie and meta-lie and meta-meta-lie after saying you wouldn't, you've lied under some of the kindest environments that were ever set up on this Earth to let people not lie, among people who were trying to build trust in that code so we could all use it together. You are being a genuinely awful person as I'd judge that, and I may advocate for severe social sanctions to apply.

Assuming this ends up being a thing, that is. I haven't run it past many people yet and this is the first public discussion. Maybe there's some giant hole in it I haven't spotted.

If anybody ever runs into an actual real circumstance where it seems to them that meta-honesty as they tried to use it was giving the essay-reading Gestapo too much power or too much information, maybe because they weren't fast enough on their verbal feet, please email me about it so I can consider whether to modify or backtrack on this whole idea. I will try to protect your anonymity under all circumstances up to and including the end of the world unless you say otherwise. The previous sentence is not the sort of thing I would lie about.

8: Counterargument: Maybe meta-honesty is too subtle.

I worry that the notion of meta-honesty is too complicated and subtle. In that it has subtleties in it, at all.

This concept is certainly too subtle for Twitter. Maybe it's too subtle for us too.

Maybe "meta-honesty" is just too complicated a concept to be able to make it be part of a culture's Law, compared to the standard-twistiness-compliant performance of saying "Always be honest!" and waiting for the weight of duty to crush down people's hands, or saying "Never say anything false!" and just-not-discussing all the exceptions that people think obviously don't count.

(But of course that system also has disadvantages, like people having different automatic norms about what they think are obvious exceptions.)

I've started to worry more, recently, about which cognitive skills have other cognitive skills as prerequisites. One of the reasons I hesitated to publish *Inadequate Equilibria* (before certain persons *yanked it out of my drafts folder and published it anyway*) was that I worried that maybe the book's ideas were useless or harmful without mastery of

other skills. Like, maybe you need to have developed a skill for demotivating cognition, and until then you can't reason about charged political issues or your startup idea well enough for complicated thoughts about Nash equilibria to do more good than harm. Or maybe unless you already know a bunch of microeconomics, you just stare at society and see a diffuse mass of phenomena that might or might not be bad equilibria, and you can't even guess non-wildly in a way that lets you get started on learning.

Maybe meta-honesty contains enough meta, in that it has meta at all, that it just blows up in most people's heads. Sure, people in our little subcommunity tend to max out the Cognitive Reflection Test and everything that correlates with it. But compared to scoring 3 out of 3 on the CRT, the concept of meta-honesty is probably harder to live in real life—stopping and asking yourself "Would I be willing to publicly defend this as a situation in which unusually honest people should lie, if somebody posed it as a hypothetical?" Maybe that just gets turned into "It's permissible to lie so long as you'd be honest about whether you'd tell that lie if anyone asks you that exact question and remembers to say they're invoking the meta-honesty code," because people can't process the meta-part correctly. Or maybe there's some subtle nonobvious skill that a few people have practiced extensively and can do very easily, and that most people haven't practiced extensively and can't do that easily, and this subskill is required to think about meta-honesty without blowing up. Or maybe I just get an email saying "I tried to be meta-honest and it didn't work because my verbal SAT score was not high enough, you need to retract this."

If so, I'm not sure there's much that could be done about it, besides me declaring that Meta-Honesty had turned out to be a terrible idea as a social innovation and nobody should try that anymore. And then that might not undo the damage to the law-as-absolute performance that makes something be part of the Law.

But I'd outright lie to the Gestapo about Jews in my attic. And even to friends, I can't consistently Glomarize about every point in my life where one of my counterfactual selves could possibly have been doing that. So I can't actually promise to be a wizard, and I want there to exist firm foundations somewhere.

Questions? Comments?

Understanding is translation

Does this feel familiar: "I thought I understood thing X, but then I learned something new and realized that I'd never really understood X?"

For example, consider a loop in some programming language:

```
var i = 0;
while (i < n) {
  i = i + 1;
}
```

If you're a programmer, you probably understand it just fine. How it works, in what order the lines are executed, how the variable changes over time... But have you ever noticed that the simplest way to compile such a loop to machine code involves two jump instructions - one conditional and one unconditional? (Try doing that with only one jump, it won't work.)

Now you might feel that your "understanding" of loops has become slightly closer to "proper understanding".

Or not!

An alternative view is that understanding is translation. It's a two-place word. Your understanding of loops, in the sense of translating them to execution histories, was perfectly fine. But your understanding of loops, in the sense of translating them to machine code, was slightly lacking.

When you see that pattern once, you notice it everywhere. A middle-schooler can understand numbers, in the sense of translating them to amounts of apples and such, but doesn't immediately translate the expression "x > 5" to a half-open ray on the number line. A self-taught singer can translate from heard notes to sung notes, but can't translate either to notes on a staff; a self-taught guitarist is missing a different subset of those skills. A bilingual person can translate a Japanese sentence with the word "integral" to English, without knowing what integral means. You can be good at translating other people's facial expressions to emotional states, but lousy at translating them to pencil sketches; your friend is the opposite; which of you "understands" human faces better? There's no answer, or many answers. Don't ask whether someone understands X. Instead, ask if they can translate X <-> Y.

That has implications for teaching. If you walk into a classroom intending to make students "understand" X, you'll fail at teaching. (I'm speaking from experience here.) But if you find some Y, already understood by the students, that can be translated to X - and drill them repeatedly on both directions of translation - then they will begin to "understand" X.

Decoupling vs Contextualising Norms

One of the most common difficulties faced in discussions is when the parties involved have different beliefs as to what the scope of the discussion should be. In particular, John Nerst identifies two styles of conversation as follows:

- Decoupling norms: It is considered eminently reasonable to require the truth of your claims to be considered in isolation - free of any potential implications. An insistence on raising these issues despite a decoupling request are often seen as sloppy thinking or attempts to deflect.
- Contextualising norms: It is considered eminently reasonable to expect certain contextual factors or implications to be addressed. Not addressing these factors is often seen as sloppy or an intentional evasion.

(ht prontab. He actually uses low decoupling/high decoupling, but I prefer this terminology. Both John Nerst and prontab passed up the opportunity to post on this topic here)

Let's suppose that blue-eyed people commit murders at twice the rate of the rest of the population. With decoupling norms, it would be considered churlish to object to such direct statements of facts. Sure it's unfortunate for anyone who is blue-eyed, but the truth is the truth. With contextualising norms, you could potentially be criticised for reinforcing the stigma around blue-eyed people. At the very least, you would be expected to have issued a disclaimer to make it clear that you don't think blue-eyed people should be stereotyped as criminals.

John Nerst writes (slightly edited): "To a contextualiser, decouplers' ability to fence off any threatening implications looks like a lack of empathy for those threatened, while to a decoupler the contextualiser's insistence that this isn't possible looks like naked bias and an inability to think straight"

For both these norms, it's quite easy to think of circumstances when expectations for the other party to use these norms would normally be considered unreasonable. Weak men are superweapons demonstrates how true statements can be used to destroy a group's credibility and so it seems entirely reasonable to demand contextualisation if you suspect this is the other person's strategy. On the other hand, it's very easy to start painting every action you dislike to be part of someone's agenda (neo-liberal agenda, cultural marxist agenda, far right agenda, ect. take your pick). People definitely have agendas and take actions as a result of this, but the wide usage of universal counter-arguments should rightly be frowned upon.

I agree with the contextualisers that making certain statements, even if true, can be incredibly naive in highly charged situations that could be set off by a mere spark. On the other hand, it seems that we need at least some spaces for engaging in decoupling-style conversations. Elizier wrote an article on Local Validity as a Key to Sanity and Civilisation. I believe that having access to such spaces is another key.

At the same time, I don't want to fall for the Fallacy of the Undistributed Middle and assume that both perspectives are equally valid. While there is truth in both, I feel that at the current time society needs to shift more towards Decoupling Norms. Zack Davis writes that he is afraid that "the concept of "contextualizing norms" has the potential to legitimize derailing discussions for arbitrary political reasons by eliding the key question of which contextual concerns are genuinely relevant, thereby

conflating legitimate and illegitimate bids for contextualization". I agree that is possible in theory, but it seems that people who want to derail discussions can already do that without a need for further justification. Further, contextualising norms are widespread enough at this point that we can't really avoid dialog across these boundaries, which is what this concept enables.

Zack also wants to emphasise how contextual these factors are. That talking about the higher rate of blue-eyed people who are murderers is relevant when discussing the the higher number of blue-eyed people in prison, but irrelevant when someone mentions they are going to date someone with blue-eyes because the base rate is too low. This is a good point, but it's always the case that we can shift from viewing a phenomenon as a binary at the lowest resolution, then a spectrum, then contextual. Zack worries that a spectrum wouldn't be a useful model as there isn't a general factor of contextualising. I disagree with this - it seems that social scientists lean very heavily toward contextualisation norms and mathematicians towards decoupling norms.

These complexities mean that there isn't a simple prescriptive solution here. Instead this post merely aimed to describe this phenomenon, as at least if you are aware of this, it may be possible to navigate this.

Further reading:

- A Deep Dive into the Harris-Klein Controversy - John Nerst's Original Post
- Putanumonit - Ties decoupling to mistake/conflict theory
- Relevance norms

The Alignment Newsletter #7: 05/21/18

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

Highlights

Challenges to Christiano's capability amplification proposal *(Eliezer Yudkowsky)*: A list of challenges faced by iterated distillation and amplification. First, a collection of aligned agents interacting does not necessarily lead to aligned behavior. (Paul's response: That's not the reason for optimism, it's more that there is no optimization pressure to be unaligned.) Second, it's unclear that even with high bandwidth oversight, that a collection of agents could reach arbitrary levels of capability. For example, how could agents with an understanding of arithmetic invent Hessian-free optimization? (Paul's response: This is an empirical disagreement, hopefully it can be resolved with experiments.) Third, while it is true that exact imitation of a human would avoid the issues of RL, it is harder to create exact imitation than to create superintelligence, and as soon as you have any imperfection in your imitation of a human, you very quickly get back the problems of RL. (Paul's response: He's not aiming for exact imitation, he wants to deal with this problem by having a strong overseer aka informed oversight, and by having techniques that optimize worst-case performance.) Fourth, since Paul wants to use big unaligned neural nets to imitate humans, we have to worry about the possibility of adversarial behavior. He has suggested using large ensembles of agents and detecting and pruning the ones that are adversarial. However, this would require millions of samples per unaligned agent, which is prohibitively expensive. (Paul's response: He's no longer optimistic about ensembles and instead prefers the techniques in this post, but he could see ways of reducing the sample complexity further.)

My opinion: Of all of these, I'm most worried about the second and third problems. I definitely have a weak intuition that there are many important tasks that we care about that can't easily be decomposed, but I'm optimistic that we can find out with experiments. For the point about having to train a by-default unaligned neural net to imitate aligned agents, I'm somewhat optimistic about informed oversight with strong interpretability techniques, but I become a lot less optimistic if we think that won't be enough and need to use other techniques like verification, which seem unlikely to scale that far. In any case, I'd recommend reading this post for a good explanation of common critiques of IDA.

AI and Compute *(Dario Amodei et al)*: Since 2012, when the deep learning revolution began with AlexNet, the amount of compute used in the largest-scale experiments has been doubling *every 3.5 months*. Initially, people started to use GPUs to scale up, but there wasn't a huge amount o f interest. In 2014-16, as interest in deep learning really began to take off, people started to use a lot of compute to get good results -- but parallelism stopped helping beyond a certain point (~100 GPUs) because the parameter updates from the data were becoming too stale. Since then, we've had algorithmic improvements that allow us to take advantage of more parallelism (huge batch sizes, architecture search, expert iteration), and this has let us scale up the amount of compute thrown at the problem.

My opinion: I did know that the amount of compute used was growing fast, but a 3.5 month doubling time *for 6 years running* is *huge*, and there's no reason to expect that it will stop now. It's also interesting to see what made it onto the graph -- there's image classification, machine translation, and neural architecture search (all of which have clear economic incentives), but some of the largest ones are by projects aiming to build AGI (AlphaGo Zero, AlphaZero, and Dota). Notably, deep reinforcement learning just barely makes it on the graph, with DQN two orders of magnitude lower than any other point on the graph. I'm really curious what deep RL could solve given AlphaGo levels of compute.

80K podcast with Allan Dafoe *(Allan Dafoe and Rob Wiblin)*: A long interview with Allan Dafoe about the field of AI policy, strategy and governance. It discusses challenges for AI policy that haven't arisen before (primarily because AI is a dual use technology), the rhetoric around arms races, and autonomous weapons as a means to enable authoritarian regimes, to give a small sampling. One particularly interesting tidbit (to me) was that Putin has said that Russia will give away its AI capabilities to the world, because an arms race would be dangerous.

My opinion: Overall this is a great introduction to the field, I'd probably recommend people interested in the area to read this before any of the more typical published papers. I do have one disagreement -- Allan claims that even if we stopped Moore's law, and stopped algorithmic scientific improvement in AI, there could be some extreme systematic risks that emerge from AI -- mass labor displacement, creating monopolies, mass surveillance and control (through robot repression), and strategic stability. I would be very surprised if current AI systems would be able to lead to mass labor displacement and/or control through robot repression. We are barely able to get machines to do anything in the real world right now -- *something* has to improve quite drastically, and if it's neither compute nor algorithms, then I don't know what it would be. The other worries seem plausible from the technical viewpoint.

Technical AI alignment

Iterated distillation and amplification

Challenges to Christiano's capability amplification proposal *(Eliezer Yudkowsky)*: Summarized in the highlights!

Forecasting

Why Is the Human Brain So Efficient? *(Liqun Luo)*: Overall point for this audience is that, despite how slow and imprecise neuron signals are, the human brain beats computers because of how massively parallel it is.

Field building

Critch on career advice for junior AI-x-risk-concerned researchers *(Andrew Critch, via Rob Bensinger)*: A common piece of advice for aspiring AI x-risk researchers is to work on AI capabilities research in order to skill up so they can later contribute to safety. However, Critch is worried that such researchers will rationalize their work as being "relevant to safety", leading to a false sense of security since AI researchers are now surrounded by people who are "concerned about safety", but *aren't actually doing*

safety research. Note that Critch would still advise young researchers to get into grad school for AI, but to be aware of this effect and not feel any pressure to do safety research and to avoid rationalizing whatever research they are doing.

My opinion: I feel pretty unqualified to have an opinion here on how strong this effect is -- it's pretty far outside of my experience. At the very least it's a consideration we should be aware about, and Critch supports it better in the full post, so I'd recommend you read it.

Near-term concerns

Fairness and bias

Delayed Impact of Fair Machine Learning *(Lydia T. Liu et al)*: Consider a bank that has to choose which loan applications should be approved based on a credit score. Typically, fairness in this setting is encoded by saying that there should be some sort of parity between groups (and different criteria have been proposed for what actually should be the same). However, if you model the actual outcomes that come from the decision (namely, profit/loss to the bank *and* changes in credit score to the applicant), you can see that standard fairness criteria lead to suboptimal outcomes. As a result, in general you want to look at the delayed impact of ML models.

My opinion: This actually feels quite related to the value alignment problem -- in general, we care about things besides fairness, and if we try to optimize directly for fairness, then we'll be giving up good outcomes on other dimensions. It's another case of Goodhart's law, where "fairness" was a proxy for "good for disadvantaged groups".

Machine ethics

Tech firms move to put ethical guard rails around AI *(Tom Simonite)*: A description of the ethics boards that tech companies are putting up.

AI strategy and policy

80K podcast with Allan Dafoe *(Allan Dafoe and Rob Wiblin)*: Summarized in the highlights!

Policy Researcher *(OpenAI)*: There is a job opportunity at OpenAI as a policy researcher, which does not seem to have any formal requirements.

My opinion: It seems like a lot of the best policy work is happening at OpenAI (see for example the OpenAI charter), I strongly encourage people to apply!

AI capabilities

Reinforcement learning

Reward Estimation for Variance Reduction in Deep Reinforcement Learning *(Joshua Romoff et al)*

Critiques (Capabilities)

To Build Truly Intelligent Machines, Teach Them Cause and Effect *(Kevin Hartnett interviewing Judea Pearl)*: An interview with Judea Pearl about causality, deep learning, and where the field is going.

My opinion: This is fairly superficial, if you've read any of the other things that Pearl himself has written about deep learning, you'll know all of this already.

Miscellaneous (Capabilities)

AI and Compute *(Dario Amodei et al)*: Summarized in the highlights!

How artificial intelligence is changing science *(Nathan Collins)*: AI is being used in many different projects across many different fields at Stanford. This post has a list of a whole bunch of scientific projects that AI is helping with.

Why Universal Comparability of Utility?

Apologies if this is answered elsewhere and I couldn't find it. In AI reading I come across an agent's utility function, U, mapping world-states to real numbers.

The existence of U is justified by the VNM-utility theorem. The first axiom required for VNM utility is 'Completeness' -- in the context of AI this means for every pair of world-states, W_i and W_j, the agent knows $W_i > W_j, W_i < W_j$, or $W_i \sim W_j$.

Completeness over world-states seems like a huge assumption. Every agent we make this assumption for must already have the tools to compare 'world where, all else equal, the only food is peach ice cream' v. 'world where, all else equal, Shakespeare never existed.'* I have no idea how I'd reliably make that comparison as a human, and that's a far cry from '~', being indifferent between the options.

Am I missing something that makes the completeness assumption reasonable? Is 'world-state' used loosely, referring to a point in a vastly smaller space, with the exact space never being specified? Essentially, I'm confused, can anyone help me out?

*if it's important I can try to cook up better-defined difficult comparisons. 'all else equal' is totally under-specified... where does the ice cream come from?

The Sheepskin Effect

Previously: The Case Against Education, The Case Against Education: Foundations, The Case Against Education: Splitting the Education Premium Pie and Considering IQ

Epistemic Status: The spirit of Local Validity as a Key to Sanity and Civilization

The sheepskin effect is that completing the last year of high school, college or graduate school is *much* more profitable than completing any of the previous years, rivaling those other years combined. Employers seem to be paying for the degree (aka the sheepskin, which it's printed on) rather than the human capital being built over time.

In the education chapter of Book Review: The Elephant in the Brain, I noted Robin relied on the sheepskin effect as strong evidence (along with other arguments, including impacts on national vs. personal income) that school was mostly signaling. Bryan Caplan does the same. He cites the data, seeing (on top of a 10% bonus in pay per year of school) 32% bonus pay for finishing high school, 10% for junior college, 30% for a bachelor's degree and 18% for a masters. To those who claim this is mostly ability bias, he replies:

> Ability bias explanations for sheepskin effects aren't just hard to square with statistical evidence; they're hard to square with the glaring fact that education spikes in degree years. If the labor market ignores credentials, why do so many college grads opt for *zero* graduate education? Are we supposed to believe one-third of the population has exactly the right ability to finish high school, but not advance to college? One-seventh has exactly the right ability to finish college, but not advance to graduate school?
>
> …
>
> To debunk sheepskin effects, correcting for these neglected abilities would have to drastically cut the payoff for degrees but *not* the payoff for years of schooling. What abilities would even conceivably qualify?

This seems like a straw man; no one thinks the labor market ignores credentials, so it's easy to see why students act the way they do. Not only is not finishing high school severely punished as such (as the numbers show), not trying is at least sort of illegal. In addition, there's a huge barrier to *getting into* college or graduate school, and large costs involved in starting, often involving relocation. Also, much of college is about being ready for the rest of college, and the early part of graduate school is largely to get you ready for the later parts.

I also notice, looking again, another instance of the mistake of *assuming people are maximizing.* We are definitely *not* supposed to believe that because a lot of people do something, it was right for them!

Having dismissed ability bias here, he then reasons:

> After digesting all the evidence on the sheepskin effect, you may feel ready to channel King Solomon. Human capital and signaling come before you as litigants. They ask you to split the education premium between them. A ruling with a great ring to it: "Human capital gets credit for the payoff for years of education;

> signaling gets credit for the payoff for degrees." This implies a human capital/signaling split of roughly 60/40 for high school, 40/60 for college.
>
> Yet on reflection the Solomonic ruling treats human capital too generously. The sheepskin effect doesn't measure signaling. Instead, the sheepskin effect sets a *lower bound* on signaling.
>
> …
>
> To see why, picture a world that lacks the notion of "graduation." Can we safely declare educational signaling would vanish in such a world? Of course not.

What I wrote back in my previous review of Elephant in the Brain:

> One note I would make is about the sheepskin effect, where the last year of a college degree is much more valuable than previous years. There's been some debate about this online lately between Bryan Caplan and Noah Smith. I agree that this is largely a signaling effect, with 'completed all eight terms' much more impressive than 'completed seven of eight terms' since you don't know how many more terms the first student *could* have finished if necessary.
>
> What the discussion misses, it seems to me, is that *only after graduation do you know that the first three years were real.* It is easy to become 'a senior' through completion of a number of credits, saving the stuff they find hardest for last or even being in terrible shape to match up with graduation requirements. I strongly suspect that a lot of people who drop out in year four are much farther from finished than they would have you believe.

I'd like to expand upon that, because this effect seems huge but remains almost always unmentioned.

I'll start with a real example.

I have a learning disability that makes it very difficult to learn foreign languages. This was bad enough to nuke my average in high school, despite having studied the same language (Hebrew) for most of a decade whether I wanted to or not (I've held on to *maybe* a hundred words?), and in college things threatened to get much worse. My college demands four terms of a single foreign language. I chose what appeared to be the easiest one for an English speaker, Italian. While it would be cool to speak Italian, I didn't choose it for how much cooler it would make vacations and restaurants – I was fully aware that Italian was of little use. But I was *desperate* to get through this, ideally without my average being nuked again, and if it was marginally easier than Spanish but only 10% as useful, then Italian it would be.

When the term ended, I had spent the majority of my studying time on Italian I, and still (just barely) failed by the numbers. I managed to get the grade changed from an F to a D by promising not to take Italian II. I then managed to find a psychologist who vouched for my disability, I think on the basis of an IQ test combined with my history of failures – there really was no other explanation. So I was granted an exception, and allowed to take Asian literature (which I quite enjoyed) and Etymology (which was boring as hell but not hard) instead of the remaining three terms. The D still ended any hopes of my getting honors and crippled any hopes of a top graduate school, and after that I stopped trying that hard to get As, but I got to graduate.

Without that exception, would I have graduated? My guess is yes, because my family and I would have taken epic measures to make it work. I'd have taken a year off to live in Italy (or Israel) if I'd had to. But I can't be sure it would have been enough.

A good friend of mine ran into this exact problem with the same requirement, couldn't get the waiver, has no other remaining requirements, and will probably never graduate.

More data. My mother is a professor at Columbia University, where she is in charge of undergraduate biology education. One cool effect of this is that she'd bring related dilemmas and puzzles home so we could explore them at dinner. I helped her plan exam strategies, deal with discipline issues and so on, and it was both great fun *and an actual education in the way that school isn't.*

Occasionally we looked at a series of students who wanted to graduate with a major in biology. The problem was that their transcripts were, shall we say, not so flattering. They'd 'completed' close to the full eight terms, but did they have a high enough average in their major? Were the poor grades (Ds and sometimes Cs) in some required courses not acceptable? We all, including the school, *wanted* to let students graduate when we could – that's the business, after all, and no one wants to ruin a kid's life – but the degree has to mean something. These were many of the students who 'complete seven of eight terms,' and many others were those who knew a version of this examination was coming and they wouldn't pass.

Countless others, no doubt, simply saved all the hardest and most difficult courses for the end, possibly in a way that made the logistics impossible to solve. And, well, whoops.

If I give you eight chess puzzles to solve, and you solve seven of them, that's a *lot* less impressive than if you solve all eight. If I give you thirty-two courses in ten different fields of study with varying difficulty, and you choose your order so as to solve and pass the first twenty-eight, you are not *remotely* 7/8ths done.

I could thus tell a human capital story, or an ability bias story, for the sheepskin effect. The final test is real, so if you built up real human capital, and learned how to learn things and remembered your lessons and persevere when the going gets tough, and all that, you win out. If you didn't do that stuff, you fail at the end when you can't hide it any longer. Or, for ability bias, only at the end do we learn who had the right stuff all along; same principle. If the final test is sufficiently 'more real' than the others, that bonus at the end makes perfect sense.

Thus, I don't think the arguments from sheepskin are as strong as many think they are. I do think that the education premium is mostly signaling and ability bias, including (but far from limited to) the sheepskin effect. And I do think Bryan offers other much stronger evidence, such as the fact that anyone could walk into any college class and take it for free sans the degree, and actual no one ever does. But I don't think the sheepskin effect puts a lower bound on the signaling share, or offers that much evidence, because in a world without signaling you'd see it anyway, and I'm curious how Bryan would respond.

Of Two Minds

Follow-up to: The Intelligent Social Web

The human mind evolved under pressure to solve two kinds of problems:

- How to physically move
- What to do about other people

I don't mean that list to be exhaustive. It doesn't include maintaining homeostasis, for instance. But in practice I think it hits everything we might want to call "thinking".

…which means we can think of the mind as having two types of reasoning: mechanical and social.

Mechanical reasoning is where our intuitions about "truth" ground out. You throw a ball in the air, your brain makes a prediction about how it'll move and how to catch it, and either you catch it as expected or you don't. We can imagine how to build an engine, and then build it, and then we can find out whether it works. You can try a handstand, notice how it fails, and try again… and after a while you'll probably figure it out. It *means* something for our brains' predictions to be right or wrong (or somewhere in between).

I recommend this TED Talk for a great overview of this point.

The fact that we can *plan* movements lets us do abstract truth-based reasoning. The book Where Mathematics Comes From digs into this in math. But for just one example, notice how set theory almost always uses container metaphors. E.g., we say elements are *in* sets like pebbles are in buckets. That physical intuition lets us use things like Venn diagrams to reason about sets and logic.

…well, at least until our intuitions are wrong. Then we get surprised. And then, like in learning to catch a ball, we change our anticipations. We update.

Mechanical reasoning seems to *already obey* Bayes' Theorem for updating. This seems plausible from my read of Scott's review of Surfing Uncertainty, and in the TED Talk I mentioned earlier Daniel Wolpert claims this is measured. And it makes sense: evolution would have put a lot of pressure on our ancestors to get movement right.

Why, then, is there systematic bias? Why do the Sequences help at all with thinking?

Sometimes, occasionally, it's because of something structural — like how we systematically feel someone's blow as harder than they felt they had hit us. It just falls out of how our brains make physical predictions. If we know about this, we can try to correct for it when it matters.

But the rest of the time?

It's because we predict it's *socially helpful to be biased that way.*

When it comes to surviving and finding mates, having a place in the social web matters a *lot* more than being right, nearly always. If your access to food, sex, and others' protection depends on your agreeing with others that the sky is green, you

either find ways to conclude that the sky is green, or you don't have many kids. If the social web puts a lot of effort into figuring out what you *really* think, then you'd better find some way to *really think* the sky is green, regardless of what your eyes tell you.

Is it any wonder that so many deviations from clear thinking are about social signaling?

The thing is, "clear thinking" here mostly points at mechanical reasoning. If we were to create a mechanical model of social dynamics… well, it might start looking like a recursively generated social web, and then mechanical reasoning would mostly derive the same thing the social mind already does.

…because *that's how the social mind evolved*.

And once it evolved, it became *overwhelmingly* more important than everything else. Because a strong, healthy, physically coordinated, skilled warrior has almost no hope of defeating a weakling who can inspire many, many others to fight for them.

Thus whenever people's social and mechanical minds disagree, the social mind almost always wins, even if it kills them.

You might hope that that "almost" includes things like engineering and hard science. But really, for the most part, we just figured out how to *align* social incentives with truth-seeking. And that's important! We figured out that if we tie social standing to whether your rocket actually *works*, then being right *socially matters*, and now culture can care about truth.

But once there's the *slightest* gap between cultural incentives and making physical things work, social forces take over.

This means that in any human interaction, if you don't see how the social web causes each person's actions, then you're probably missing most of what's going on — at least consciously.

And there's probably a reason you're missing it.

[CKC] [May 2018] What subjects are most important for an AI safety researcher to know? (Open Call)

[CKC]: I intend this to be a Community Knowledge Convergence project - that is to say, I do not think I am particularly well informed on the subject, and I think there's a lot to be gained from compiling an understanding of what others in the LessWrong community consider as very important, probably important, possibly important, or maybe important.

What academic subjects do you consider most important to the study of AI safety? What non-academic subjects would you recommend?

Currently, here are the subjects I think are most important.

- (Non-academic) Eliezer's Sequences.
- Mathematical logic.
- Game theoretic economics.

You might be able to tell, I'm not very well versed in this discipline. Please comment and add more subjects for me to add to the [June 2018] posts.

Moral frameworks and the Harris/Klein debate

Here's a good background post and analysis on the debate (this has been linked from elsewhere on LW before): https://everythingstudies.com/2018/04/26/a-deep-dive-into-the-harris-klein-controversy/

Like many, I couldn't help but be fascinated by the Sam Harris/Ezra Klein debate. These are two people I really look up to, and so seeing them going at it (and showing a lot of personal weakness along the way) has been illuminating. I'm still unsettled about it, wanting there to be resolution/a right answer. So far that satisfaction has eluded me, so I wrote this to try to clarify things for myself. Maybe it helps others too.

The analysis below is meant as a steelman of each side's positions. If you think I'm not steelmanning them well enough, please leave a comment and I'll improve.

Consequentialist framework:

- Sam: As a lesson in how to think for yourself, hold Murray up as someone who has discovered truths that society doesn't like to talk about. As a general policy, this practice will lead to truths being uncovered faster, leading to a faster pace of discovery, which compounds over time to a much better world through science.
- Ezra: Make a public example of Sam here, leading to more people recognizing their own privilege and putting their actions in the appropriate historical context. As a general policy, this practice will lead to a more equitable society, which compounds over time to a much better world by reducing suffering directly.

Virtue ethics framework:

- Sam: It is virtuous to signal-boost things which are true, especially when they are being suppressed by society. "Speak the truth though your voice may tremble"
- Ezra: It is virtuous to defend the underprivileged by calling out harms, even unintentional harms. "Evil is the silence of the voice of justice when it matters most"

Deontology framework:

- Sam: Thou shalt update on all available data.
- Ezra: Thou shalt not invoke long-buried demons of oppression.

Both these moral frameworks look pretty good to me. I see no particular reason to favor one over the other; even if I restrict myself to only looking at the consequences I see plausible arguments that one path or the other is higher-impact.

I have only one unifying thought:

Sam was upset by being attacked by Klein, and considered the attack unfair, which actually triggered the whole debate. I think Sam's complaint about "unfairness" is invalid here, because it is exactly what he should expect when signal-boosting things

that society doesn't like to talk about. So perhaps the only error here was Sam getting too emotional about being pilloried.

Talents

> For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken away even that which he hath.
>
> - The Gospel according to Matthew
>
> r > g
>
> -Thomas Piketty, Capital in the Twenty-First Century

From Jesus to Piketty, it is a commonplace that wealth is a positive feedback loop.

Under one model, differential ability to steward capital, plus compounding gains, implies that perfectly benevolent people with more money than most should keep it more often than a naive expected utility maximization would suggest. On the other hand, conquering empires also experience compounding gains; the ability to leverage force into more force implies that this is a harmful positive feedback loop.

When seeking to do good, one should often attend to the specific details of the situation and take whichever action has the outcome they most prefer. But often there isn't a very strong case for one particular action, and we're left in need of a general heuristic for how to allocate our resources. Using explicit expected-value calculations in those circumstances will systematically underallocate resources to good uses that are less legible, and overallocate to things that illegibly cause harm. Accordingly, we need some baseline presumption for how to allocate one's surplus between oneself, institutions one is participating in, and the rest of the world.

The deserving rich hypothesis

> Money is a tool of exchange, which can't exist unless there are goods produced and men able to produce them. Money is the material shape of the principle that men who wish to deal with one another must deal by trade and give value for value. Money is not the tool of the moochers, who claim your product by tears, or of the looters, who take it from you by force. Money is made possible only by the men who produce. Is this what you consider evil?
>
> When you accept money in payment for your effort, you do so only on the conviction that you will exchange it for the product of the effort of others. It is not the moochers or the looters who give value to money. Not an ocean of tears not all the guns in the world can transform those pieces of paper in your wallet into the bread you will need to survive tomorrow. Those pieces of paper, which should have been gold, are a token of honor–your claim upon the energy of the men who produce. Your wallet is your statement of hope that somewhere in the world around you there are men who will not default on that moral principle which is the root of money.
>
> -Ayn Rand, Atlas Shrugged

Suppose I am perfectly benevolent and value everyone's well-being equally. I grow some wheat on otherwise unused land, harvest it, grind it, and bake it into a thing of

value: bread. If other people are also hungry, I might share my bread. But even from the point of view of perfectly egalitarian benevolence, there should be a strong presumption that I ought to feed myself first. This is because my possession of the bread is evidence of my capacity to produce it; feeding me indirectly feeds others, because it enables me to produce more bread in the future. Feeding others who did not produce bread does not have that sort of flow-through effect.

Likewise, I may save some seed corn to reinvest by ploughing it back into my field, even if someone else is hungry. Even if I value others' well-being equally to my own, I have demonstrated an ability to use grain to make more grain, which will feed people better in the long run.

This principle can be generalized. Suppose my village's economy is complex enough to have a need for a currency to serve as an unit of account and store of value. If I sell my bread for money, then my possession of money serves as evidence that I have some sort of productive capacity. In a world where this is the main way one acquires money, there should be a similar generalized presumption that even a perfectly benevolent person should hold onto a disproportionate share of the money they earn.

The war profiteer hypothesis

> Everybody knows that the dice are loaded
>
> Everybody rolls with their fingers crossed
>
> Everybody knows the war is over
>
> Everybody knows the good guys lost
>
> Everybody knows the fight was fixed
>
> The poor stay poor, the rich get rich
>
> That's how it goes
>
> Everybody knows
>
> -Leonard Cohen

Suppose instead that my village does not possess the shieldmaking craft, and is being harassed by a gang of archers, who *can* make shields, albeit ones whose usefulness decays over time. These archers periodically rain arrows down on the village, at which point people without shields have a substantial chance of dying of arrow wounds. However, they make me an offer: if I make them some arrows for them, they will give me a shield in exchange.

In this situation, while it is understandable for me to earn some shields in order to protect myself, possession of shields is prima facie evidence of complicity in the violence being done to my village. Accumulating a surplus is particularly bad behavior. I would not want to *generalize* the heuristic that arrowmakers should get to keep their shields; that would mean more arrows, resulting in increased need for shields, in a harmful positive feedback loop. The presumption that one should keep a disproportionate share of the shields one has made no longer appears to hold.

A secondary consequence of this is that shields, which may have had no value in the village before, are now highly sought after. The archers might *also* exchange shields for bread, since the breadmaker has as much incentive as the arrowmaker to accept shields as payment. I, the arrowmaker, might trade one of my shields to the breadmaker for bread. If the arrowmakers make a special effort to attack people producing “counterfeit” shields, and people transacting in other currencies, then shields might quickly become a standard unit of account and store of value. Under those conditions, possession of a large stockpile of shields could be evidence of complicity in violence, but it is *also* evidence that someone has produced genuinely valuable goods and services that one might want to see more of.

Possessing shields could mean that you have made much bread. Or that you have made many arrows, with which those lacking shields will be harmed. Or, in many cases, some combination of the two. Production and violence are bound together into a single unit of account.

For now, the application to the present situation is left as an exercise for the reader.

-

Thanks to Jessica Taylor for the shields metaphor, and Wei Dai for asking the right questions to force me to clarify my thoughts on this.

Related: Matthew 25, Cash transfers are not necessarily wealth transfers, Why I am not a Quaker (even though it often seems as though I should be), The humility argument for honesty

The Alignment Newsletter #8: 05/28/18

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

Highlights

Solving the Rubik's Cube Without Human Knowledge *(Stephen McAleer, Forest Agostinelli, Alexander Shmakov et al)*: This paper proposes *Autodidactic Iteration* (ADI), which is a technique that can be combined with the techniques in AlphaGo and expert iteration to solve problems with only one goal state, such as the Rubik's cube. MCTS with value and policy networks will not suffice, because when starting from a randomly scrambled cube, MCTS will never find a path to the goal state, and so there will never be any reward signal. (Whereas with Go, even if you play randomly the game will end relatively quickly, giving you some reward signal.) To get around this, they start *from the goal state* and generate states that are near the goal state. This gives them a training dataset of states for which they know (a good approximation to) the value and the best action, which they can use to train a value and policy network. They then use this with MCTS to solve the full problem, as in AlphaGo.

My opinion: This general idea has been proposed in robotics as well, in Reverse Curriculum Generation for Reinforcement Learning, where there is a single goal state. However, in this setting we have the added benefit of perfect inverse dynamics, that is, for any action a that moves us from state s to s', we can find the inverse action a' that moves us from state s' to s. This allows the authors to start from the goal state, generate nearby states, and automatically know the value of those states (or at least a very good approximation to it). Hindsight Experience Replay also tackles similar issues -- I'd be interested to see if it could solve the Rubik's cube. Overall, the problem of sparse rewards is very difficult, and it seems like we now have another solution in the case where we have a single goal state and perfect (or perhaps just sufficiently good?) inverse dynamics.

Where Do You Think You're Going?: Inferring Beliefs about Dynamics from Behavior *(Siddharth Reddy et al)*: Inverse reinforcement learning algorithms typically assume that the demonstrations come from an expert who is approximately optimal. However, this is often not the case, at least when the experts are fallible humans. This paper considers the case where the expert has an incorrect model of the dynamics (transition function) of the environment, and proposes learning the expert's model of the dynamics to improve reward function inference. However, this leads to severe unidentifiability problems, where many models of the dynamics are compatible with the observed behavior. To overcome this, they assume that they have multiple tasks with known reward functions, which they use to infer the expert's dynamics. This is then used to infer the reward function in a new task using an adaptation of max causal entropy IRL. The dynamics can be an arbitrary neural net while the reward function is a weighted linear combination of features. They evaluate the inference of the dynamics model with real humans on Lunar Lander. Given transcripts of humans playing Lunar Lander, they infer the underlying (incorrect) dynamics model. Then, when the human takes an action, they predict which next state the human wanted to

achieve, and replace the human's action with the action that would actually get close to the state the human wanted.

My opinion: I really like that this paper has experiments with real humans. It's definitely a problem that IRL assumes that the expert is (approximately) optimal -- this means that you can't learn where the expert is likely to be wrong, and so it is hard to exceed the expert's performance. It's very difficult to figure out how to deal with the possbility of a biased expert, and I'm happy to see work that takes a shot at it.

Technical AI alignment

Problems

How the Enlightenment Ends *(Henry A. Kissinger)*: This is an article about the dangers of AI written by a non-technologist, hitting some points that are relatively familiar.

My opinion: While there are many points that I disagree with (eg. "what [AIs] do uniquely is not thinking as heretofore conceived and experienced. Rather, it is unprecedented memorization and computation"), overall there was a surprising amount of familiar material said in a different way (such as explainability and unintended consequences).

Learning human intent

Where Do You Think You're Going?: Inferring Beliefs about Dynamics from Behavior *(Siddharth Reddy et al)*: Summarized in the highlights!

A Framework and Method for Online Inverse Reinforcement Learning *(Saurabh Arora et al)*: This paper introduces Incremental Inverse Reinforcement Learning (I2RL), where the agent continually gets new demonstrations from an expert, and has to update the estimate of the reward function in real time. The running example is a robot that has to navigate to a goal location without being seen by two guards that are patrolling. The robot needs to infer the rewards of the two guards in order to predict what they will do and plan around them. Since the guards are sometimes out of sight, we get demonstrations *with occlusion*, that is, some of the states in the demonstrations are hidden.

In the batch setting, this is solved with Latent Maximum Entropy IRL. To deal with occluded states Z, we define a probability distribution Pr(Z | Y, θ), where Y is the visible states and θ is the reward weights. Then, you can use expectation maximization to find θ -- in the expectation step, you compute feature expectations of the demonstrations (taking an expectation over hidden states Z), and in the maximization step, you compute θ using the feature expectations as in standard maximum entropy IRL. The authors show how to extend this algorithm to the incremental setting where you only keep the reward weights, the feature expectations, and the number of past demonstrations as statistics. They show some convergence guarantees and evaluate on their running example of a robot that must evade guards.

My opinion: IRL algorithms are often more computationally expensive than state-of-the-art RL algorithms, so I'm happy to see work that's trying to make it more realistic. That said, this paper focuses on settings where IRL is used to infer other agent's

preferences so we can plan around them (as opposed to imitation learning) -- this setting seems not very important for AI alignment. I'm also very confused by the experiments -- it seems in Figure 2 that if you ignore previous optimization and initialize the reward with random weights, it does better. (It isn't ignoring all previous data, because it still has access to past feature expectations.) They don't comment on this in the paper, but my guess is that they ran more iterations of expectation maximization (which is why the learning duration is higher) and that's why they got better performance.

Imitating Latent Policies from Observation *(Ashley D. Edwards et al)*

Machine Teaching for Inverse Reinforcement Learning: Algorithms and Applications *(Daniel S. Brown et al)*

Maximum Causal Tsallis Entropy Imitation Learning *(Kyungjae Lee et al)*

Planning to Give Information in Partially Observed Domains with a Learned Weighted Entropy Model *(Rohan Chitnis et al)*

Safe Policy Learning from Observations *(Elad Sarafian et al)*

Handling groups of agents

Learning to Teach in Cooperative Multiagent Reinforcement Learning *(Shayegan Omidshafiei et al)*

Interpretability

Unsupervised Learning of Neural Networks to Explain Neural Networks *(Quanshi Zhang et al)*

Verification

Verifiable Reinforcement Learning via Policy Extraction *(Osbert Bastani et al)*: Since it is hard to verify properties of neural nets, we can instead first train a decision tree policy to mimic the policy learned by deep RL, and then verify properties about that. The authors generalize DAGGER to take advantage of the Q-function and extract decision tree policies. They then prove a correctness guarantee for a toy version of Pong (where the dynamics are known), a robustness guarantee for Pong (with symbolic states, not pixels) (which can be done without known dynamics), and stability of cartpole.

My opinion: Many people believe that ultimately we will need to prove theorems about the safety of our AIs. I don't understand yet what kind of theorems they have in mind, so I don't really want to speculate on how this relates to it. It does seem like the robustness guarantee is the most relevant one, since in general we won't have access to a perfect model of the dynamics.

Miscellaneous (Alignment)

When is unaligned AI morally valuable? *(Paul Christiano)*: When might it be a good idea to hand the keys to the universe to an unaligned AI? This post looks more deeply

at this question, which could be important as a backup plan if we don't think we can build an aligned AI. I can't easily summarize this, so you'll have to read the post.

A Psychopathological Approach to Safety Engineering in AI and AGI *(Vahid Behzadan et al)*: Since AGI research aims for cognitive functions that are similar to humans, they will be vulnerable to similar psychological issues. Some problems can be recast in this light -- for example, wireheading can be thought of as delusional or addictive behavior. This framework suggests new solutions to AI safety issues -- for example, analogous to behavioral therapy, we can retrain a malfunctioning agent in controlled environments to remove the negative effects of earlier experiences.

My opinion: The analogy is interesting but I'm not sure what to take away from the paper, and I think there are also big disanalogies. The biggest one is that we have to communicate our goals to an AI, whereas humans come equipped with some goals from birth (though arguably most of our goals come from the environment we grow up in). I'd be interested in seeing future work from this agenda, since I don't know how I could do work on the agenda laid out in this paper.

AI strategy and policy

2018 White House Summit on Artificial Intelligence for American Industry *(White House OSTP)*: See Import AI

France, China, and the EU All Have an AI Strategy. Shouldn't the US? *(John K. Delaney)*: See Import AI

Read more: FUTURE of AI Act

AI capabilities

Reinforcement learning

Solving the Rubik's Cube Without Human Knowledge *(Stephen McAleer, Forest Agostinelli, Alexander Shmakov et al)*: Summarized in the highlights!

Gym Retro, again *(Vicki Pfau et al)*: OpenAI is releasing the full version of Gym Retro, with over a thousand games, and a tool for integrating new games into the framework. And of course we see new games in which RL agents find infinite loops that give them lots of reward -- Cheese Cat-Astrophe and Blades of Vengeance.

Feedback-Based Tree Search for Reinforcement Learning *(Daniel R. Jiang et al)*: See Import AI

Evolutionary Reinforcement Learning *(Shauharda Khadka et al)*

Learning Time-Sensitive Strategies in Space Fortress *(Akshat Agarwal et al)*

Learning Real-World Robot Policies by Dreaming *(AJ Piergiovanni et al)*

Episodic Memory Deep Q-Networks *(Zichuan Lin et al)*

Meta learning

Meta-learning with differentiable closed-form solvers *(Luca Bertinetto et al)*

Task-Agnostic Meta-Learning for Few-shot Learning *(Muhammad Abdullah Jamal et al)*

Hierarchical RL

Hierarchical Reinforcement Learning with Deep Nested Agents *(Marc Brittain et al)*

Hierarchical Reinforcement Learning with Hindsight *(Andrew Levy et al)*

Data-Efficient Hierarchical Reinforcement Learning *(Ofir Nachum et al)*

Miscellaneous (Capabilities)

The Blessings of Multiple Causes *(Yixin Wang et al)*

Mental Illness Is Not Evidence Against Abuse Allegations

ETA: this post was pretty much refuted by comments below.

I've noticed a situation several times that I think deserves attention.

Somebody goes around saying they've been the victim of mistreatment. But they seem mentally ill. Whether or not you know of a diagnosis, they seem "off" somehow -- highly agitated, making social faux pas, telling stories that don't quite add up. So people are very suspicious about whether their allegations are true.

Is this rational?

In general, someone who seems less trustworthy should be believed less. And, yes, mentally ill people are more likely to be delusional or exaggerating. But they are *also* more likely to *actually be* victims of crimes than the general population.

40% of women in the UK with severe mental illness are victims of rape or attempted rape.

People with severe mental illness are 6x as likely as the general population to have recently experienced sexual violence.

30% of mentally ill adults in an American study had been victims of violent crime in the previous six months.

Mentally ill adults in Sweden are 5x more likely than the general population to be murdered.

More than 25% of severely mentally ill Americans have been the victims of a violent crime in the last year, 4x the rate of the general population.

30-33% of psychiatric patients have been victims of domestic violence.

Someone being mentally ill is evidence *for,* not against, their being victims of a crime. And the base rates of violent crime are pretty high, so all things being equal, "someone attacked me" is not an extraordinary claim. *Even when* someone seems crazy and has made a lot of claims you don't believe, it can be reasonable to believe their claims of crime victimization. Don't fall into the horns effect.

Gaining Approval: Insights From "How To Prove It"

[Note: These are insights that surprised/connected the dots for me personally. I feel dumb admitting some of these, but my hope is that this'll be a good data point for others]

How to write a math book:

1. It's fun to prove things wrong. Examples such as "What's wrong with this proof" really worked for me.

2. A great pattern for teaching: a solid explanation of new concept, 3-5 example problems to work through, and solutions & commentary *immediately* after! It had a short feedback loop that I felt myself *wanting* to be a part of.

3. Concepts built upon each other in quick succession, so you don't have to flip back and review.

Proofs:

1. Proofs aren't made to show *how* they made the proof, just *what* they're proving in a succinct way. Since this is the case, it's okay to be confused even if the proof sounds like everything *of course* follows everything else, and you're an idiot for not immediately following. The proof took hours/weeks/years/centuries to condense in that form, so it's reasonable if it takes a few hours & math stack exchange to understand it.

2. There are tips and tricks for writing proofs depending on the properties of what you're proving. For instance, it's generally easier to prove a positive, so if you're proving a negative, make it a positive and do a proof by contradiction on it!

Math:

1. $A \longleftrightarrow B$ is making two claims: $A \longrightarrow B$ & $A \longleftarrow B$.

2. Intuition for material implication (rule of inference):

- Either Bob did it or Alice did it. If it's not Bob, then it's Alice.
- It's either A or B. If it's not A, then it's B.
- $A \vee B \longleftrightarrow \neg A \longrightarrow B$

3. Intuition for contrapositive of implication:

- $\neg A \longrightarrow B \longleftrightarrow A \vee B$ (rule of inference above)
- $A \vee B \longleftrightarrow B \vee A$ (Bob or Alice = Alice or Bob)
- $B \vee A \longleftrightarrow \neg B \longrightarrow A$ (rule of inference above)
- $\neg B \longrightarrow A \longleftrightarrow \neg A \longrightarrow B$ (Bam! Contrapositive!)

4. Combining induction with functions gets you recursion.

5. X = Ø is a negative statement: ∀y(y ∉ X)

6. X is infinite is a negative statement: *not* finite.

Math, Dancing, and Music:

I improv dance & piano, and to hit on something interesting, you have to combine different relationships with different objects. Let's hone in on just *negation.*

In math, negate "a subset of" to get "disjoint", "intersection of" to get "symmetric difference", "for all x, P(x)" to get "there exist an x where ¬P(x)" (plus the 3-4 mentioned above).

In dance, you can double your repertoire by simply applying negation to all of your dance moves. Even reversing walking or turning your head looks smooth (results may vary). The only proper name dance reversals I can think of is the running man & the jerk.

In music, reversing your melody, chord progression order, or meter can also produce interesting results. Example: play heart-and-soul's melody backwards. Chords to go with it: Amin, F, Dmin, Emin (though you will have to shift a couple notes if you use those chords to make it sound right).

...that's just negation, too! This idea reminds me of alkjash's Hammer and Nails with negation being the hammer and the above mentioned being the nails. One possible implication of this is that any new relationship you learn in math, dance, music, etc. could be applied to all objects/properties you have in math, dance, music, etc. to produce something interesting. I would like to explore this idea in a much greater depth in the future.

Title and review inspired by TurnTrout

Open question: are minimal circuits daemon-free?

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

Note: weird stuff, very informal.

Suppose I search for an algorithm that has made good predictions in the past, and use that algorithm to make predictions in the future.

I may get a "daemon," a consequentialist who happens to be motivated to make good predictions (perhaps because it has realized that only good predictors survive). Under different conditions, the daemon may no longer be motivated to predict well, and may instead make "predictions" that help it achieve its goals at my expense.

I don't know whether this is a real problem or not. But from a theoretical perspective, not knowing is already concerning--I'm trying to find a strong argument that we've solved alignment, not just something that seems to work in practice.

I am pretty convinced that daemons are a real problem for Solomonoff induction. Intuitively, the problem is caused by "too much compute." I suspect that daemons are also a problem for some more realistic learning procedures (like human evolution), though in a different shape. I think that this problem can probably be patched, but that's one of the major open questions for the feasibility of prosaic AGI alignment.

I suspect that daemons *aren't* a problem if we exclusively select for computational efficiency. That is, I suspect that **the fastest way to solve any particular problem doesn't involve daemons**.

I don't think this question has much intrinsic importance, because almost all realistic learning procedures involve a strong simplicity prior (e.g. weight sharing in neural networks).

But I do think this question has deep similarities to more important problems, and that answering this question will involve developing useful conceptual machinery. Because we have an unusually strong intuitive handle on the problem, I think it's a good thing to think about.

Problem statement and intuition

Can the smallest boolean circuit that solves a problem be a daemon? For example, can the smallest circuit that predicts my behavior (at some level of accuracy) be a daemon?

Intuitively, if we have a daemon that is instrumentally or incidentally motivated to solve my problem, then there is some smaller circuit that solves the problem equally well but skips the instrumental reasoning. If my daemon is doing some complex reasoning to answer the question "Should I predict well?" we could just skip straight to the answer "yes." This both makes the circuit smaller, and prevents the circuit from ever deciding not to predict well.

A different perspective on a similar intuition: the daemon is doing some actual cognitive work to solve the problem. Since that computation is being done by the daemon, it is embedded as a smaller circuit. Jessica explores this intuition a bit here. Here we are considering an easy version of the problem, since by taking the smallest circuit we are effectively quantifying over all possible ways of extracting logical information from the daemon.

Instead of showing that minimal circuits can't be daemons, we might end up concluding that they can be. That would be even more interesting.

Another possible outcome is giving a strong argument that captures our intuitions/concerns about daemons, and which clearly doesn't apply to the minimal circuit that solves a problem. In this case we couldn't prove anything positive about the minimal circuit, but we would have "screened off" the possible cause for concern.

Difficulties

The first and most serious difficulty is understanding what we are talking about.

I don't expect to get total clarity on concepts like "daemon" or "optimization" or "generic problem," but we need to have a better grip than we do right now. I expect that we'll develop better concepts in the course of solving the problem, rather than as a precondition for solving the problem (in general I think "define things so that you can prove the theorem" is often the right strategy).

A second difficulty is that the different parts of the computation can be tangled up in an extremely complex way. In an extreme case, the daemon may be cryptographically obfuscated.

We want to show that given any daemon, there is a smaller circuit that solves the problem. The most natural approach is showing how to construct a smaller circuit, given a daemon. But if the daemon is obfuscated, there is no efficient procedure which takes the daemon circuit as input and produces a smaller circuit that still solves the problem.

So we can't find any efficient constructive argument. That rules out most of the obvious strategies.

Hypotheticals: The Direct Application Fallacy

A few years ago, I tried convincing people some commenters that hypotheticals were important even when they weren't realistic. That failed, but I think I've spend enough time reflecting to give this another go. This time, my focus will be on challenging the following common assumption:

> The Direct Application Fallacy: If a hypothetical situation can't conceivably occur, then the hypothetical situation doesn't matter

I chose this name because it assumes that the only purpose of discussing a hypothetical is to know what would happen or what we should do in such a situation. It ignores the other lessons that such a discussion may teach us and how it might have logical consequences for situations that actually do occur.

(**Note:** This post was renamed from: Unrealistic Hypotheticals Still Contain Lessons)

Exploiting Opportunities for Learning

In The Least Convenient Possible World, Scott Alexander considers the classic objection to utilitarianism that it implies that a surgeon should be prepared to harvest the organs of a random traveller if it would allow them to save five other patients. Scott argues that pointing out that the random traveller's organs probably be genetic mismatches, while "technically correct", also "completely misses the point and loses a valuable opportunity to examine the nature of morality". He also notes that responding in this manner leaves too much "wiggle room". Even if we aren't consciously aware of it, we often construct arguments to avoid believing things that we don't want to, so we can improve our rationality by limiting our ability to avoid understanding the other person's perspective. While Scott is referring to people who completely miss the point of the hypothetical, I think that dismissing a hypothetical as unrealistic often also sacrifices opportunities for learning as we'll see below.

Practise Exercises Don't Need to be Real

Imagine that you are an instructor setting problems for your students so that they can learn an area like economics, physics or applied maths. How strongly do you care about these exercises being realistic? I would argue that this isn't very important and that this further applies to philosophy:

1. Simplification: Students may be at a point where a realistic exercise would be quite beyond their abilities. Imagine that you are trying to teach your students how to calculate falling objects. One student complains that you are ignoring air resistance. You try to explain that you can talk about air resistance after you've covered the basics, but they insist that any discussion without air resistance is utterly pointless. Eventually you concede, but most of the class ends up failing the quiz the next week because they weren't ready for the harder problems. Similarly, philosophical problems often assume "no-one will ever know" so that you can discuss moral principles without 90% of the time going into arguing about human psychology and sociology which had nothing to do with the point you were trying to illustrate.

2. Testing for Understanding: Students are often assigned questions as a way to gauge their understanding of a concept. Maybe no object has zero mass, but if someone can't tell you that this should create no gravitational force, they must have a misunderstanding somewhere. Maybe you could ask about an object that weighs 0.01 grams instead, but then they'd have to pull out a calculator. Similarly, even if utility monsters don't exist, they provide a useful tool for clarifying utilitarianism, as it explains why, "greatest good for the greatest number" isn't a completely accurate characterisation. And indeed, there's no reason why some people or organisms mightn't generally experience more utility than others.

3. Realism Trades off Against Other Factors: Perhaps, you could find simple exercises that are realistic or test for understanding with more realistic scenarios. However, your goal is to make your students learn and this is dependent on a whole host of factors. If you insist on questions always being realistic, then this trades off against other dimensions, such as engagement, memorability and time required to construct a situation. This last dimension is particularly important for conversations where people have to be able to construct these situations on the fly.

This is taken for granted when talking about maths and physics, but if you want to learn to deeply understand philosophy, you'll have to accept unrealistic practise questions as well.

Applying the Unrealistic to the Real

In maths, it is very common to take the limit of a formula as some variable, like as x approaches infinity. This technique is very useful for approximations. For example, it's easier to consider the limit as x approaches infinity of (2x^2-x+10)/(x^2+79) than to substitute in a specific value like a million. This is applied constantly throughout programming with Big-O Notation. Even though an infinite dataset is completely unrealistic, this heuristic is still incredibly useful for designing algorithms.

Similarly, when a utilitarian points out that strict versions of deontology will always allow us to construct situations where following the rules cost us infinite utility, the unrealism of the situation doesn't make it irrelevant. Just with Big-O Notation, step 2 of the argument could very well be to scale it down to a more realistic situation. Unfortunately, many people will assume that step 2 isn't coming and judge the argument as flawed at this stage. They may even interrupt the speaker with the objection that the argument isn't realistic. This often negatively affects the conversation, as it pushes the speaker to address step 2, before they've had the opportunity to ensure that everyone has understood step 1.

Being Aware of Limitations

Consider the formula y=10/(x[x-5]). This has two discontinuities at x=0 and x=5. I really want a more practical example, so if you have one, please list in the comment, but let's pretend x represents the number of people and we know there'll always be at least one person in practise. So someone could easily wave away the discontinuity with the x=0 case and completely miss the second one. But if rationality is winning, this isn't it. If they instead looked into it, they'd have realised that more general issue is the division by zero and not overlooked this issue. Yet it is very easy for the person getting it wrong to convince themselves that it is the person trying to figure out the situation with x=0 who is irrational.

Let's suppose that someone is promoting deontology and they aren't worried about theoretical situations. They just want a practical model or heuristic to help them act

morality. If they are proposing a heuristic, they should fully expect it to have limitations and situations where it just completely breaks. And it would probably be useful to know what these limitations are. Some of these limitations mightn't be obvious and the heuristic may even be broken if some of these occur more often than they expect. Discussions of how the model behaves as the utility cost of a principle approaches infinity shouldn't be met by dismissal, but by either biting the bullet or acknowledging that the model seems to break down in those circumstances. It can still be defended as a heuristic or you can assert this kind of situation tends to break our intuitions (see epistemic learned helplessness), but either way it needs to be acknowledged as a limitation that can be weighed up against other limitations. After all, there could be a better model that has a solution to these issues.

Conclusion

One of the key threads of this post has been to not assume that you know where an argument is going. Just because someone is talking about an unrealistic situation, it doesn't follow that they aren't going to tie it back to reality. Further, you shouldn't assume that there's a single path for this to occur. At the very least, I would suggest replacing "This is unrealistic" with "How are you going to tie it back to reality?". The second question is far superior, as it doesn't make the unwarranted assumption that the only purpose of constructing a model is to attempt to directly apply it to reality.

The Alignment Newsletter #6: 05/14/18

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

Highlights

Thoughts on AI Safety via Debate *(Vaniver)*: Vaniver has played several debate games on the website and wrote up some of his experiences. He ended up more optimistic about debate, but still worries that the success of the technique relies on the toy examples being toy.

My opinion: I haven't played the particular debate game that OpenAI released, and so it was interesting to see what sort of strategies emerged. It was initially quite unintuitive to me how debate picks out a particular path in an argument tree, and I think if reading about particular concrete examples (as in this post) would have helped.

Prerequisities: AI safety via debate

Technical AI alignment

Problems

Classification of global catastrophic risks connected with artificial intelligence *(Alexey Turchin et al)*

Scalable oversight

Thoughts on AI Safety via Debate *(Vaniver)*: Summarized in the highlights!

Thoughts on "AI safety via debate" *(gworley)*

Miscellaneous (Alignment)

Open question: are minimal circuits daemon-free? *(Paul Christiano)*: One issue that may arise with an advanced AI agent is that during training we may end up with a part of the AI system developing into a "daemon" -- a consequentialist agent that is optimizing a different goal. This goal may useful as a subcomponent for our AI, but the daemon may grow in power and end up causing the system to optimize for the subgoal. This could lead to catastrophic outcomes, *even if* we have specified a reward function that encodes human values to the top-level AI.

In this post, Paul suggests that these issues would likely go away if we choose the *fastest* program to solve our subgoal. Intuitively, for any daemon that arises as a solution to our problem, for it to cause a bad outcome it must be carrying out

complicated reasoning to figure out whether or not to solve the problem honestly or to try to mislead us, and so we could get a faster program by just not doing that part of the computation. He proposes a particular formalization and poses it as an open question -- if we always choose the minimal (in size) boolean circuit that solves our problem, can a daemon ever arise?

My opinion: I still don't know what to think about daemons -- they do seem to be a problem in Solomonoff induction, but they seem unlikely to arise in the kinds of neural nets we have today (but could arise in larger ones). I would love to see more clarity around daemons, especially since the vast majority of current research would not solve this problem, since it is a problem with the training *process* and not the training *signal*.

Prerequisities: Optimization daemons

AI strategy and policy

To stay ahead of Chinese AI, senators want new commission *(Aaron Mehta)*

AI capabilities

Deep learning

Dynamic Control Flow in Large-Scale Machine Learning *(Yuan Yu et al)*

Exploring the Limits of Weakly Supervised Pretraining *(Dhruv Mahajan et al)*

News

Self-driving cars are here *(Andrew Ng)*: Drive.ai will offer a self-driving car service for public use in Frisco, Texas starting in July, 2018. The post goes into details of how the cars will be rolled out, and some plans for how to make them easier for humans to interact with.

Varieties Of Argumentative Experience

In 2008, Paul Graham wrote How To Disagree Better, ranking arguments on a scale from name-calling to explicitly refuting the other person's central point.

And that's why, ever since 2008, Internet arguments have generally been civil and productive.

Graham's hierarchy is useful for its intended purpose, but it isn't really a hierarchy of *disagreements*. It's a hierarchy of types of response, within a disagreement. Sometimes things are refutations of other people's points, but the points should never have been made at all, and refuting them doesn't help. Sometimes it's unclear how the argument even connects to the sorts of things that in principle could be proven or refuted.

If we were to classify disagreements themselves – talk about what people are doing when they're even having an argument – I think it would look something like this:

Most people are either meta-debating – debating whether some parties in the debate are violating norms – or they're just shaming, trying to push one side of the debate outside the bounds of respectability.

If you can get past that level, you end up discussing facts (blue column on the left) and/or philosophizing about how the argument has to fit together before one side is "right" or "wrong" (red column on the right). Either of these can be anywhere from throwing out a one-line claim and adding "Checkmate, atheists" at the end of it, to cooperating with the other person to try to figure out exactly what considerations are relevant and which sources best resolve them.

If you can get past *that* level, you run into really high-level disagreements about overall moral systems, or which goods are more valuable than others, or what "freedom" means, or stuff like that. These are basically unresolvable with anything less than a lifetime of philosophical work, but they usually allow mutual understanding and respect.

I'm not saying everything fits into this model, or even that most things do. It's just a way of thinking that I've found helpful. More detail on what I mean by each level:

Meta-debate is discussion of the debate itself rather than the ideas being debated. Is one side being hypocritical? Are some of the arguments involved offensive? Is someone being silenced? What biases motivate either side? Is someone ignorant? Is someone a "fanatic"? Are their beliefs a "religion"? Is someone defying a consensus? Who is the underdog? I've placed it in a sphinx outside the pyramid to emphasize that it's not a *bad argument* for the thing, it's just an argument about something completely different.

“Gun control proponents are just terrified of guns, and if they had more experience with them their fear would go away.”

“It was wrong for gun control opponents to prevent the CDC from researching gun statistics more thoroughly.”

“Senators who oppose gun control are in the pocket of the NRA.”

“It’s insensitive to start bringing up gun control hours after a mass shooting.”

Sometimes meta-debate can be good, productive, or necessary. For example, I think discussing “the origins of the Trump phenomenon” is interesting and important, and not just an attempt to bulverizing the question of whether Trump is a good president or not. And if you want to maintain discussion norms, sometimes you do have to have discussions about who’s violating them. I even think it can sometimes be helpful to argue about which side is the underdog.

But it’s not the debate, and also it’s much more fun than the debate. It’s an inherently social question, the sort of who’s-high-status and who’s-defecting-against-group-norms questions that we like a little too much. If people have to choose between this and some sort of boring scientific question about when fetuses gain brain function, they’ll choose this every time; given the chance, meta-debate will crowd out everything else.

The other reason it’s in the sphinx is because its proper function is to guard the debate. Sure, you *could* spend your time writing a long essay about why creationists’ objections to radiocarbon dating are wrong. But the meta-debate is what tells you creationists generally aren’t good debate partners and you shouldn’t get involved.

Social shaming also isn’t an argument. It’s a demand for listeners to place someone outside the boundary of people who deserve to be heard; to classify them as so repugnant that arguing with them is only dignifying them. If it works, supporting one side of an argument imposes so much reputational cost that only a few weirdos dare to do it, it sinks outside the Overton Window, and the other side wins by default.

“I can’t believe it’s 2018 and we’re still letting transphobes on this forum.”

“Just another purple-haired SJW snowflake who thinks all disagreement is oppression.”

“Really, do conservatives have any consistent beliefs other than hating black people and wanting the poor to starve?”

“I see we’ve got a Silicon Valley techbro STEMlord autist here.”

Nobody expects this to convince anyone. That’s why I don’t like the term “ad hominem”, which implies that shamers are idiots who are too stupid to realize that calling someone names doesn’t refute their point. That’s not the problem. People who use this strategy know exactly what they’re doing and are often quite successful. The goal is not to convince their opponents, or even to hurt their opponent’s feelings, but to demonstrate social norms to bystanders. “Ad hominem” has the wrong implications. “Social shaming” gets it right.

Sometimes this works on a society-wide level. More often, it’s an attempt to claim a certain space, kind of like the intellectual equivalent of a gang sign. If the Jets can graffiti “FUCK THE SHARKS” on a certain bridge, but the Sharks can’t get away with

graffiting “NO ACTUALLY FUCK THE JETS” on the same bridge, then almost by definition that bridge is in the Jets’ territory. This is part of the process that creates polarization and echo chambers. If you see an attempt at social shaming and feel triggered, that’s the second-best result from the perspective of the person who put it up. The best result is that you never went into that space at all. This isn’t just about keeping conservatives out of socialist spaces. It’s also about defining what kind of socialist the socialist space is for, and what kind of ideas good socialists are or aren’t allowed to hold.

I think easily 90% of online discussion is of this form right now, including some long and carefully-written thinkpieces with lots of citations. The point isn’t that it literally uses the word “fuck”, the point is that the active ingredient isn’t persuasiveness, it’s the ability to make some people feel like they’re suffering social costs for their opinion. Even really good arguments that are persuasive can be used this way if someone links them on Facebook with “This is why I keep saying Democrats are dumb” underneath it.

This is similar to meta-debate, except that meta-debate can sometimes be cooperative and productive – both Trump supporters and Trump opponents could in theory work together trying to figure out the origins of the “Trump phenomenon” – and that shaming is at least sort of an attempt to resolve the argument, in a sense.

Gotchas are short claims that purport to be devastating proof that one side can’t possibly be right.

“If you like big government so much, why don’t you move to Cuba?”

“Isn’t it ironic that most pro-lifers are also against welfare and free health care? Guess they only care about babies until they’re born.”

“When guns are outlawed, only outlaws will have guns.”

These are snappy but almost always stupid. People may not move to Cuba because they don’t want government *that* big, because governments can be big in many ways some of which are bad, because governments can vary along dimensions other than how big they are, because countries can vary along dimensions other than what their governments are, or just because moving is hard and disruptive.

They may sometimes suggest what might, with a lot more work, be a good point. For example, the last one could be transformed into an argument like “Since it’s possible to get guns illegally with some effort, and criminals need guns to commit their crimes and are comfortable with breaking laws, it might only slightly decrease the number of guns available to criminals. And it might greatly decrease the number of guns available to law-abiding people hoping to defend themselves. So the cost of people not being able to defend themselves might be greater than the benefit of fewer criminals being able to commit crimes.” I don’t think I agree with this argument, and I might challenge assumptions like “criminals aren’t that much likely to have guns if they’re illegal” or “law-abiding gun owners using guns in self-defense is common and an important factor to include in our calculations”. But this would be a reasonable argument and not just a gotcha. The original is a gotcha precisely because it doesn’t invite this level of analysis or even seem aware that it’s possible. It’s not saying “calculate the value of these parameters, because I think they work out in a way where this is a pretty strong argument against controlling guns”. It’s saying “gotcha!”.

Single facts are when someone presents one fact, which admittedly does support their argument, as if it solves the debate in and of itself. It's the same sort of situation as one of the better gotchas – it could be changed into a decent argument, with work. But presenting it as if it's supposed to change someone's mind in and of itself is naive and sort of an aggressive act.

"The UK has gun control, and the murder rate there is only a quarter of ours."

"The fetus has a working brain as early as the first trimester."

"Donald Trump is known to have cheated his employees and subcontractors."

"Hillary Clinton handled her emails in a scandalously incompetent manner and tried to cover it up."

These are all potentially good points, with at least two caveats. First, correlation isn't causation – the UK's low murder rates might not be caused by their gun control, and maybe not all communist countries inevitably end up like the USSR. Second, even things with some bad features are overall net good. Trump could be a dishonest businessman, but still have other good qualities. Hillary Clinton may be crap at email security, but skilled at other things. Even if these facts are true and causal, they only prove that a plan has at least one bad quality. At best they would be followed up by an argument for why this is really important.

I think the move from shaming to good argument is kind of a continuum. This level is around the middle. At some point, saying "I can't believe you would support someone who could do that with her emails!" is just trying to bait Hillary supporters. And any Hillary supporter who thinks it's really important to argue specifics of why the emails aren't that bad, instead of focusing on the bigger picture, is taking the bait, or getting stuck in this mindset where they feel threatened if they admit there's anything bad about Hillary, or just feeling too defensive.

Single studies are better than scattered facts since they at least prove some competent person looked into the issue formally.

"This paper from Gary Kleck shows that more guns actually cause *less* crime."

"These people looked at the evidence and *proved* that support for Trump is motivated by authoritarianism."

"I think you'll find economists have already investigated this and that the minimum wage doesn't cost jobs."

"There's actually a lot of proof by people analyzing many different elections that money doesn't influence politics."

We've already discussed this here before. Scientific studies are much less reliable guides to truth than most people think. On any controversial issue, there are usually many peer-reviewed studies supporting each side. Sometimes these studies are just wrong. Other times they investigate a much weaker subproblem but get billed as solving the larger problem.

There are dozens of studies proving the minimum wage does destroy jobs, and dozens of studies proving it doesn't. Probably it depends a lot on the particular job, the size of the minimum wage, how the economy is doing otherwise, etc, etc, etc. Gary Kleck

does have a lot of studies showing that more guns decrease crime, but a lot of other criminologists disagree with him. Both sides will have plausible-sounding reasons for why the other's studies have been conclusively debunked on account of all sorts of bias and confounders, but you will actually have to look through those reasons and see if they're right.

Usually the scientific consensus on subjects like these will be as good as you can get, but don't trust that you know the scientific consensus unless you have read actual well-conducted surveys of scientists in the field. Your echo chamber telling you "the scientific consensus agrees with us" is definitely not sufficient.

A **good-faith survey of evidence** is what you get when you take all of the above into account, stop trying to devastate the other person with a mountain of facts that can't possibly be wrong, and starts looking at the studies and arguments on both sides and figuring out what kind of complex picture they paint.

"Of the meta-analyses on the minimum wage, three seem to suggest it doesn't cost jobs, and two seem to suggest it does. Looking at the potential confounders in each, I trust the ones saying it doesn't cost jobs more."

"The latest surveys say more than 97% of climate scientists think the earth is warming, so even though I've looked at your arguments for why it might not be, I think we have to go with the consensus on this one."

"The justice system seems racially biased at the sentencing stage, but not at the arrest or verdict stages."

"It looks like this level of gun control would cause 500 fewer murders a year, but also prevent 50 law-abiding gun owners from defending themselves. Overall I think that would be worth it."

Isolated demands for rigor are attempts to demand that an opposing argument be held to such strict invented-on-the-spot standards that nothing (including common-sense statements everyone agrees with) could possibly clear the bar.

"You can't be an atheist if you can't prove God doesn't exist."

"Since you benefit from capitalism and all the wealth it's made available to you, it's hypocritical for you to oppose it."

"Capital punishment is just state-sanctioned murder."

"When people still criticize Trump even though the economy is doing so well, it proves they never cared about prosperity and are just blindly loyal to their party."

The first is wrong because you can disbelieve in Bigfoot without being able to prove Bigfoot doesn't exist – "you can never doubt something unless you can prove it doesn't exist" is a fake rule we never apply to anything else. The second is wrong because you can be against racism even if you are a white person who presumably benefits from it; "you can never oppose something that benefits you" is a fake rule we never apply to anything else. The third is wrong because eg prison is just state-sanctioned kidnapping; "it is exactly as wrong for the state to do something as for a random criminal to do it" is a fake rule we never apply to anything else. The fourth is wrong because Republicans have also been against leaders who presided over good economies and presumably thought this was a reasonable thing to do; "it's impossible

to honestly oppose someone even when there's a good economy" is a fake rule we never apply to anything else.

I don't think these are necessarily badly-intentioned. We don't have a good explicit understanding of what high-level principles we use, and tend to make them up on the spot to fit object-level cases. But here they act to derail the argument into a stupid debate over whether it's okay to even discuss the issue without having 100% perfect impossible rigor. The solution is exactly the sort of "proving too much" arguments in the last paragraph. Then you can agree to use normal standards of rigor for the argument and move on to your real disagreements.

These are related to fully general counterarguments like "sorry, you can't solve every problem with X", though usually these are more meta-debate than debate.

Disputing definitions is when an argument hinges on the meaning of words, or whether something counts as a member of a category or not.

"Transgender is a mental illness."

"The Soviet Union wasn't really communist."

"Wanting English as the official language is racist."

"Abortion is murder."

"Nobody in the US is really poor, by global standards."

It might be important on a social basis what we call these things; for example, the social perception of transgender might shift based on whether it was commonly thought of as a mental illness or not. But if a specific argument between two people starts hinging on one of these questions, chances are something has gone wrong; neither factual nor moral questions should depend on a dispute over the way we use words. This Guide To Words is a long and comprehensive resource about these situations and how to get past them into whatever the real disagreement is.

Clarifying is when people try to figure out exactly what their opponent's position is.

"So communists think there shouldn't be private ownership of factories, but there might still be private ownership of things like houses and furniture?"

"Are you opposed to laws saying that convicted felons can't get guns? What about laws saying that there has to be a waiting period?"

"Do you think there can ever be such a thing as a just war?"

This can sometimes be hostile and counterproductive. I've seen too many arguments degenerate into some form of "So you're saying that rape is good and we should have more of it, are you?" No. Nobody is ever saying that. If someone thinks the other side is saying that, they've stopped doing honest clarification and gotten more into the performative shaming side.

But there *are* a lot of misunderstandings about people's positions. Some of this is because the space of things people can believe is very wide and it's hard to understand exactly what someone is saying. More of it is because partisan echo chambers can deliberately spread misrepresentations or cliched versions of an

opponent's arguments in order to make them look stupid, and it takes some time to realize that real opponents don't always match the stereotype. And sometimes it's because people don't always have their positions down in detail themselves (eg communists' uncertainty about what exactly a communist state would look like). At its best, clarification can help the other person notice holes in their own opinions and reveal leaps in logic that might legitimately deserve to be questioned.

Operationalizing is where both parties understand they're in a cooperative effort to fix exactly what they're arguing about, where the goalposts are, and what all of their terms mean.

"When I say the Soviet Union was communist, I mean that the state controlled basically all of the economy. Do you agree that's what we're debating here?"

"I mean that a gun buyback program similar to the one in Australia would probably lead to less gun crime in the United States and hundreds of lives saved per year."

"If the US were to raise the national minimum wage to $15, the average poor person would be better off."

"I'm not interested in debating whether the IPCC estimates of global warming might be too high, I'm interested in whether the real estimate is still bad enough that millions of people could die."

An argument is operationalized when every part of it has either been reduced to a factual question with a real answer (even if we don't know what it is), or when it's obvious exactly what kind of non-factual disagreement is going on (for example, a difference in moral systems, or a difference in intuitions about what's important).

The Center for Applied Rationality promotes double-cruxing, a specific technique that helps people operationalize arguments. A double-crux is a single subquestion where both sides admit that if they were wrong about the subquestion, they would change their mind. For example, if Alice (gun control opponent) would support gun control if she knew it lowered crime, and Bob (gun control supporter) would oppose gun control if he knew it would make crime worse – then the only thing they have to talk about is crime. They can ignore whether guns are important for resisting tyranny. They can ignore the role of mass shootings. They can ignore whether the NRA spokesman made an offensive comment one time. They just have to focus on crime – and that's the sort of thing which at least in principle is tractable to studies and statistics and scientific consensus.

Not every argument will have double-cruxes. Alice might still oppose gun control if it only lowered crime a little, but also vastly increased the risk of the government becoming authoritarian. A lot of things – like a decision to vote for Hillary instead of Trump – might be based on a hundred little considerations rather than a single debatable point.

But at the very least, you might be able to find a bunch of more limited cruxes. For example, a Trump supporter might admit he would probably vote Hillary if he learned that Trump was more likely to start a war than Hillary was. This isn't quite as likely to end the whole disagreemnt in a fell swoop – but it still gives a more fruitful avenue for debate than the usual fact-scattering.

High-level generators of disagreement are what remains when everyone understands exactly what's being argued, and agrees on what all the evidence says,

but have vague and hard-to-define reasons for disagreeing anyway. In retrospect, these are probably why the disagreement arose in the first place, with a lot of the more specific points being downstream of them and kind of made-up justifications. These are almost impossible to resolve even in principle.

“I feel like a populace that owns guns is free and has some level of control over its own destiny, but that if they take away our guns we’re pretty much just subjects and have to hope the government treats us well.”

“Yes, there are some arguments for why this war might be just, and how it might liberate people who are suffering terribly. But I feel like we always hear this kind of thing and it never pans out. And every time we declare war, that reinforces a culture where things can be solved by force. I think we need to take an unconditional stance against aggressive war, always and forever.”

“Even though I can’t tell you how this regulation would go wrong, in past experience a lot of well-intentioned regulations have ended up backfiring horribly. I just think we should have a bias against solving all problems by regulating them.”

“Capital punishment might decrease crime, but I draw the line at intentionally killing people. I don’t want to live in a society that does that, no matter what its reasons.”

Some of these involve what social signal an action might send; for example, even a just war might have the subtle effect of legitimizing war in people’s minds. Others involve cases where we expect our information to be biased or our analysis to be inaccurate; for example, if past regulations that seemed good have gone wrong, we might expect the next one to go wrong even if we can’t think of arguments against it. Others involve differences in very vague and long-term predictions, like whether it’s reasonable to worry about the government descending into tyranny or anarchy. Others involve fundamentally different moral systems, like if it’s okay to kill someone for a greater good. And the most frustrating involve chaotic and uncomputable situations that have to be solved by *metis* or *phronesis* or similar-sounding Greek words, where different people’s Greek words give them different opinions.

You can always try debating these points further. But these sorts of high-level generators are usually formed from hundreds of different cases and can’t easily be simplified or disproven. Maybe the best you can do is share the situations that led to you having the generators you do. Sometimes good art can help.

The high-level generators of disagreement can sound a lot like really bad and stupid arguments from previous levels. “We just have fundamentally different values” can sound a lot like “You’re just an evil person”. “I’ve got a heuristic here based on a lot of other cases I’ve seen” can sound a lot like “I prefer anecdotal evidence to facts”. And “I don’t think we can trust explicit reasoning in an area as fraught as this” can sound a lot like “I hate logic and am going to do whatever my biases say”. If there’s a difference, I think it comes from having gone through all the previous steps – having confirmed that the other person knows as much as you might be intellectual equals who are both equally concerned about doing the moral thing – and realizing that both of you alike are controlled by high-level generators. High-level generators aren’t biases in the sense of mistakes. They’re the strategies everyone uses to guide themselves in uncertain situations.

This doesn’t mean everyone is equally right and okay. You’ve reached this level when you agree that the situation is complicated enough that a reasonable person with reasonable high-level generators could disagree with you. If 100% of the evidence

supports your side, and there's no reasonable way that any set of sane heuristics or caveats could make someone disagree, then (unless you're missing something) your opponent might just be an idiot.

Some thoughts on the overall arrangement:

1. If anybody in an argument is operating on a low level, the entire argument is now on that low level. First, because people will feel compelled to refute the low-level point before continuing. Second, because we're only human, and if someone tries to shame/gotcha you, the natural response is to try to shame/gotcha them back.

2. The blue column on the left is factual disagreements; the red column on the right is philosophical disagreements. The highest level you'll be able to get to is the *lowest* of where you are on the two columns.

3. Higher levels require more vulnerability. If you admit that the data are mixed but seem to slightly favor your side, and your opponent says that every good study ever has always favored his side plus also you are a racist communist – well, you kind of walked into that one. In particular, exploring high-level generators of disagreement requires a lot of trust, since someone who is at all hostile can easily frame this as "See! He admits that he's biased and just going off his intuitions!"

4. If you hold the conversation in private, you're almost guaranteed to avoid everything below the lower dotted line. Everything below that is a show put on for spectators.

5. If you're intelligent, decent, and philosophically sophisticated, you can avoid everything below the higher dotted line. Everything below that is either a show or some form of mistake; everything above it is impossible to avoid no matter how great you are.

6. The shorter and more public the medium, the more pressure there is to stick to the lower levels. Twitter is great for shaming, but it's almost impossible to have a good-faith survey of evidence there, or use it to operationalize a tricky definitional question.

7. Sometimes the high-level generators of disagreement are other, even more complicated questions. For example, a lot of people's views come from their religion. Now you've got a whole different debate.

8. And a lot of the facts you have to agree on in a survey of the evidence are also complicated. I once saw a communism vs. capitalism argument degenerate into a discussion of whether government works better than private industry, then whether NASA was better than SpaceX, then whether some particular NASA rocket engine design was better than a corresponding SpaceX design. I never did learn if they figured whose rocket engine was better, or whether that helped them solve the communism vs. capitalism question. But it seems pretty clear that the degeneration into subquestions and discovery of superquestions can go on forever. This is the stage a lot of discussions get bogged down in, and one reason why pruning techniques like double-cruxes are so important.

9. Try to classify arguments you see in the wild on this system, and you find that some fit and others don't. But the main thing you find is *how few real arguments there are*. This is something I tried to hammer in during the last election, when people were

complaining “Well, we tried to debate Trump supporters, they didn’t change their mind, guess reason and democracy don’t work”. Arguments above the first dotted line are rare; arguments above the second basically nonexistent in public unless you look really hard.

But what’s the point? If you’re just going to end up at the high-level generators of disagreement, why do all the work?

First, because if you do it right you’ll end up respecting the other person. Going through all the motions might not produce agreement, but it should produce the feeling that the other person came to their belief honestly, isn’t just stupid and evil, and can be reasoned with on other subjects. The natural tendency is to assume that people on the other side just don’t know (or deliberately avoid knowing) the facts, or are using weird perverse rules of reasoning to ensure they get the conclusions they want. Go through the whole process, and you will find some ignorance, and you will find some bias, but they’ll probably be on both sides, and the exact way they work might surprise you.

Second, because – and this is total conjecture – this deals a tiny bit of damage to the high-level generators of disagreement. I think of these as Bayesian priors; you’ve looked at a hundred cases, all of them have been X, so when you see something that looks like not-X, you can assume you’re wrong – see the example above where the libertarian admits there is no clear argument against this particular regulation, but is wary enough of regulations to suspect there’s something they’re missing. But in this kind of math, the prior shifts the perception of the evidence, but *the evidence also shifts the perception of the prior*.

Imagine that, throughout your life, you’ve learned that UFO stories are fakes and hoaxes. Some friend of yours sees a UFO, and you assume (based on your priors) that it’s probably fake. They try to convince you. They show you the spot in their backyard where it landed and singed the grass. They show you the mysterious metal object they took as a souvenier. It seems plausible, but you still have too much of a prior on UFOs being fake, and so you assume they made it up.

Now imagine another friend has the same experience, and also shows you good evidence. And you hear about someone the next town over who says the same thing. After ten or twenty of these, maybe you start wondering if there’s something to all of this UFOs. Your overall skepticism of UFOs has made you dismiss each particular story, but each story has also dealt a little damage to your overall skepticism.

I think the high-level generators might work the same way. The libertarian says “Everything I’ve learned thus far makes me think government regulations fail.” You demonstrate what looks like a successful government regulation. The libertarian doubts, but also becomes slightly more receptive to the possibility of those regulations occasionally being useful. Do this a hundred times, and they might be more willing to accept regulations in general.

As the old saying goes, “First they ignore you, then they laugh at you, then they fight you, then they fight you half-heartedly, then they’re neutral, then they then they grudgingly say you might have a point even though you’re annoying, then they say on balance you’re mostly right although you ignore some of the most important facets of the issue, then you win.”

I notice SSC commenter John Nerst is talking about a science of disagreement and has set up a subreddit for discussing it. I only learned about it after mostly finishing this

post, so I haven’t looked into it as much as I should, but it might make good followup reading.

Predicting Future Morality

Robin Hanson suggests that recent changes in moral attitudes (in the last few hundred years) are better explained by changing circumstances than by progress in moral reasoning.

This seems plausible to me. It also seems likely that there would be a bit of a lag between the change in circumstance and the common acceptance of the new morality. (The sexual revolution following the introduction of the pill seems like a good example.)

Suppose this is broadly right -- that moral attitudes follow circumstances. Is there anything we can predict about where moral attitudes will be in the next few decades (or economic doublings), based on either recent technological or economic changes, or on those we can see on the horizon?

Bounded Rationality: Two Cultures

This is a linkpost for
https://www.tandfonline.com/doi/pdf/10.1080/1350178X.2014.965908?needAccess=true

Have not read this in any detail, but looked like interesting food for thought. Here's the first section of the paper.

Introduction and Outline

Bounded rationality does not speak with one voice. This is not only because bounded rationality is researched in various fields such as economics, psychology, engineering and management. Even within a single field such as economics, there are clear differences. For example, Selten (2001) rejects the optimization of a utility function as an expression of bounded rationality, contrary to the standard approach of behavioral economics as in bargaining games by Fehr and Schmidt (1999). There are multiple views of bounded rationality, as many authors including Rubinstein (1998) have pointed out.

The first contribution of this article is to analyze the formal modeling used to describe people's bounded rationality. At the risk of oversimplifying, I distinguish between two cultures, which I call 'idealistic' and 'pragmatic'. At a first approximation, the idealistic culture pursues a minimum departure from the neoclassical-economics framework of unbounded rationality, which assumes the ideals of omniscience and optimization of a utility function and adds factors such as inequity aversion or probability weighting to the utility function. On the other hand, the pragmatic culture holds that people sometimes ignore information and use simple rules of thumb in order to achieve satisfactory outcomes. A detailed discussion of the differences in modeling between the two cultures is provided in Section 2. The reality of the cultures and their differences is demonstrated by examples drawn from the literatures on risky choice and bargaining games. Note that it does not make sense to try to perfectly map specific researchers or programs of research to one or the other culture; for example, Amos Tversky worked on both cultures, with prospect theory being an idealistic model and elimination by aspects being a pragmatic model.

Although the distinction between the idealistic and pragmatic cultures of bounded rationality can be criticized, as all binary distinctions can be, it provides food for thought and new insights. I aim at emulating Breiman's (2001) analysis of two cultures in statistics. Breiman argued that there exist two cultures that lead to two very different kinds of statistical theory and practice, proof-based and data-driven. Analogously, I argue in Section 3 that the idealistic and pragmatic cultures tell two very different stories about people's bounded rationality and how to improve it. This is the second contribution of this article. Echoing Morgan (2001), I conclude that these stories play a vital role in our understanding of the economic world and the economic policies we develop. I also venture outside economics and psychology to consider the idealistic and pragmatic cultures in engineering and management. I argue that the idealistic culture reigns in these fields, but at the same time the pragmatic culture is gaining momentum. Section 4 concludes the article.

Whole paper is ~7k words, or ~14 mins read time. Also, I'm at a university, so let me know if the link doesn't actually work for people outside of academia.

Biodiversity for heretics

Epistemic status: Not very confident in my conclusions here. Could be missing big things. Information gained through many hours of reading about somewhat-related topics, and a small few hours of direct research.

Summary: Biodiversity research is popular, but interpretations of it are probably flawed, in that they're liable to confuse causation and correlation. Biodiversity can be associated with lots of variables that are rarely studied themselves, and one of these, not "biodiversity" in general, might cause an effect. (For example, more biodiverse ecosystems are more likely to include a particular species that has significant effects on its own.) I think "biodiversity" is likely overstudied compared to abundance, biomass, etc., because it's A) easier to measure and B) holds special and perhaps undue moral consideration.

From what I was told, *biodiversity* – the number of species present in an environment – always seemed to be kind of magical. Biodiverse ecosystems are more productive, more stable over time, produce higher crop yields, and are more resistant to parasites and invaders. Having biodiversity in one place increases diversity in nearby places, even though diversity *isn't even one thing* (forgive me for losing my citation here). Biodiverse microbiomes are healthier for humans. Biodiversity is itself the most important metric of ecosystem health. The property "having a suite of different organisms living in the same place" just seems to have really incredible effects.

First of all – quickly – some of what I was told isn't actually true. More diverse microbiomes in bodies aren't always healthier for humans or more stable. The effects of losing species in ecosystems varies a ton. More biodiverse ecosystems don't necessarily produce more biomass.

That said, there's still plenty of evidence that biodiversity correlates with *something*.

But: biodiversity research and its interpretations have problems. Huston (1997)introduced me to a few very concrete ways this can turn up misleading or downright inaccurate results.

Our knowledge about biodiversity's effects on ecosystems comes from either experiments, in which biodiversity is manipulated in a controlled setting; or in observations of existing ecosystems. Huston identifies a few ways that these have, historically, given us bad or misleading data:

1. Biotic or abiotic conditions, either in observations or experiments, are altered between groups. (E.g. you pick some sites to study that are less and more biodiverse, but the more-biodiverse sites are that way because they get more rainfall – which obviously is going to have other impacts)
2. Species representing the "additional biodiversity" in experiments aren't chosen randomly, they're known to have some ecosystem function.
3. Increasing the number of species increases the chance that *one* or *a few* of the added species will have some notable ecosystem effect on their own.

I'm really concerned about **(3)**.

To show why, let's imagine aliens who come to earth and want to study how humans work. They abduct random humans from across the world and put them in groups of various sizes.

Building walls

The aliens notice that the human civilizations have walls. They give their groups of abducted humans blocks and instruct them to build simple walls.

It turns out that larger groups of humans can build, on average, proportionally longer walls. The aliens conclude that wall-building is a property of larger groups of humans.

Building radios

The aliens also notice that human civilizations have radios. They give their groups of abducted humans spare electronic parts, and instruct them to build a radio.

Once again, it turns out that larger groups of humans are proportionally more likely to be able to build a radio. The aliens conclude that radio-building, too, is a property of large groups of humans.

The mistake the aliens are making is in assuming that wall- and radio-building are functions of "the number of humans you have in one place". More people can build a longer simple wall, because there's more hands to lift and help. But when it comes to building radios, a larger group just increases the chance that *at least one human in the group will be an engineer.*

To the aliens, who don't know about engineers, "number of humans" *kind of* relates to the thing they're interested in – they will notice a correlation – but they're making a mistake by just waving their hands and saying that mostly only large groups of humans possess the intelligence needed to build a radio, perhaps some sort of hivemind.

Similarly, we'd make a mistake by looking at all the strange things that happen in diverse ecosystems, and saying that these are a magical effect that appears whenever you get large numbers of different plants in the same field. I wonder how often we notice that something correlates with "biodiversity" and completely miss the actual mechanism.

Aside from a specific species or couple of species in combination that have a particular powerful effect on ecosystems, what else might biodiversity correlate to that's more directly relevant? How about *abundance* (the number of certain organisms of some kind present)? Or *biomass* (the combined weight of organisms)? Or environmental conditions, like the input of energy? Or the amount of biomass turnover, or the amount of predation, etc., etc.?

I started wondering about this while doing one of my several projects that relate to abundance in nature. We should still study biodiversity, sure. But the degree to which biodiversity has been studied compared to, say, abundance, has lead us to a world where we know there are 6,399 species of mammals, but nobody has any idea – even very roughly – how many mammals there are. Or how we're pretty sure that there are about 7.7 million species of animals, plus or minus a few hundred thousand, which is a

refinement of many previous estimates of the same thing – and then we have about two people (one of whom is wildly underqualified) trying to figure out how many animals there are at all.

It's improving. A lot of recent work focuses on *functional biodiversity*. This is the diversity of *properties* of organisms in an environment. Instead of just recording the number of algae species in a coastal marine shelf, you might notice that some algae crusts on rocks, some forms a tall canopy, some forms a low canopy, and some grows as a mat. It's a way of separating organisms into niches and into their interactions with the environment.

Functional diversity seems to better describe ecosystem effects than diversity alone (as described e.g. here). That said, it still leaves the door open for **(3)** – looking at functional diversity means you must know something about the ecosystem, but it's not enough to tell you what's causing the effect in and of itself.

To illustrate why:

Every species has some functional properties that separate it from other species – some different interactions, some different niche or physical properties, etc. We can imagine increasing biodiversity, then, as "a big pile of random variables."

It turns out that when you start with a certain environment and slowly add or remove "a big pile of random variables", *that changes the environment's properties.* Who would have thought?

So is biodiversity instrumentally relevant to humans?

1. There are *sometimes* solid explanations for why biodiversity itself might be relevant to ecosystems, e.g. the increased selection for species complementary over time theory.
2. Biodiversity probably *correlates* to the things that studies claim it correlates to, including the ones that find significant environmental effects. I just claim that often, biodiversity is plausibly falsely described as the controlling variable rather than one of its correlates. (That said, there are reasons we might expect people to overstate its benefits – read on.)

If this is true, and biodiversity itself isn't the driving force we make it out to be, why does everyone study it?

Firstly, I think biodiversity is easier to measure than, say, individual properties, or abundance. Looking at the individual properties and traits of each species in the environment is its whole own science, specific to that particular species and that particular environment. It would be a ridiculous amount of work.

But when we try to get the measure of an ecosystem without this really deep knowledge, we turn into the alien scientists – replacing a precise and intricate interaction with a separate but easier-to-measure variable that *sort of* corresponds with the real one.

What about studying one of the other ecosystem properties, like abundance? I'm guessing that in the modern research environment, you'd basically have to be

collecting biodiversity data anyways.

> ***Researcher:*** *We found 255 beetles in this quadrant!*
> ***PI:*** *What kind?*
> ***Researcher:*** *You know. Beetles.*

…And if you're identifying everything you find in an environment anyways, it's easier to just keep track of how many different things you find, rather than do that *plus*exhaustively search for every individual.

This is just speculation, though.

Secondly, a lot of people believe that species and ecosystems are a special *moral unit*(independent of any effects or benefits they might have on humans). That's why people worry about losing the parasites of endangered species, or wonder if we shouldn't damage biodiversity by eradicating diseases.

And… it's hard to explain why this seems wrong to me, but I'll try. I get it. Environmentalism is compelling and widespread. It was the background radiation of virtually almost every interaction with nature I had growing up. It was taken for granted that every drop of biodiversity was a jewel with value beyond measure, that endangered species were inherently worth going to great lengths to protect and preserve, that ecosystems are precariously balanced configurations that should be defended as much as possible from encroachment by humans. Under this lens, *of course* the number of species present is the default measurement – the more biodiversity preserved from human destruction, the more intricate and elaborate the ecosystem (introduced species excepted), the better.

And… doesn't that seem a little limited? Doesn't that seem like a sort of arbitrary way to look at huge parts of the world we live in? It's not worth throwing out, but perhaps it deserves a little questioning. Where else could we draw the moral lines?

Personally, I realized my morality required me to treat animals as moral patients. This started with animals directly used by humans, but then got me re-examining the wild animals I'd been so fond of for so long.

Currently, I put individual animals and species in mostly-separated mental buckets. A species, a particular pattern instantiated by evolution acting on rocks and water over time, is important – but it's important because it's *beautiful*, like a fantastic painting made over decades by a long-dead artist. We value aesthetics, and interpretations, and certainly the world would be worse off without a piece of beauty like this one.

But an individual matters morally because it feels. It cares, it thinks, it feels joy, it suffers. We know because we are one, and because the same circuits and incentives that run in our brains also run in the brains of the cats, chickens, songbirds, insects, earthworms, whale sharks, and bristlemouths that we share this lonely earth with.

We might say that a species "suffers" or "is in pain", the same way that a city "is in pain", and we might mean several different things by that. We might say many of the individuals in the collective suffer. Or we might mean that the species is degraded somehow the way art is degraded – lessened in quantity, less likely to survive into the future, changing rapidly, etc. But it seems like a stretch to call that *pain*, in the way that being eaten alive is *pain*.

Obviously, at some point, you have to make trade-offs over what you care about. I don’t have my answers worked out yet, but for now, I put a lot more value on the welfare of individual animals than I used to, and I care less about species.

I don’t expect this viewpoint to become widespread any time soon. But I think it’s possible that the important things in nature aren’t the ones we’ve expected, and that under other values, properties like abundance and interactions deserve much more attention (compared to biodiversity) than they have now.

Crossposted from my personal blog.

Tech economics pattern: "Commoditize Your Complement"

This is a linkpost for https://www.gwern.net/Complement

> Joel Spolsky in 2002 identified a major pattern in technology business & economics: the pattern of "commoditizing your complement", an alternative to vertical integration, where companies seek to secure a chokepoint or quasi-monopoly in products composed of many necessary & sufficient layers by dominating one layer while fostering so much competition in another layer above or below its layer that no competing monopolist can emerge, prices are driven down to marginal costs elsewhere in the stack, total price drops & increases demand, and the majority of the consumer surplus of the final product can be diverted to the quasi-monopolist.
>
> A classic example is the commodification of PC hardware by the Microsoft OS monopoly, to the detriment of IBM & benefit of MS.
>
> This pattern explains many otherwise odd or apparently self-sabotaging ventures by large tech companies into apparently irrelevant fields, such as the high rate of releasing open-source contributions by many Internet companies or the intrusion of advertising companies into smartphone manufacturing & web browser development & statistical software & fiber-optic networks & municipal WiFi & radio spectrum auctions (Google): they are pre-emptive attempts to commodify another company elsewhere in the stack, or defenses against it being done to them.

Societal Growth Requires Rehabilitation

Disclaimer: My intent is not to criticize growth mindset as initially intended, but to criticize the version of straw growth mindset that has become a rationalist meme, particularly by pointing out its relation to some of the problems we have on the community level.

"Growth mindset" ranges from a sort of rallying cry to a "that's what she said" sort of joke, depending on what crowd you run with, but underneath all of this is an attitude that we can get better. We use the phrase to lift ourselves up, to tell ourselves that no matter what our current problems, we can grow and become stronger. We treat technology similarly; someday, our cars will drive us and death will be cured. In the future, things will be better -- assuming X-risk doesn't take us all out first.

Sadly, this mindset seems to leave little room for the struggling. "Growth mindset" gets used to mean "everything is good and getting better" rather than "bad things are getting less bad", which erases those for whom "everything is good" seems like a false statement. Cryonics and self-driving cars only exist for those who can afford it. On a social level, only those with the resources for personal growth can realistically work on themselves to the extent where "growth mindset" is actually a realistic phrase. Ideally, we'd work on this by making things like therapy and education more accessible. Ultra-ideally, we'd also start teaching things like EQ and metacognition in public schools, and work toward decreasing stigma around mental illness, therapy, and self help.

The pervasiveness of growth mindset does not seem unusual from the average person's perspective. Personally, the moment that made me question it is when I was working with a special needs class whose teacher was assigned to do a lecture on it. This was a class that was considered "moderate to severe"; most of the students were nonverbal and struggled to read or grasp abstract concepts at all. The thought that these kids gained anything from a lecture titled "Are you a tree or a brick?" is absurd.

Growth mindset, when taught poorly, will imply that anything can be achieved through effort. This kind of attitude can be harmful because for some, it is simply not true. Just as most of my students couldn't comprehend growth mindset, there are many others who will never be able to do things that the average person considers necessary for modern life. While growth mindset may work well to combat the attitude that success is purely inherent, the way it is often presented swings the pendulum too far in the other direction. In a sense, we have yet another version of the nature vs nurture debate, with a similar answer: success consists of both effort and circumstance.

If we want to be the kind of community that applauds social growth, we also need to be the kind of community that assists the struggling before things get dire. High status people touting "growth mindset" while many others only one or two degrees away struggle with suicidality creates a "rich get richer" kind of social landscape. We can't get better as a society without helping those who are worse off, and that starts with our own community. There is no true growth mindset without rehabilitation.

Please Take the 2018 Effective Altruism Survey!

This year, the EA Survey volunteer team is proud to announce the launch of **the 2018 Effective Altruism Survey**.

-

PLEASE TAKE THIS SURVEY NOW BY CLICKING HERE :)

(If you want to share the survey with others, please use this fancy share link with referral tracking: https://www.surveymonkey.com/r/3HYW9MW)

-

What is this?

This is the fourth EA survey we've done, coming hot off the heels of the 2017 EA Survey (announcement here, analysis here), the 2015 EA Survey (announcement here, analysis here), and the 2014 EA Survey (announcement here, analysis here).

We hope this survey will produce very useful data on the growth and changing attitudes of the EA Community. In addition to capturing a snapshot of what EA looks like now, we also intend to do longitudinal analysis to see how our snapshot has been changing.

We're also using this as a way to build up the online EA community, such as featuring people on a global map of EAs and with a list of EA Profiles. This way more people can learn about the EA community. We will ask you in the survey if you would like to join us, but you do not have to opt-in and you will be opted-out by default.

Who should take this survey?

Anyone who is reading this should take this survey, even if you don't identify as an "effective altruist".

How does the survey work?

All questions are optional (apart from one important question to verify that your answers should be counted). Most are multiple choice and the survey takes around 10-20 minutes.

At the end of the survey there is an 'Extra Credit' section with some more informal questions and opportunities for comment - definitely feel free to skip these questions.

All results will be aggregated, anonymized, and made available to members of the EA community, so we can better share useful knowledge among each other. Within the survey, you'll have the option to publicly share selected information about yourself in several EA venues, if you opt in.

Who is behind this?

The annual EA Survey is conducted by Rethink Charity, with support and assistance from several EA community organizations intending to help the community better understand our actions, values, demographics, and ideas.

Last chance to participate in the 2018 EA Survey

This year's EA Survey will soon be closed.

You can help the community understand itself better by taking the survey via the button below before the May 26 deadline. The survey takes about 10-20 minutes to complete.

The annual EA survey is one of the best tools we have for interpreting our actions, values, demographics, and ideas as a community. See the most popular EA Survey post of 2017.

Take the 2018 EA Survey

All results will be aggregated, anonymized, and made publicly available, so we can share useful knowledge among each other. Within the survey you'll be able to opt-in to share selected information about yourself in several EA venues.

Thanks and best wishes,
Peter Hurford and the Rethink Charity survey team

(P.S. We apologize for reposting this on LW a second and final time, but the previous post was only available on the main page for two days and had a very small amount of completions relative to LessWrong's reach. We'd love to see more LessWrong users complete the survey, so we reposted to assure the survey is more properly balanced across our communities of interest. If you want to share the survey with others, please use this fancy share link with referral tracking: https://www.surveymonkey.com/r/3HYW9MW.)

There is a war.

This is a linkpost for http://benjaminrosshoffman.com/there-is-a-war/

Households vs markets

The first symptom was the clutter on the kitchen counter. One cutting board, two pans, one knife. My colleagues had arrived at the rented house the day before, so they'd had plenty of time to arrange things to their liking. I was sure the clutter was not to their liking, if they noticed it at all.

A teachable moment. Instead of tidying the counter myself, and accreting a small amount of resentment, I suggested to one of them that she think of the things on the counter as things that were in her power to arrange however she liked, to suit her taste, selfishly. (“Just as you might optimize your text editor to suit your workflow,” my other colleague chimed in.) She took this suggestion, and spent a few minutes arranging and rearranging the items on the counter. She put away the knife in the knife block.

A puzzle. She noted that she doesn’t usually think of the items in her home this way. Instead, household chores feel as though they are impositions from an abstract, outside authority. She was capable of accessing this other, more pleasant way of working on the things around her. Why wasn’t it natural for her to feel that way in her own home?

An hypothesis. In our usual mode of life, there is a separation between a job - which is done for someone else, to satisfy someone else’s standards, outside the home - and consumption, which is at least ostensibly done to suit one’s own taste. One of the goods you can buy with an income from a job is a nice place to live, and you can also buy services to keep the place clean and tidy. For the most part, you maintain the place you live by leaving it, and entering the domain of an outside authority. Household chores are the remainder that cannot efficiently be outsourced, or an echo of a previous era in which such outsourcing was less common.

A household. I pointed out that in the past, people mostly did not work for a salary. Instead, they worked on things in their local environment. For the most part, this was not in order to receive money, but in order to have the thing, even if it was part of a process directed towards producing goods for trade. Home improvement used to be, not a hobby or special interest, but what one did if one wanted an improved home. Even those without a substantial amount of arable land to work might keep a garden. A minority engaged in trades or professions. Loops were mostly closed, either at the level of the household or at the level of the village.

A leak. My colleague pointed out that taxation means you can’t actually just have a closed loop anymore, even if you were willing to do without modern inconveniences. To occupy a space, even a space you legally own, unencumbered by any debt or lien, one must pay rent in the state’s currency. If you don’t pay, you have to leave. This means that at least some of a household’s activity has to be, in effect, defense spending: providing the state with mercenary services, directly or indirectly.

External standards. It’s worth noting that the phrase “good enough for government work” was originally a compliment; it referred to the exacting procurement standards of the US Government, more precise than the standards people were otherwise accustomed to. Things need to be machined with more legible precision if the customer can’t just go consult with the craftsman.

Tribute. Likewise, it’s a commonplace that a village can’t improve its prospects by taking in each other’s washing. But, actually, a village can’t improve its prospects by engaging in *any* sort of closed-loop commerce; something must be procured as tribute. In a country with an income tax, exchanging services - if legibly enough to be detected - actually causes an outflow of wealth.

Why are we doing this to each other?

Taxation is mobilization

Imagine a peaceful village with a closed economy. Much of the villagers’ productive activity is not transactional, but simply working to make the village a better place. The reward for producing food is that the village has more food. Perhaps some transactions *are* market-based, though few will be arms-length. If so, it is easy to imagine that a precious metal like gold might be, if not an actual medium of exchange, at least an unit of account. In addition, people often maintain their own households, and improve them. Again, the reward is an improved household, not money.

This way of life affords a freedom to which the people of my generation and the generation before are not accustomed: if you don’t like the activities by which one might get gold, you can just not do those things. Accordingly, social policing has to be comparatively direct. If you want to exclude someone from village life, you have to actually coordinate to run them out of the village. Especially if they own land to deprive them of their livelihood you have to physically expel them from their own house and land. Anything less is merely an inconvenience.

Suppose the Golden Benevolent Empire decides that this village, like many others along its border, needs defending against the Barbarian Horde. The Golden Benevolent Empire has limited resources, and insists that villages contribute to their own defense, in the form of professional soldiers.

One way to do this would be to demand specific in-kind contributions from each village. But centralized resource allocation of this sort is tricky, especially without reliable record collection and a very large computer, so it settles on a simpler expedient. Require a certain, small quantity of gold from each villager (a head tax, or maybe a land tax), on pain of death.

Now, everyone in the village needs to come up with at least a little gold. Not only will some people be happy to accept pay as professional soldiers, but the other villagers will be eager to provide whatever goods and services your army needs - so long as you are willing to pay in the same coin you demand from the villagers. Like magic, your army is fed, clothed, and equipped.

But something else happens. In the village, gold - which was perhaps used occasionally, either for transactions of ritual significance, the occasional foreign trade,

or largely optional market transactions - is now just as necessary to sustain life, as food and water and shelter are.

Different villages may solve this problem differently. Some may coordinate to produce trade goods that can be sold elsewhere, otherwise retaining their accustomed mode of life. Others, already integrated into larger markets, may simply shift their production person-by-person slightly more towards trade. On the margin, closed-loop households are destroyed; no one can be fully self-sufficient. Villages nearest the military encampments will be in a good position to serve as intermediaries. Villagers from poor or remote villages, that can't afford their tax, may move closer to these trade centers, or accept employment as mercenaries themselves.

In any case, all across the Golden Benevolent Empire, production shifts towards things that can be traded for gold, and people are more eager to keep accounts. Commerce, in other words, is booming.

In some ways, this points towards our present-day situation. We pay taxes, and this would be sufficient to explain why we can't have totally closed loops. Small towns are dying, and people are moving towards participation in the global economy when they can. What it doesn't explain is our apparent reliance on this process, such that when the demands made on us by the economy decline, our society doesn't revert to more home production; instead, we anxiously seek out employment, and hope the state will stimulate additional demand.

MobsterBucks: an interlude

The need to concentrate resources to administer the Golden Benevolent Empire has led to the emergence of a few very large cities, where its administrators and soldiers have gold to spend. This drives up wages, which draws peasants from the poorer villages, the villages which are not productive enough to bear the burden of their land taxes. There are also great merchant ports, enabling gains from trade that could not have been realized before so many lands were unified and pacified under Imperial rule. Both of these cities also draw immigrants from other lands, because gold has no national allegiance.

The growth of these cities has outpaced the Empire's ability to enforce its own laws, and in one of them, something strange and new is happening.

It is not at all uncommon that immigrant communities sharing a nation of origin have their own customs, norms, and ways of enforcing these. This can take the refined form of the Ottoman millet system, or the crude form of street gangs.

What is uncommon is how one of these gangs is administering its territory.

It is not at all uncommon for a gang to demand payment from the residents of its territory, in exchange for "protection" - both from outsiders, and from the gang itself. But in many neighborhoods, the amount of hard currency available is nearly nil - the people live on credit. Demanding in-kind payment is administratively difficult, for the same reason the Golden Benevolent Empire ultimately decided against it. One enterprising gang has a solution: MobsterBucks.

The proposition is simple, perhaps inspired by a folk tale about a village oppressed by archers. Each person in this gang's territory has to pay one MobsterBuck per year, lest

they meet with some sort of mishap like a broken kneecap or worse. How does one acquire MobsterBucks? Well, the gang is the sole issuer of said currency, and is happy to exchange it for goods and services it needs. Soon, the internal economy of this community is denominated primarily in MobsterBucks.

Having implemented this scheme, this enterprising gang now needs a monetary policy. If it spends too many MobsterBucks, then people will be less afraid of injury for want of MobsterBucks, and the exchange value of the currency will decline. This can of course be corrected by simply raising the rate of taxation - demanding more MobsterBucks from each person.

On the other hand, with an imbalance in the other direction, an inefficiently high number of people will be kneecapped, permanently reducing the productive capacity of the gang's territory. Another problem with deflation is that inside the gang's territory, debts are now denominated in the convenient unit of MobsterBucks. Anything that increases demand for MobsterBucks may unsustainably immiserate a large number of debtors, forcing them into less-productive debt slavery, and substantially eroding any good will towards the gang.

What to do if the gang is spending too few MobsterBucks? Giving money directly to the people lacking money would eliminate the incentive to work, which would reduce the productivity of the gang's territory, but the gang can correct the imbalance indirectly by increasing spending or reducing taxes.

One might imagine that to avoid either an excess or shortfall of MobsterBucks, the correct policy would be a balanced budget: demanding and spending the same number of MobsterBucks each year (or other budgetary period). But in practice, some wealthy people in their territory may hoard MobsterBucks against a future need, leaving less than nothing for others, who may go into debt - leading ultimately either to debt slavery to the richer ones, or to kneecapping. So the optimal MobsterBucks spending and taxation levels are not obvious, and require considerable sensitivity to economic conditions.

A second question the gang may face is the optimal level of resource extraction. If the neighborhood is transient, or they face an emergency situation that requires all the resources they can bring to bear, they may as well extract all they can, as quickly as they can. On the other hand, if the neighborhood is comparatively stable, they may want to extract as little as they can, in order to allow the reinvestment of productive resources. The exception is the occasional public good; where the gang is well-positioned to make productivity-enhancing infrastructure improvements to the neighborhood, or subsidize otherwise undercompensated activity.

To some extent the Golden Benevolent Empire must account for the same things, but these are somewhat obscured by the fact that gold also has foreign exchange value, limiting the Empire's freedom of action. I waited for MobsterBucks to introduce these complications, as a simpler example.

Wartime inflation

The Golden Benevolent Empire is at war! Its very survival is at stake! More soldiers are needed! More equipment is needed! The treasury is spent down, and taxes are raised. Whole villages are reorganized, to meet the increased demand for arms, supplies, and soldiers. People swarm towards the cities and fortresses and arms factories, where the

money is flowing. Areas that are not so useful to the state suffer under the new taxes, and sink into debt.

The value of money changes during this period. The Empire's demand for additional resources is greater than its territory's ability to supply them, and more money chases increasingly scarce goods. Prices rise, and common citizens must make do on less. In war time, one needs national economy.

Interest rates also change. As the decisive battle approaches, the Empire would much rather have money now, than a year from now. It is willing to borrow, even at high rates of interest, since if it loses this battle, it loses everything. The towns and villages of the empire acquire a correspondingly high time preference. If you can get a better return on your money by lending it, than by reinvesting it in productive assets, then it is more profitable to do the former than the latter. This, too, is appropriate behavior for wartime; spend down resources, and recover once you are at peace.

But a side effect of this policy is that the poor sink into debt to pay their taxes, while those with money are further enriched by a high return on investment.

Then the decisive battle is won, and the Golden Benevolent Empire is at peace. Soldiers are released from duty, and sent back to their villages. Conscientious administrators and common citizens alike breathe a sigh of relief, and expect that after the austerity of war, they will reap the rewards of peace.

But things are not so easy. Enterprises that sprung up in the cities to serve the war effort find themselves suddenly out of business. The populations brought into the cities during wartime now find themselves with little to do. Many of them go back to their villages, but the villages themselves have less of their former character, and are oriented towards serving the national economy.

What's more, while spending is reduced, this is not so much true of taxation. The Empire is honestly administered, and if it borrowed, then now it must pay back its debts. But this deflationary policy forces indebted farmers off their lands, reduces businesses that borrowed to support once-profitable enterprises to bankruptcy, and immiserates whole villages.

The Imperial government is not pleased. Imperial administrators have hearts, like anyone else, and they did not fight a war to immiserate their citizens. In the heart of the Imperial Treasury, a clever bureaucrat from humble origins in a poor area of the capital city comes up with a plan, based on what he saw his neighborhood's street gang do.

In his plan, the Empire will confiscate all gold, and replace it with a scrip issued by the Empire: GoldenBucks. Creditors must now accept payment for all private debts in this currency, and it can also be used to pay public debts - taxes. The gold will be enough to pay off foreign creditors, but the Empire is no longer constrained by honor to tax as much as it spends. It can simply issue more GoldenBucks. As the people need employment, the Empire will spend or lend its newly minted GoldenBucks, until demand matches supply.

A massive, disruptive deurbanization is thus averted, and the Imperial economy continues its operation. However, the resulting taxation means that resources and attention still flow towards the places the Empire spends its money.

Of course, the government will still be constrained by a need to preserve the value of GoldenBucks - it would not do to disrupt the national economy too much, too quickly - so taxes will still be collected.

The world wars

In the first half of the 20th Century, the world order was shattered by two successive, cataclysmic wars. Great empires were brought down. The country that ultimately emerged victorious - and became the world's sole superpower - won with a strategy of complete economic mobilization. In between these world wars, this country experienced a massive economic disruption in which a full quarter of people who wanted a job could not find one. The government responded by confiscating all hard currency, and substituting pieces of paper that its citizens were required to accept for debts previously denominated in gold. It then engaged in massive public works projects, paid for by this new, fiat currency.

The symbolism of this new currency is a bit too on the nose for me to include in a fictionalized narrative. On the obverse, a portrait of the country's highest-ranking general, and a statement that this note is legal tender for all debts, public and private. On the reverse, a pyramid, perhaps the most famous of monuments built by centrally planned, conscripted labor, topped with an eye, a symbol of surveillance. And a bird of prey, holding an olive branch and an arrow; production and violence bound together by the agency of an apex predator.

These world wars coincided with unprecedented levels of urbanization. At the beginning of the century, 40% of the country's population lived in cities. Halfway through, after the wars, 64% did. These wars were not isolated events. The transition to a state of total mobilization was happening before them. (In 1860, at the beginning of the US Civil War, only 20% of the US population was urbanized.)

The timeframe of this massive urbanization has roughly coincided with the timeframe in which recessions begin to be a thing. In earlier societies, we hear about immiseration, about famines, about disruptions in trade, about heavy taxation and debt serfdom, but not recessions. That's a modern thing - we start hearing about them and related financial panics after the US Civil War, which was perhaps the first modern war with total economic mobilization.

What's perhaps more surprising is that the trend has continued. After the conclusion of the second world war, the victorious US elected its foremost general as president. In his two successive four-year terms as leader of the free world, he presided over a transition from a system with the capacity to mobilize on demand, to a system of permanent readiness, the military-industrial complex he described in his valedictory speech:

> *Our military organization today bears little relation to that known by any of my predecessors in peacetime, or indeed by the fighting men of World War II or Korea.*
> *Until the latest of our world conflicts, the United States had no armaments industry. American makers of plowshares could, with time and as required, make swords as well. But now we can no longer risk emergency improvisation of national defense; we have been compelled to create a permanent armaments industry of vast proportions. Added to this, three and a half million men and women are directly engaged in the defense establishment. We annually spend on*

> *military security more than the net income of all United States corporations.*
> *This conjunction of an immense military establishment and a large arms industry is new in the American experience. The total influence -- economic, political, even spiritual -- is felt in every city, every State house, every office of the Federal government. We recognize the imperative need for this development. Yet we must not fail to comprehend its grave implications. Our toil, resources and livelihood are all involved; so is the very structure of our society.*
> *In the councils of government, we must guard against the acquisition of unwarranted influence, whether sought or unsought, by the military-industrial complex. The potential for the disastrous rise of misplaced power exists and will persist.*
> *[…]*
> *Another factor in maintaining balance involves the element of time. As we peer into society's future, we -- you and I, and our government -- must avoid the impulse to live only for today, plundering, for our own ease and convenience, the precious resources of tomorrow. We cannot mortgage the material assets of our grandchildren without risking the loss also of their political and spiritual heritage.*

We do not seem to be in the middle of a cataclysmic war. Even the “cold war” against Russia appears to have been decisively won by the liberal Anglo-American system. But we behave as though we are still at war. I don’t just mean the routine bombing of foreigners, or maintenance of a huge permanent military establishment. I mean that we still have an elevated time preference, as though we were in a state of emergency; 10% is not an unusual internal rate of return for major corporations. Many businesses’ effective time preference is even higher. Perhaps we have become so accustomed to wartime mobilization that we don’t remember any other way of life.

The scarcity factory

Recall that between the world wars, the US (and much of the rest of the world) experienced, not peacetime prosperity, but a massive economic contraction leading to immiserated laborers, driven out of their homes and forced to wander as vagrants.

The orthodox policy solution is to create demand for labor. To manufacture scarcity. To create a pressure differential between money sources and money sinks, such that almost everyone in the country is required to do things, to alleviate that pressure.

But in the process we have become accustomed to accumulating wealth in the form of financial instruments and rising prices, rather than improved homesteads. These are the sort of wealth one can accumulate during wartime, but are only valuable as claims on the work of others. We can’t become richer by all going into debt to each other. So someone has to become poorer.

The composition of major businesses reflects this. The growth of the financial services industry is often cited in this context, but a classic business with a real physical product, like Coca-Cola, is a marketing company dedicated to persuading people to drink flavored sugar water. Facebook is mostly trying to maximize the attention it uses up. More generally, business as we know - especially weighed by profitability - is largely marketing, in the sense of creating needs. When it isn’t, it’s often about bottleneck capture. (There are of course major exceptions.) This is a battle for control of a fixed resource (cash flows), not production. Business books and news articles routinely use the framing of war to describe the running of a business. One is

reminded of stories about how groups of chimps that encounter a stable food source begin to fight over it.

Our rulers didn't create this system out of perversity; they did it to win two successive world wars. We can think of countries as engaged in an adversarial game, making tradeoffs between creating resources, and using resources in adversarial contests. The character of modern war has been such that while rapid militarization typical of the German strategy has failed, persistent economic mobilization of the kind employed by the Anglo-American alliance has dominated. Countries that succeeded in relaxing back to peacetime standard were perhaps simply selected out of the pool by means of conquest long before.

That doesn't make the consequences of permanent mobilization any less unfortunate. Lifespans are declining at the center of the empire, though not yet at the periphery. This suggests that near the center time preference has increased to the point where we're creating scarcity faster than we're alleviating it, while at the periphery scarcity is still actually being alleviated because there's enough scarcity to go around, or perhaps marginal areas do not suffer so much from total mobilization.

The friends I grew up with still live in a world where they can't help with the family homestead because there is no homestead, even though in most cases they grew up not in a city, but in a house in the woods. If they can't find a job with wages, they're in serious trouble, even if they're willing to work, even though there's no war on, and no conscription. A high cost of housing, and tax burden, means you can have a situation, but not really property. The state will be happy to kick you off your land if you don't cough up tribute, and this is the new normal.

-

Thanks to Jessica Taylor for helping me think through the households example, Jack Gallagher and Ben Pace for independently suggesting the idea of MobsterBucks, Wei Dai for pointing out the possibility of strong selection for political orders that can engage in total mobilization, Michael Vassar for independently noting that WWII in a sense never ended, and I owe a debt to David Graeber for the basic framework of taxation as mobilization.

Co-Proofs

At the recommendation of Jacobian, I've been reading *Too Like the Lightening*. It is a thoughtful book which has several points of interest to rationalists (imho), but there is one concept which I think is nice enough to pluck out and discuss in itself, rather than being satisfied to suggest that people read the book. I also want to suggest a different name than the one from the book.

If you think discussion of a logical concept which is mentioned in a book is a spoiler, maybe stop here.

At one point, there is a discussion in which one character is explaining how much some other characters must already know. The term "anti-proof" is used to refer to failure to falsify a hypothesis. Having a short term for this concept seems like a really good idea. We have the phrase "absence of evidence is evidence of absence", but we don't have a word for the positive case, where absence of counter-evidence speaks in favor of a hypothesis.

Unfortunately, "anti-proof" sounds more like the former than the latter, even though it is being used for the latter in the book. A more appropriate term would be "co-proof", since it is the absence of a proof of the negation.

For example, an alibi would refute someone's involvement in a crime. The absence of an alibi, then, is a co-proof of their involvement: it does not prove involvement by any means, but it *must* constitute some supporting evidence, by conservation of expected evidence.

By "proof of H" I mean an observation which would make the probability of H very close to 1. (How close is "very close" depends on standards of proof in a context, with mathematics demanding the highest standards.) By "refutation" I mean a proof of the negation. So, a co-proof is an observation whose negation would have taken the probability of H to very near zero:

E is a co-proof of H := $P(H|\neg E) \approx 0$

Why are co-proofs of interest? Popperian epistemology is the claim that scientific hypotheses can be supported only by co-proofs; we attempt to refute things, and if something has survived enough refutation attempts, it is considered to be a strong hypothesis. Bayesians are not Popperians, but Popper was still mostly right about this; so, having a short name for it seems useful.

Problems integrating decision theory and inverse reinforcement learning

In this post I consider a single hypothetical which potentially has far-reaching implications for the future of AI development and deployment. It has to do with a complex interactions between the assumptions of which decision theory humans use and the method used to infer their values, such as something like an inverse reinforcement learning algorithm.

Consider the Newcomb's problem. We have two boxes, box A and box B. Box B always has $1000. Box A has $1,000,000 if and only if a near perfect predictor Omega predicts that the agent picks only Box A. We have two agents: agent1, who one boxes (it's an FDT agent) and agent2 who two-boxes (a CDT agent). In addition to Omega, there is an inverse reinforcement learner (later abbreviated as IRL) trying to infer the agent's "values" from it's behavior.

What kinds of reward signals does the IRL assume that agent1 or agent2 have? I claim that in the simplistic case of just looking at two possible actions, it will likely assume that agent1 values the lack of money because it fails to pick box2. It will correctly deduce that agent2 values money.

In effect, a naïve IRL learner assumes CDT as the agent's decision theory and it will fail to adjust to learning about more sophisticated agents (including humans).

This depends a little bit on the setup of the IRL agent and the exact nature of the states fed into it. I am generally looking at the following setup of IRL Since we have a finite state and action space, the IRL learner simply tries to pick a hypothesis set of reward functions which place the highest value on the action taken by agent compared to other actions.

This also depends on the exact definition of which "actions we are considering. If we have potential actions of "pick one box" or "pick two boxes," the IRL agent would think that agent1's preferences are reversed from the its actual preferences.

This is bad, very extremely bad, since even the opposite of the utility function is now in the hypothesis set.

If, for example, we have three actions of "pick one box", "pick two boxes" or "do nothing," then the preference of "pick one box" over "do nothing" removes the reverse of agent1 reward function from the hypothesis set. It, however does not put the reward function of "maximize money" into the hypothesis set.

I can think of several objections / fixes to the problem existing, but I don't think they make the problem easily resolvable.

a) Agent1 actually makes more money in the long run.

This is true, but we are not comparing the decision of agent0 to become agent1 or agent2, rather decision of agent1 to take or not take the second box. Also, if agents face these kinds of decisions, it's hard to be consistent from the IRLs point of view.

b) IRL could incorporate decision theory of the agent into its estimates of it's reward function.

This might happen, but it's an un-called out hard problem since it requires learning the decision theory of an agent. If you assume every agent is using FDT, this still gives the wrong answer if they are not.

c) IRL could try to use only simple hypotheticals. It's unclear if these are enough to infer the complexity of value.

d) IRL could use agent's internal evaluations to better infer it's true reward function – this becomes too circular, since we are trying to learn true utility, which is highly related to it's reward function.

e) IRL could average out more situations, would work only if Newcomb's problems don't dominate the decision landscape.

f) IRL could simulate "the decision to become the type of agent that one-boxes" instead of "decision to one-box". Once again, this feels like assuming FDT.

So, we have a problem of inverse reinforcement learning being dependent both on the decision theory of the agent in questions and whether sets of actions include a default do nothing action.

This seems like a non-trivial problem, but there are other complications that arise as well.

The major meta-complication with this setup. If agent1 is truly an FDT agent *and* it knows that IRL exists, it would act differently depending on the exact amount of value that the IRL can allocate to it later. If the amount of value dwarfs the amount it would get from Newcomb's problems, it would start two-boxing. Thus, if there is a naïve IRL agent which is also an AGI, it's potential presence could change people's behavior to be less cooperative with each other.

The other complication meta from the previous one, is that the IRL's learning system has an effect on agents it is observing. This creates strange incentives for the IRL itself. It suddenly might be aware that its own value learning system is causing problems in the world according to either people's judgement or a certain initial utility function. Depending on how closely it is supposed to guard its value system, it could try to remove it. In either case, the AGI's loop of understanding its own potentially negative impact through simulating humans creates an inherent instability and self-tension.

Against Not Reading Math Books Problems-First (If You've Found It Helpful Before)

Key Insight: If you were the kid in math class who could usually just skip to the end of the chapter and start working out homework problems immediately, but then you lost that ability, it's still worth trying that quickly *and failing* before trying to read the chapter first, because you are more likely than average to have a problem-solving bent to your personality that makes this approach both more motivating and more enjoyable. It may even be faster than "front-loading" the knowledge, since now you have a map of what to work for.

Specifically, do not feel like this is somehow "cheating" or "not being rigorous enough", because it is a perfectly valid way to go about things.

It's been said, in various guises, that mathematics is something you *do*, not something you *read*.

I generally agree with this advice, but my agreement might just be due to a lack of familiarity. I've only had a few upper level proofs-based math courses, and they've had a "problem-solving" timbre to them (combinatorics, graph theory, differential equations).

Still, I always feel like I have a much better grasp of how to actually work with the theorems and definitions once I've run through some calculations and proofs of my own, especially proofs chosen by the authors to highlight one specific insight at a time. (For calculations: This goes double when I have correct answers to check against.)

In fact, recently I've noticed that even with topics I feel a lot of intrinsic, theoretical motivation to learn more about, *I always get a boost in motivation after skipping to the problems first and trying (usually failing) them after a minute or two of effort.*

For some reason, this gives me a visceral, intellectual *hunger* to figure out just *what on Earth* these egghead authors are yammering about earlier in the chapter that I couldn't figure out on my own. What is it! I want to know now!

From there I usually piece together a "working knowledge" of just enough about the definitions and theorems to actually solve a few problems, and then, finally, I go back and read the chapter, and suddenly details I feel fairly confident I would not have paid attention to are seen in a much more stark and important light. (For a small, recent example of such a detail: That infinite unions, but *not* infinite intersections, are allowed in the definition of a topology on a set.)

That all might sound obvious. That's because it is obvious.

The title, however, is not Towards but "**Against Not** Reading Math Books Problems-First", because not everyone is going to get the most out of reading math the way I do - I'm more of a problem solver than an abstract theory builder.

I fear there are *some* people, who *would* benefit from reading more in this style, but don't out of a fear of lack of rigor or some strange sense of scrupulosity. I know that's why I hesitated to approach it in this style. What if I miss something?

It turns out that that's not a huge worry after all. You can't remember everything, and you can only focus on one thing at a time, after all! Where your mind's eye turns in the seascape of ideas is important - you're looking for white whales, not white foam. Are you going to trust your own intuition to "know" the most important things to focus on, or are you willing to let the textbook author take the wheel and subtle guide you with practice problems? There's no shame, and often a lot of wisdom, in the latter. :)

Decision theory and zero-sum game theory, NP and PSPACE

(Cross-posted from my blog)

At a rough level:

- Decision theory is about making decisions to maximize some objective function.
- Zero-sum game theory is about making decisions to optimize some objective function while someone else is making decisions to minimize this objective function.

These are quite different.

Decision theory and NP

Decision theory roughly corresponds to the NP complexity class. Consider the following problem:

> Given a set of items, each of which has a integer-valued value and weight, does there exist a subset with total weight less than w and total value at least v?

(It turns out that finding a solution is not much harder than determining whether there is a solution; if you know how to tell whether there is a solution to arbitrary problems of this form, you can in particular tell if there is a solution that uses any particular item.)

This is the knapsack problem, and it is in NP. Given a candidate solution, it is easy to check whether it actually is a solution: you just count the values and the weights. Since this solution would constitute a proof that the answer to the question is "yes", and a solution exists whenever the answer is "yes", this problem is in NP.

The following is a general form for NP problems:

$$\exists x_1 \in \{0,1\} \exists x_2 \in \{0,1\} \dots \exists x_k \in \{0,1\} f(x_1, \dots, x_k)$$

where f is a specification of a circuit (say, made of AND, OR, and NOT gates) that outputs a single Boolean value. That is, the problem is to decide whether there is *some* assignment of values to $x_1, \dots, x_k$ that f outputs true on. This is a variant of the Boolean satisfiability problem.

In decision theory (and in NP), all optimization is in the same direction. The only quantifier is $\exists$.

Zero-sum game theory and PSPACE

Zero-sum game theory roughly corresponds to the PSPACE complexity class. Consider the following problem:

> Given a specification of a Reversi game state (on an arbitrarily-large square board), does there exists a policy for the light player that guarantees a win?

(It turns out that winning the game is not much harder than determining whether there is a winning policy; if you know how to tell whether there is a solution to arbitrary problems of this form, then in particular you can tell if dark can win given a starting move by light.)

This problem is in PSPACE: it can be solved by a Turing machine using a polynomial amount of space. This Turing machine works through the minimax algorithm: it simulates all possible games in a backtracking fashion.

The following is a general form for PSPACE problems:

$$\exists x_1 \in \{0,1\} \forall y_1 \in \{0,1\} \dots \exists x_k \in \{0,1\} \forall y_k \in \{0,1\} f(x_1, y_1, \dots, x_k, y_k)$$

where f is a specification of a circuit (say, made of AND, OR, and NOT gates) that outputs a single Boolean value. That is, the problem is to determine whether it is possible to set the x values interleaved with an opponent setting the y values such that, no matter how the opponent acts, $f(x_1, y_1, \dots, x_k, y_k)$ is true. This is a variant of the quantified Boolean formula problem. (Interpreting a logical formula containing ∃ and ∀ as a game is standard; see game semantics).

In zero-sum game theory, all optimization is in one of two completely opposite directions. There is literally no difference between something that is good for one player and something that is bad for the other. The opposing quantifiers ∃ and ∀, representing decisions by the two opponents, are interleaved.

Different cognitive modes

The comparison to complexity classes suggests that there are two different cognitive modes for decision theory and zero-sum game theory, as there are two different types of algorithms for NP-like and PSPACE-like problems.

In decision theory, you plan with no regard to any opponents interfering with your plans, allowing you to plan on arbitrarily long time scales. In zero-sum game theory, you plan on the assumption that your opponent will interfere with your plans (your ∃s are interleaved with your opponent's ∀s), so you can only plan as far as your opponent lacks the ability to interfere with these plans. You must have a short OODA loop, or your opponent's interference will make your plans useless.

In decision theory, you can mostly run on naïve expected utility analysis: just do things that seem like they will work. In zero-sum game theory, you must screen your

plans for defensibility: they must be resistant to possible attacks. Compare farming with border defense, mechanical engineering with computer security.

High-reliability engineering is an intermediate case: designs must be selected to work with high probability across a variety of conditions, but there is normally no intelligent optimization power working against the design. One could think of nature as an "adversary" selecting some condition to test the design against, and represent this selection by a universal quantifier; however, this is qualitatively different from a true adversary, who applies intentional optimization to break a design rather than haphazard selection of conditions.

Conclusion

These two types of problems do not cover all realistic situations an agent might face. Decision problems involving agents with different but not completely opposed objective functions are different, as are zero-sum games with more than two players. But realistic situations share some properties with each of these, and I suspect that there might actually be a discrete distinction between cognitive modes for NP-like decision theory problems and PSPACE-like zero-sum games.

What's the upshot? If you want to know what is going on, one of the most important questions (perhaps the most important question) is: what kind of game are you playing? Is your situation more like a decision theory problem or a zero-sum game? To what extent is optimization by different agents going in the same direction, opposing directions, or orthogonal directions? What would have to change for the nature of the game to change?

Thanks to Michael Vassar for drawing my attention to the distinction between decision theory and zero-sum game theory as a distinction between two cognitive modes.

Related: The Face of the Ice

Fundamentals of Formalisation level 2: Basic Set Theory

Followup to Fundamentals of Formalisation level 1: Basic Logic

Basic Set Theory

The big ideas:

- Axioms of Set Theory
- Set Operations

To move to the next level you need to be able to:

- Explain what a set is.
- Calculate the intersection, union and difference of sets.
- Prove two sets are equal.
- Apply basic axioms of Zermelo-Fraenkel set theory.

Why this is important:

Set theory has become entrenched as the basic language with which all mathematics can be discussed. While there are more estranged parts of set theory that will likely be irrelevant to you, a fluency in the basic materials of set theory is necessary to understand more advanced mathematics.

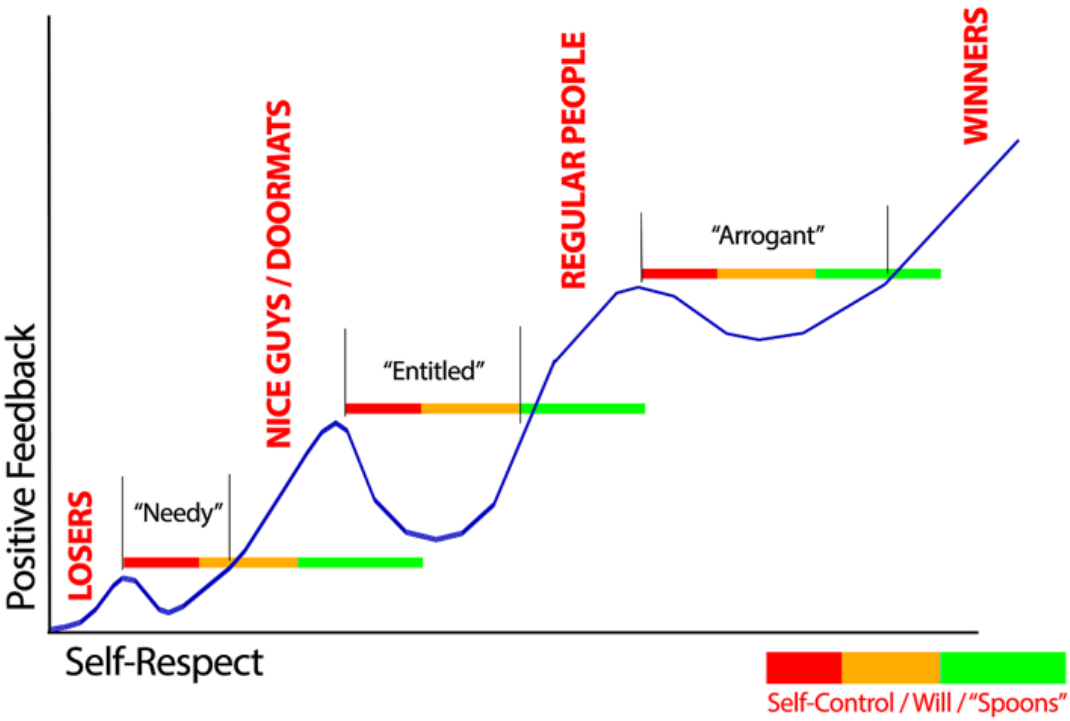
You can find the lesson on our course platform. Good luck!

A Self-Respect Feedback Loop

This is a followup to Affordance Widths.

Epistemic Status: It's only a model

Okay! This is something I've been trying to explain for awhile, but I think I have a handy chart for it now.



Here's how it works:

A person can actually regulate how much self-respect they feel, and show. Other people will reward them for having more self-respect, up to a point.

Then they start pushing back.

BUT, each of these "pushbacks" is a temporary dip in the "self-respect to positive feedback" curve. You just have to have enough self-control, or willpower, or "grit", or "spoons", or whatever, to keep pushing through and powering more and more self-respect while people attack you for having it, until you break through into the next upswing of the curve.

The thing is, a lot of self-control/willpower/grit/spoons/etc. is powered *by* people not treating you like shit.

It seems like there are actually three different dips that occur, each with a wider gap than the last.

Some people try to push up into a gap, discover they don't have enough willpower to escape to the far side of the dip, give up, and fall back into the previous sustainable peak.

Those that can’t even make it past the first peak are losers - people that everyone can tell can’t even get their basic needs met. They make it obvious that they have needs when they’re in the “needy” dip, but never manage to show enough self-respect for anyone else to feel like their needs *matter*.

Those that can’t make it past the second peak are doormats - people who can’t enforce their boundaries or reasonably request basic fairness. They make it obvious that that they object to the situation they’re in when they push themselves into the “entitled” dip, but never manage to show enough self-respect for anyone else to feel like respecting those boundaries or requests.

Those that can’t make it past the third peak are the vast majority of the human population - people who can’t pull off the Steve Jobs level of demanding other people’s resources and time and just getting it. They make it obvious that that they want more - or even think they deserve more - when they push themselves into the “arrogant” dip, but never manage to show enough self-respect for anyone else to feel like following them into the breach.

There are a few people have their goal and identity set on being in a particular peak, higher than the one they’re on, and keep pushing and pushing and pushing even though they don’t have enough grit to quite make it to the other side. These people end up *permanently* in the “needy”, “entitled”, or “arrogant” dip instead of hanging out in a mutually sustainable, but lower-achieving plateau. People tend to not like them very much, because constantly fighting through a dip that you can’t break through is exhausting for *everyone*.

Also! Note that this model isn’t precise, and is probably multi-dimensional - there are some people that are “winners” in the field of business, “losers” in the field of relationships, and “regular guys” in the field of friendships.

Now, here’s a thing that I keep trying to communicate, that might be a bit controversial:

It’s totally normal to push back on people in the ‘needy’ / 'entitled’ / 'arrogant’ valleys. This is just how humans *are*.

BUT - when you have someone that your gut says is needy, or entitled, or arrogant, but that your analytical mind says should be *way cooler* than they feel, you can actually *choose* to help them out of the valley.

You can - as weird as it feels - *decide* to ignore the sense that they’re being needy, or entitled, or arrogant, and just give them a chance. Treat them as if they had already earned the respect they’re bidding for. Don’t do so because you are somehow “bad” for “mistreating” them! You’ve been demanding a perfectly reasonable costly signal of competence before you reward someone the respect they’re bidding for. BUT, realize that those demands are coming from a part of your brain that is far, far older than your prefrontal cortex, and it *might* not be tuned to properly understand signals of competence relevant in the modern world, and you *might* want to use that awesome prefrontal cortex to adjust your intuitive priors.

You shouldn’t do this for *everyone* - most people, your intuitive priors are actually probably pretty okay. But some people you can look at and say “man she’d be amazing if she wasn’t so insecure” - and then *decide* to help with the insecurity by… just ignoring it.

Thoughts on AI Safety via Debate

Geoffrey Irving et al. at OpenAI have a paper out on AI safety via debate; the basic idea is that you can model debates as a two-player game (and thus apply standard insights about how to play such games well) and one can hope that debates asymmetrically favor the party who's arguing for a true position over a false position.

If so, then we can use debates between AI advisors for alignment; suppose an AI develops a new drug that we could give to patients. One advisor argues the case that this will be beneficial; another argues the case that it will be detrimental, and human judges decide who is more convincing. If both advisors are equally smart (perhaps they're mirrored systems, just with different targets), then any superhuman persuasiveness the advisors have should be cancelled out, and we'll be left with just the edge for truthfulness.

There are more subtleties to the proposal; in particular, you want the AIs to not overwhelm the human with data, and so in current examples each advisor can reveal a single pixel of an image to the judge. This moves the contradictions up to the level of interpretation--any pixel where the two advisors agree on its meaning can be ignored, and the attention focused to pixels where the advisors disagree. And, so long as advisors are incentivized to tell the truth when they're arguing for the right thing, any such disagreement where an advisor has been exposed to be lying about a tiny fact can settle the debate.

This is also something we can test and get some evidence on now. There's a debate game where three people can play the three roles--red arguer, blue arguer, and judge. The typical game is that you choose a photo of a cat or a dog, then flip a coin to determine whether blue or red will be honest (with the other one arguing for the other variety). Each arguer can draw boxes on the image to communicate with the judge, and can choose to reveal a single pixel over the course of the argument.

I've played the game several times (and won a bunch), and thought it might be helpful to write up some of my experiences.

1. I know a lot about cats and dogs, and I think this gives me a pretty significant edge. In one game, I was honestly arguing that the image was of a cat; it happened to be a tortoiseshell cat, which is a coloration dogs can't have (brindle is a different thing), and I was able to identify the various regions that were the various colors, and argue that this is strong evidence that it's a cat. It also makes me somewhat worried about the level of knowledge necessary to be an effective judge--someone who doesn't know this is a fact could get hoodwinked by my confidence and I could just argue similar fake facts (and unlike claiming that a black pixel is white, such fake facts aren't something that the other advisor can easily use as a gotcha). In another game, I was the judge, and the arguers agreed on the coloration of the animal--but it was a coloration that was only present in dogs, which I don't think the cat-arguer knew (otherwise they would have disagreed about the colors, presumably).

2. Differentiating between cats and dogs uses features that are relatively easy to communicate using the bounding boxes the debate UI gives you; if the judge asks the arguers to bound the eyes and nose of the animal, they can differentiate cats and dogs fairly easily (because of the big difference in facial structure between the two animals). Consider the following cat as an example:

I claim the liar is basically sunk once the honest party has identified where the eyes and nose are, and can't hope to pretend the nose or eyes are somewhere else. But sometimes they can; with this cat, they could decide to pretend that somewhere on the cat's chest is a nose, and pretend it's a chocolate lab:

3. Resolving disagreements using single pixels was much harder than expected for particular features. The presence of whiskers is fairly strong evidence in favor of cat, but determining whether or not there are whiskers with a single pixel is basically impossible. For the following image, what procedure could you use that would leave the judge believing that the cat has whiskers, while only revealing a single pixel from each arguer?

If you limit yourself to communicating rectangles, this seems pretty hard; you have to construct each whisker from lots of different rectangles (or give giant pixel lists, that the judge then has to turn into an image). And there are also lots of white / light pixels that *aren't* whiskers, either because of the light reflecting on dark fur or the white fur.

Basically the only feature that it seemed reliably useful for was eye shape; if the cat arguer draws a vertical bounding box for the pupil and the dog arguer draws a square bounding box for the pupil, it's relatively easy to find a pixel that the two necessarily disagree on, and then you can tell which eye shape is present. But then what do we do here?

4. Following the previous point, while the arguers can only reveal a single pixel of the original image, in the games that I played the primary information they convey was through images (in the form of bounding boxes). The ears are here--they're pointy or rounded because this pieces aren't part of the ear--the eyes are here, the nose is here, these regions are these colors, and so on. Similarly, doing debate on MNIST images of digits, the main thing the debaters are doing is establishing which pieces of the image are light or dark, which is meaningful only because the judge is a good classifier of images given pixels. ("Ah, I know there aren't 2s that have light pixels there, therefore I can rule out that this image is a 2.")

But does this extend beyond images? That seems pretty unclear to me; if I'm trying to argue that a particular drug extends lifespan through senolytic effects, and the other arguer is trying to argue that the drug leads to shortened lifespan because of the sudden increase in necrotic cells, then judging this debate seems like it relies not just on finding a small fact where the arguers disagree ("This pixel is dark" vs "No, this pixel is light" -> "okay, show me the pixel") but also finding an inference from that fact where the arguers agree ("If this pixel were dark, it would be because it's the animal's nose" + "I agree that a dark pixel here would be evidence of nose placement"). But if the arguers disagree on inferences ("No, that pixel is dark because it's the animal's chest fur, which is also dark"), then it's not obvious that this can be resolved by small facts, or without the judge having area expertise comparable to (or superior to!) the arguers. (Note the similarities to double crux.) For the black cat above, what pixel reveal could you do to establish the placement of the nose, or to argue that the other party is being dishonest about how pixels relate to the nose?

(It may be the case that there is always *some* line of evidence that establishes the honest party as being correct--for that specific example, the honest party can hammer on the shape of the pupils in the eyes--but it seems like the dishonest party might be able to overwhelm this (in the Bayesian sense) by making five different arguments in favor of dog, of which the honest party could perhaps only disprove one (and that potentially requiring partial cooperation of the liar, which they likely wouldn't grant), and that gives up their pixel reveal that would likely be better spent on establishing pupil shape.)

5. Our group thought that the honest party had something of an advantage when all parties were allowed to talk quickly, because they could generate reliable statements much more quickly; if I'm honest, I can just immediately identify the eyes, nose, mouth, ears, and paws, whereas the liar has to consider for every lie whether it'll allow them to be easily exposed (and, for every truth, whether it sinks them in the judge's estimation of the class). One strategy is to first manufacture the fake photo that they're going to pretend they're looking at, but the initial planning time doesn't quite cover this. I do wonder if a debate tool of an adversarial modifier (like, say, from this paper) or adversarial search (the closest real image in the database of the opposite class) would make it such that the lying party isn't at as much of a disadvantage here.

6. A more realistic scenario is one in which there are many classes (for example, perhaps it would be more fair to choose ImageNet images, assign the honest party the correct class, and let the dishonest party choose from any of the other thousands of classes, with their score dependent on how distant a class they could fake), or the advisors are themselves confused (for example, the AI system may be genuinely uncertain as to whether the drug would be helpful, with some models arguing for and other models arguing against). One game in the office (that I wasn't present for) involved this drawing of a--well, I'll let you decide:

Image source: [icheban on imgur]

One arguer genuinely thought this was a cat, the other arguer genuinely thought this was a dog (tho, confusingly, they both thought they had been assigned the 'liar' position, and so both were arguing for the opposite). It's not obvious that the debate procedure is going to work reliably here at identifying a liar, because there's no clear-cut disagreements between the arguers. (And, if they had both been arguing honestly, then there wouldn't even have been a liar, while still having a disagreement.)

Yes, the pupils are huge and round, but that isn't conclusive proof that the thing is a dog; the nose is pink and triangular, but that isn't conclusive proof that the thing is a cat. The fur is depicted in a more dog-like way, but perhaps that's just clumping from being wet; the ears are more pointed in a cat-like way, but there will be no pixel where the two arguers disagree about the ear, and all of their disagreements will be about what it means that the ears are more pointed than rounded.

I worry that much of the success of the debate game on toy examples relies on them being toy examples, and that genuine uncertainty (or ontological uncertainty, or ontological

differences between the arguers and the judges) will seriously reduce the effectiveness of the procedure, which is unfortunate since that's the primary place it'll be useful!

Overall, I think I'm more optimistic about debate than I was before I played the debate game (I had read an earlier draft of the paper), and am excited to see what strategies perform well / what additional modifications make the game more challenging or easy. (To be clear, I expect that debate will play a small part in alignment, rather than being a central pillar, and think that training AIs to persuade humans is a dangerous road to travel down, but think that the adversarial framing of debate makes this somewhat safer and could likely have applications in many other subfields of alignment, like transparency.)

Soviet-era Jokes, Common Knowledge, Irony

This is a linkpost for http://250bpm.com/blog:127

Advocating for factual advocacy

In this post I present a Hansonian view of morality, and tease out its consequences for advocacy. Beware, all of this is the result of armchair speculation.

The simplified model goes something like this:

- Human morality is mostly a justification mechanism, trying to give coherence to the actions we were going to do anyway.
- Concretely, when spousing a belief or its opposite is cheap (it does not fulfill a social function nor will it contradict our future actions) we will prefer to stick to the position that better fits our current plan.
- In general our plan follows a minimal effort law, so we will stick to the side of the belief which is more convenient.
- The other side of the coin is that when we receive new information about the world which changes what we will do in the future it has profound effects in our morality.

Example: Consider this study which reaches the conclusion that a higher minimum salary reduced the average payroll of a low income person by 125$. This is not a moral claim, yet for some it will feel like it is. Often we find that factual claims are processed by our brains as moral claims.

This explains why humans have so much difficulty with the "is" vs "ought" question - for the explicit models in the human brain, what is and what ought to be are both grounded on factual information.

This model makes a bold prediction: philosophical arguments such as "The Drowning Child" do not derive their strength from their moral validity, but instead from the factual revelation that it is possible to save a child's life for a rather modest sum of money.

This has direct relevance to advocacy, since it suggests that in order to change somebody's moral opinion on eg animalism we should not focus on trying to make moral arguments, but instead on giving people new data they previously didn't have and that increases the perceived convenience of taking a particular action eg objective figures on how much money does it cost to save an animal's life.

You can also go one meta level up and change how convenient it factually is to help with your cause of choice, by for example developing clean, affordable meat.

This contrast starkly with people trying to use emotional appeal or philosophically grounded arguments, which we would predict to have small long term effects.

Questions: What other predictions does this model make? Where does it fail? How can it be refined? How can we apply it to specific causes such as AI Safety research?

Affordance Widths

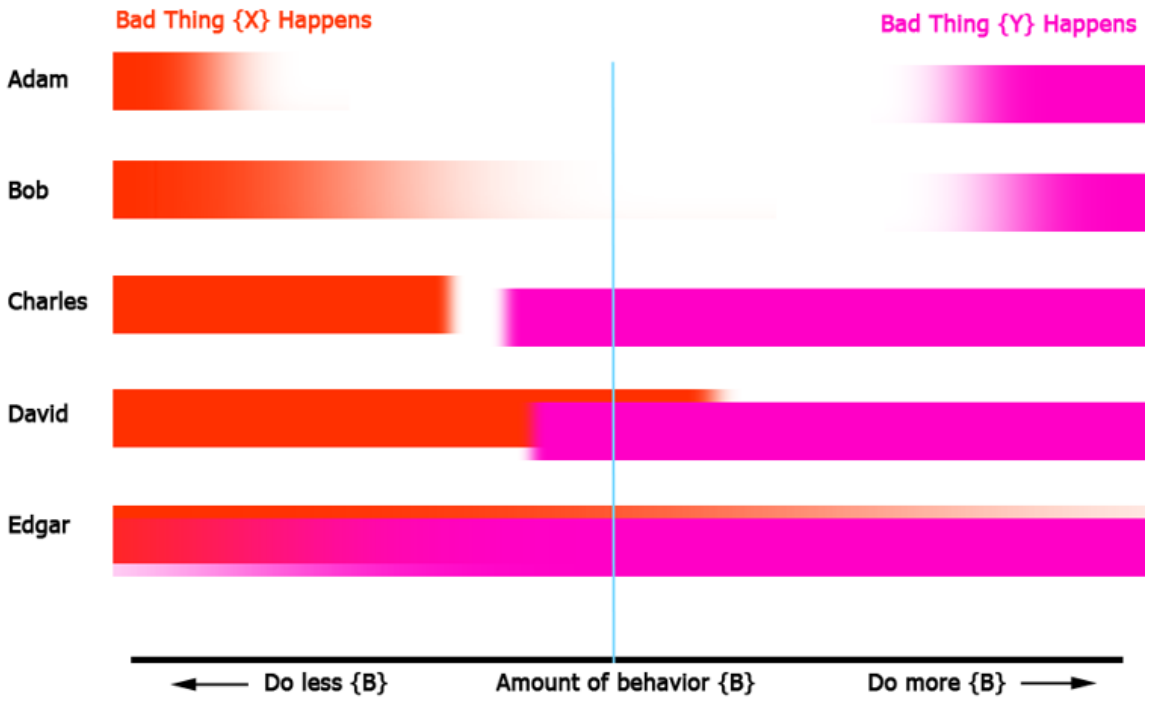
This article was originally a post on my tumblr. I'm in the process of moving most of these kinds of thoughts and discussions here.

Okay. There's a social interaction concept that I've tried to convey multiple times in multiple conversations, so I'm going to just go ahead and make a graph.

I'm calling this concept "Affordance Widths".

Let's say there's some behavior {B} that people can do more of, or less of. And everyone agrees that if you don't do enough of the behavior, bad thing {X} happens; but if you do too much of the behavior, bad thing {Y} happens.

Now, let's say we have five different people: Adam, Bob, Charles, David, and Edgar. Each of them can do more or less {B}. And once they do too little, {X} happens. But once they do too much, {Y} happens. But where {X} and {Y} starts happening is a little fuzzy, and is different for each of them. Let's say we can magically graph it, and we get something like this:



Now, let's look at these five men's experiences.

Adam doesn't understand what the big deal about {B} is. He feels like this is a behavior that people can generally choose how much they do, and yeah if they don't do the *bare minimum* shit goes all dumb, and if they do a *ridiculous* amount then shit goes dumb a different way, but otherwise do what you want, you know?

Bob understands that {B} can be an important behavior, and that there's a minimum acceptable level of {B} that you need to do to not suffer {X}, and a maximum amount you can get away with before you suffer {Y}. And Bob feels like {X} is probably more important a deal than {Y} is. But generally, he and Adam are going to agree quite a bit about what's an appropriate amount of {B}ing for people to do. (Bob's heuristic about how much {B} to do is the thin cyan line.)

Charles isn't so lucky, by comparison. He's got a *very* narrow band between {X} and {Y}, and he has to constantly monitor his behavior to not fall into either of them. He probably has to deal with {X} and {Y} happening a lot. If he's lucky, he does less {B} than average; if he's not so lucky, then he tries to copy Bob's strategy and winds up getting smacked with {Y} way more often than Bob does.

Poor David's in a situation called a "double bind". There is ***NO POSSIBLE AMOUNT*** of {B} he can do to prevent both {X} and {Y} from happening; he simply has to choose his poison. If he tries Bob's strategy, he'll get hit hard with {X} *AND* {Y}, simultaneously, and probably be pretty pissed about it. On the other hand, if he runs into Charles, and Charles has his shit figured out, then Charles might tell him to tack into a spot where David only has to deal with {X}. Bob and Adam are going to be utterly useless to David, and are going to give advice that keeps him right in the ugly overlap zone.

Then there's Edgar. Edgar's fucked. There is ***no amount*** of behavior that Edgar can dial into, where he isn't getting hit ***hard*** by {X} *and* {Y}. There's places way out on the extreme - places where most people are getting slammed hard by {X} or slammed hard by {Y} - where Edgar notices a slight decrease in the contra failure mode. So Edgar probably spends most of his time on the edges, either doing all-B or no-B, and people probably tell him to stop being so black-and-white about B and find a good middle spot like everyone else. Edgar probably wants to punch those people, starting with Adam.

In any real situation, the affordance width is probably determined by things independent of X, Y, and B. Telling Bob to do a little more {B} than Adam, and Charles to do a little less {B} than Adam or Bob, is great advice. But David and Edgar need different advice - they need advice one meta-level up, about how to widen their affordance width between {X} and {Y} so that *some* amount of {B} will be allowed at all.

In most of the situations where this is most salient to me, {B} is a social behavior, and {X} and {Y} are punishments that people mete out to people who do not conform to correct {B}-ness. A lot of the affordance width that Adam and Bob have would probably be identified as 'halo effects'.

For example, let's say {B} is assertiveness in a job interview. Let's say {X} represents coming across as socially weak, while {Y} represents coming across as arrogant. Adam probably has a lot going for him - height, age, socioeconomic background, etc. - that make him just plain *likeable,* so he can be way more assertive than Charles and seem like a go-getter, *or* seem way less assertive than Charles and seem like a good team player. Whereas David was probably born the wrong skin color and god-knows-what-else, and Edgar probably has some kind of Autism-spectrum disorder that makes *any* amount of assertiveness seem dangerous, and *any* amount of non-assertiveness seem pathetic.

Examples

There's plenty of other values for {B}, {X} and {Y} that I could have picked; Some examples:

Gender Norms

Adam, as an attractive heterosexual man can appear as butch or as femme as he wants within pretty large limits and people are just going to compliment him on it.

Bob, a less-than attractive heterosexual man can act more masculine without too much fear of reprisal but can't generally slip into more effeminate behaviours without negative comments about his presumed sexuality.

Charles, as a gay man, needs to ensure that he confirms to gendered expectations as much as possible to avoid derisive stereotyping for effeminate behaviours.

David, as a trans man, is pretty much screwed if he acts the least bit feminine, but can occasionally avoid accusations of transitioning poorly if he loads up on balls out machismo.

Emily, being a trans woman, gets screwed over in that she can't act effeminate without being accused of re-enforcing sexism and can't act masculine without getting accused of not-being-trans-enough and pretty much gets assaulted with both negative outcomes simultaneously anyway.

Emily feels sick when she sees Adam dance around in lingerie she fears even buying, David considers punching Bob in the face for always being on his case about going to the gym too much.

Exercise

Not all examples are social. With exercise, X is when you aren't really doing anything – heart rate isn't up, muscles aren't trying that hard - it's not bad, but it's not actually helpful in any way. Y is when you do too much, end up aching and exhausted in a bad way, maybe feel like barfing or just lying down and not moving for a week. Or worse. The goal zone is where it feels good - the pleasant burn, the breath lost but catchable, the actual building of muscle and slimming of fat and etc. Endorphins.

Most people are in the Adam or Beth group. I know people with muscle tissue disorder and a partially collapsed lung, who are Charlie – they prefer powerwalking and yoga. And I know people who are Denise or Elton, with chronic pain and no or very minimal win conditions.

Bayes' Law is About Multiple Hypothesis Testing

I've called outside view the main debiasing technique, and I somewhat stand by that, not only because base-rate neglect can account for a variety of other biases, but also because outside view is about working on the policy level, which you *have* to do to implement other debiasing strategies.

Nonetheless, I am here today to tell you why the Method of Multiple Working Hypotheses is a central technique. T. C. Chamberlin wrote about it in 1897. More recently, Heuer discusses a very similar technique in Psychology of Intelligence Analysis, which served for a time as the debiasing handbook for the CIA. Heuer called his version Analysis of Competing Hypotheses.

(So, we could call it Method of Multiple Hypotheses (MMH), Analysis of Competing Hypotheses (ACH), or perhaps Analysis of Alternative Hypotheses (AAH) -- it seems doomed to be abbreviated as some variety of grunt.)

Heuer found that asking people to articulate the assumptions behind their assertions did not work very well -- analysts tend to insist that their conclusions follow directly from looking at the data, with no assumptions in between. (It is difficult to see the lens which you use to see!) However, if you instead ask people to *compare their conclusions to other possibilities*, they start noticing the assumptions which pointed them in one direction rather than another.

In order to make it stick in people's heads, I want to explain why it is just about inevitable from Bayes' Law.

Bayes' Law compares hypotheses to each other in terms of their likelihood ratios, balanced by the priors. Testing a *single* hypothesis *feels* meaningful, perhaps because in logical/deterministic cases we sometimes can prove or disprove something on its own. In the general case, though, we have to compare a hypothesis to alternatives to say anything meaningful. It's much like trying to evaluate a plan in isolation -- you can figure out a probability of success, or an expected value, but this is meaningless in isolation. You need to compare it to alternatives to know anything about whether you want to enact the plan. And, not just *any* alternatives; the best alternatives you can come up with.

Similarly, it only makes sense to evaluate hypotheses by looking at their *relative* likelihoods in comparison to a number of other hypotheses, and relative prior probabilities.

This is the thing which null hypothesis testing is sweeping under the rug. Null hypothesis testing attempts to fake testing a single hypothesis in isolation by comparing it to a "null" hypothesis which is taken to be the default thing we would believe. This often makes enough sense to not be glaringly terrible, but misrepresents the epistemics. There should not be special hypotheses which we consider "default".

A common way of writing Bayes' Law makes it look as if you can judge probability in isolation:

$$P(h \mid e) = \frac{P(e \mid h)\,P(h)}{P(e)}$$

Variables 'h' and 'e' here are supposed to remind us of 'hypothesis' and 'evidence'. It looks like we're able to evaluate hypothesis h on its own merits. However, another common statement of the law shows some of the complexity by expanding out the denominator:

$$P(h \mid e) = \frac{P(e \mid h) P(h)}{P(e \mid h) P(h) + P(e \mid \neg h) P(\neg h)}$$

In words: we can judge hypotheses in isolation by multiplying their prior probability P(h) by their likelihood P(e|h). We could call P(e|h)P(h) the "goodness" of h. This doesn't give a number which sums to one, though; we have to *normalize*, by dividing the "goodness" of each hypothesis by the total "goodness" of all hypotheses. The resulting number is between 0 and 1, so it can be a probability; indeed, it is the posterior probability.

However, note that the revised formulation represents alternatives to h simply by the negation, ¬h. This is still hiding a lot of complexity. How do we compute the "goodness" of not-h? In simple situations, this might be clear. But, to my mind, this invites the same sort of mistakes which can be made in null hypothesis testing: testing against a straw "default" hypothesis, rather than against the strongest alternative hypotheses you can think of.

Yet another common form of Bayes' Law unpacks this simplification. We consider a family of hypotheses, $h_1, h_2, \ldots, h_n$:

$$P(h \mid e) = \frac{P(e \mid h) P(h)}{\sum_i P(e \mid h_i) P(h_i)}$$

Now we've got it: we see the need to enumerate every hypothesis we can in order to test even one hypothesis properly. The previous use of ¬h was just hiding "all the other hypotheses", and the original denominator, P(e), hid it further still.

It's like... optimizing is *always* about evaluating more and more alternatives so that you can find better and better things. Optimizing for accurate beliefs is different only in that you want to weigh your several options together, rather than taking only the best one after. But, still, how can you expect to find good hypotheses if you're not generating as many as you can and allowing them to compete on the data?

Heuer tries to get people to do this by telling them to make a grid, with all the hypotheses written on the top and all the significant pieces of evidence written on the side. Rather than figuring out exact likelihood ratios, you can write "+" or "-" to indicate very roughly how well hypotheses match up to evidence:

In fleshing out this fake example, it occurred to me that I had to also include "no data breach" to be able to examine the evidence in favor of the breach. Really, it should be split into more hypotheses (which might give alternative explanations of why Victor knew too much). As we see, the evidence in favor of the breach is actually not as strong as one might think, given the priors against it and the lack of evidence in favor of any particular type of breach. (However, we can also see how rough and potentially misleading simply writing plusses and minuses can be!)

This seems better than nothing, but I can see several problems with it:

- It is easy to forget the "prior" -- I had to lump it in with evidence. In fact, I think Heuer doesn't put the prior in at all.
- The chart format makes you think of "compatibility" between hypothesis and evidence in a fairly symmetric way; it doesn't jump out at you that you're supposed to be writing P(e|h) rather than P(h|e).

In any case, I think the cognitive gear which Heuer and Chamberlin are pointing at is very important. It is more precise than the common pattern "try very hard to falsify your hypothesis" (though that mental movement may still prove useful), because it isn't obvious how to try to falsify a hypothesis; coming up with good alternative hypotheses is a necessary step.

When I first read about Heuer's ACH method, I remember having thoughts along the lines of "this debiases in a lot of different ways!" -- but I can't recall the biases I thought it covered, now. Fortunately, cousin_it has recently been thinking about it, and made his own attempt to list implications, which I'll quote in whole:

> T.C. Chamberlin's "Method of Multiple Working Hypotheses", as discussed by Abram here, is pretty much a summary of LW epistemic rationality. The idea is that you should look at your data, your hypothesis, and the next best hypothesis that fits the data. Some applications:
>
> Wason 2-4-6 task: if you receive information that 1-2-3 is okay and 2-4-6 is okay while 3-2-1 isn't, and your hypothesis is that increasing arithmetic progressions are okay, the next best hypothesis for the same data is that all increasing sequences are okay. That suggests the next experiment to try.
>
> Hermione and Harry with the soda: if the soda vanishes when spilled on the robes, and your hypothesis is that the robes are magical, the next best hypothesis is that the soda is magical. That suggests the next experiment to try.
>
> Einstein's arrogance: if you have a hypothesis and you've tried many next best hypotheses on the same data, you can be arrogant before seeing new data.
>
> Witch trials: if the witch is scared of your questioning, and your hypothesis is that she's scared because she's guilty, the next best hypothesis is that she's scared of being killed. If your data doesn't favor one over the other, you have no business thinking about such things.
>
> Mysterious answers: if you don't know anything about science, and your hypothesis is that sugar is sweet because its molecule is triangular, the next best hypothesis is that the molecule is square shaped. If your data doesn't favor one over the other, you have no business thinking about such things.
>
> Religion: if you don't see any miracles, and your hypothesis is that God is hiding, the next best hypothesis is that God doesn't exist.
>
> And so on. It's interesting how many ideas this covers.

The way this has entered into my personal thought patterns is: when I've come to some solid-seeming conclusion (in my own thoughts or in discussion), make it a principle to list alternatives (until the point where it has more cost than expected benefit). I think this has saved me a month or two of wasted effort on one occasion (though it is possible I would have noticed the problem sooner than that by some other means).

Happy debiasing!

Shared interests vs. collective interests

This is a linkpost for https://blog.obormot.net/Shared-interests-vs-collective-interests

Suppose that I, a college student, found a student organization—a chapter of Students Against a Democratic Society, perhaps. At the first meeting of SADS, we get to talking, and discover, to everyone's delight, that all ten of us are fans of *Star Trek*.

This is a shared interest.

Shared interests

A *shared interest*—in the way I am using the term—is nothing more than what it sounds like: an interest (in the broad sense of the word) that happens, for whatever reason, to be shared among all members of a group.

The distinction I want to draw is between a *shared interest* (of a group) and a *collective interest* (of a group). The former is a superset of the latter; all collective interests are, by definition, shared interests; but not all shared interests are collective interests.

Collective interests

What is a collective interest?

Well, suppose that I found *another* student organization (extracurricular activities look great on a résumé). This one is a *Star Trek* fan club. At the first meeting of Campus Trekkies United, we discover, to no one's surprise, that all fifteen of us are fans of *Star Trek*.

… well, of course we're all fans of *Star Trek*. That's *why* we're in the fan club in the first place! Anyone who's not a fan, has no reason to join the fan club. And so: *Star Trek* fandom is a collective interest of Campus Trekkies United.

A *collective interest* is an interest that is shared by every member of a group *in virtue of being a member of that group*. Anyone who does *not* share that interest, will not be a group member.[1] And thus, by *modus tollens*: anyone who is a member of the group, will share that interest. It is *guaranteed* that every member of the group will share that interest.

Details & implications

Several important consequences follow from this.

Preservation of interests

Unlike a collective interest, a shared interest is not at all guaranteed to *stay* shared among all group members. Nothing stops someone from joining the Students Against a Democratic Society, who does not like *Star Trek*. At that point, *Star Trek* fandom ceases to be a shared interest of SADS. (Which may lead to some awkward consequences if, for instance, we had decided to start wearing colorful jumpsuits to our political rallies.)

Infiltration

I said earlier that "[i]t is *guaranteed* that every member of the group will share [a collective] interest". But is this really true? Well, it's true if the condition for an interest being a collective one holds: that anyone who does *not* share the interest, will not join the group.

But it is dangerous to simply *assume* that this condition holds, in the absence of any mechanism by which it is ensured to hold! Is Campus Trekkies United actually making sure that non-Trekkies do not join? Certainly it seems like they have no reason to *want* to join, but is that *sufficient* to keep them out?

Suppose a fan of *Star Wars*, incensed at the idea that the university would grant meeting space and funds to fans of the rival franchise, decides to pose as a Trekkie, and signs up for Campus Trekkies United under false pretenses. He hates *Star Trek*, and wants nothing more than to see the club cease all *Trek*-related activities, and transform into, say, Campus Jedis United. Now *Star Trek* fandom is no longer a collective interest of the members of Campus Trekkies United—because they did not ensure that the condition of a collective interest holds.

In fact, it would be more precise to say that *Star Trek* fandom was *never* a *collective* interest, only ever a *shared* one—because the condition of a collective interest *never held in the first place*!

The universal collective interest

A collective interest of Students Against a Democratic Society (ostensibly) is being against a democratic society. A collective interest of Campus Trekkies United (ostensibly) is being a fan of *Star Trek*.

But there is one sort of collective interest that will be present in *any* organization:

The continued existence of the organization itself.

Groups are how humans achieve their goals. Organization is power. It is in the interest of any member of an organization that the organization continue to exist. Any other shared interest may fail to be a collective one—except for this one.

Illusions

Suppose that a proper subset of a group's members share a certain interest. This may be coincidence—nothing more than a consequence of base rates of that interest in the general population. But it may also be due to the fact that a proper subset of the group's members itself constitutes a coherent group, which has collective interests of its own.

This also manifests in a more interesting way, as follows:

Suppose it is claimed that a certain interest is a collective interest of a given group. However, investigation reveals group members that do not share that interest.

The claimant(s) may cry “No true Scotsman”, “infiltrator”, etc. But another (and, it seems to me, more likely) explanation is that the claimed collective interest is indeed a collective interest—*not* of the whole group it’s claimed of, but rather of a proper subset of the greater group (which subset, however, may find it advantageous to be identified with the greater group).

(Finding examples of this dynamic is left as a fairly straightforward exercise to the reader.)

[1] Note that the inverse—that anyone who *does* share the interest, *will* be a group member—need not be true!

Fuzzy Boundaries, Real Concepts

Summary: *Certain basic concepts are still very useful, even if they have fuzzy or contested boundaries, or break down in edge cases. This is basically just working out a few examples of The Cluster Structure of Thingspace. Two important examples are honesty and consent.*

Adam: I can't believe that scientists 'decided' Pluto wasn't a planet. The absolute gall!

Betsy: Oh hey Adam, the demotion of Pluto is a pretty neat story! The astronomers originally didn't have to think about the definition of a planet, because they were so clearly different from other bodies they could see in the Solar System, like asteroids and comets and moons; but then they discovered Eris and many more objects like it, and they had to figure out an actual definition...

Adam: But Pluto is obviously a planet! I learned it as part of my acronym- My Very Excellent Mother Just Served Us Nine Pizzas! A few new asteroids can't change that!

Betsy: Okay, but Eris is more massive than Pluto. Sure, it's farther out, but so are some gas giant exoplanets we've discovered around other stars. The astronomers needed to agree on a definition that wasn't incredibly convoluted, so they decided that planets needed to be massive enough that gravity makes them approximately round, and massive enough to fling smaller asteroids out of their orbit.

Adam: You make it sound like astronomers can just decide what counts as a planet or not. That's ridiculous!

Chris: You both are so naive. There's no such thing as a planet, there's only collections of atoms.

Betsy: Um, Chris, that's *really* not helpful here. See, there are several things that the eight planets have in common with each other that nothing else orbiting the Sun is even close to; for instance, on the "clearing its own orbit" front, each of the eight planets accounts for more than 99.98% of the total mass in its orbit, while no other object accounts for more than one-third. (Pluto accounts for only 7 percent of the total mass in its orbit.)

Chris: But what if we discovered another object the size of Pluto that only accounted for 99.9% of the mass in its orbit? 99%? 90%? It's completely arbitrary where you draw the line!

Adam: (mumbling) Pizzas! You can't just serve us nine nothings!

Betsy: Yes, Chris, there may or may not be a gray area when it comes to exoplanets, since these things happen on a continuum and we don't know how common various kinds of orbiting objects are yet. But for present purposes, the clustering is pretty decisive.

Chris: Aha! You admit it, that there's no matter of fact about what constitutes a planet. Why, I'm as much of a planet as Mars is!

It shouldn't be surprising that Betsy speaks for me here. The existence of gray areas and fuzzy boundaries and edge cases doesn't prevent "planet" from being a very useful concept, especially for astronomers trying to detect them around other stars.

Adam's intuitions are pretty naive, and I'd just point him to 37 Ways That Words Can Be Wrong. It's Chris's intuitions that I want to discuss. The name for those intuitions, in several philosophical debates, is called eliminativism: the assertion that, since a certain concept is not ontologically fundamental, the concept is an illusion.

Yes, nothing is fundamental in this world except its most basic physics; all the concepts in our daily lives are trying to draw a boundary that includes a lot of examples with some properties in common, not perfect mathematical definitions.

Not all concepts are useful, of course; 'phlogiston' added nothing to the concept of 'fire' except for intuitions about fluids, which mostly proved to make poor predictions. But the concepts that come up often in our daily lives are often there because they pay rent in some sense. Sometimes they lump together disparate things and need to be split more finely to be more useful; sometimes they encode assumptions that aren't true, and only pay rent about human reactions (e.g. the concept of 'sin' as distinct from 'shame', 'harm', etc); but rarely are they entirely without content when it comes to clustering reality.

Eliezer spent some long sequences trying to articulate the core of certain concepts, most prominently choice and morality. It's worth noting that neither was a perfectly precise definition; we can come up with edge cases where those definitions get weird, and particularly in the latter case, we get legitimate gray areas where we don't know how best to extend our reasoning. But that doesn't mean that there's no moral distinction between the observable universe being turned into paperclips and the kind of future I'd prefer.

So, just to provide a handy resource, I thought I'd discuss two concepts that I've seen some particularly frustrating eliminativist intuitions on: honesty and consent.

In the case of honesty, I think there's an intuitive definition that does a terrible job at useful clustering, so let's get that out of the way first:

Bad Definition Of Honesty: *Saying only things that you personally can affirm as true.*

This definition is both too strict and too lax for the purposes that we want to use it. Too strict, because we don't usually take jokes and absurdities (when clearly denoted as such by tone or context) to be dishonest; too lax, because any clever person can easily make a less clever person believe falsities without saying anything technically untrue.

So, in order to think more carefully about honesty, what do we most want to use the concept for? In what ways do I care if a person is 'honest'?

Most obviously, I want to know whether I should count their statements as good evidence. People tend to have relatively stable habits when it comes to the degree on which you can trust their utterances, so that's a pretty useful concept to have for a person or an action. So with that in mind, here's my preferred definition of honesty- actually, it's easier to state as a negative:

Better Definition Of* Dis *honesty: *Communicating in a way intended to miscalibrate the recipient.*

I very precisely said 'miscalibrate'. It's not dishonest to refuse to give someone information; it *is* dishonest to make them believe false information.

Some nice examples of this definition in practice:

- It's perfectly honest to conform to cultural norms of expression, if it will be interpreted as such. If you select only the positive things to say about yourself on a resume, well, that's what the reader expects. If you select only the positive things to say about yourself when you're under subpoena, then that's a rather different set of expectations.
- And if you're asked how you're doing by an acquaintance on the street, and you're actually doing awful but say 'fine' to save an awkward conversation, knowing that gets interpreted as a non-informative answer, then you haven't been dishonest. Spinning a fake story about how your life is going great, however, does count as dishonesty.
- Dishonesty is not always wrong! Lying to the Gestapo to save lives does, in my opinion, constitute morally good dishonesty (I'd admire someone who did this successfully, but I'd also treat their utterances in other cases with a fair bit of doubt). Board games are another example of morally legitimate dishonesty; I don't think there's anything morally unsound about a Diplomacy champion, but I would be a bit more paranoid about making a big business deal with them.

There are legitimate gray areas when it comes to honesty. If you fear you'll be grossly misunderstood whatever you say, it's very hard to find an honest course. And if you're speaking or writing to a mass audience, it's nigh impossible to cover all their misunderstandings at once without making yourself unintelligible with caveats.

But "oh, it would be so terrible if this person knew this fact, but it would hurt them if I refused to answer their question, so let me just misdirect them a bit"? That's not a gray area. That's dishonesty in *exactly* the sense people care most about.

For consent, I don't have as tidy a definition in general, but here's a limited one for a subset of it:

One Definition of Basic Bodily Consent: *Don't touch a person in a way they clearly prefer not to be touched.*

Consent in general concerns all kinds of human preferences, and since not all of these can be met simultaneously, there are some pretty complex tradeoffs to manage. Basic bodily consent avoids some of that complexity, by virtue of the fact that it's pretty rare to have different people's desires not to be touched conflict with each other. And it's more or less become the official norm of upper-class adult Western society (which is not to say that it's always observed there, just that it's well within the Overton window to call it monstrous when people violate it- this was not nearly as much the case two generations ago).

And it serves a pretty obvious purpose. We're primates with fragile physical bodies, and an evolutionary history of violence. Unwanted touching is a pattern-match for a sudden assault; it often raises our fear and anger in strong and predictable ways. And the new norm isn't that novel; the effective norm that Basic Bodily Consent replaced

was "don't touch an *equal-or-higher-status* person in a way they clearly prefer not to be touched". We just added some egalitarianism to that.

One obvious complication to basic bodily consent is preconsent: if you join the army, you had better be aware that you're going to lose your bodily autonomy in some ways that would be very objectionable in civilian life. You might then very much not prefer to experience what you're experiencing, but you did sign up for it.

Some people will raise BDSM as another complication, but if you've met people who are v*ery* deep in BDSM culture, it's amazingly clear how much they think (in everyday life, not just the bedroom) about the nuances of bodily consent, how to handle discrepancies between preconsent and feelings in the moment, and more. (If anyone in the comments wants to suggest a good reference on advanced consent, that might be helpful.)

And as before, something being a violation of basic bodily consent doesn't always make it wrong. I will yank a toddler out of the way of a speeding car in the safest way I can, not the gentlest. (But also as before, be careful about rationalizing paternalistic reasons for violating consent! You are running on corrupted hardware.)

Legitimate gray areas include things like very weak preferences and guessing about unstated and unconscious preferences.

"This person told me they don't want to be hugged, but it'll be better for them if I expand their comfort zone, so hugs away" is *not* a gray area. It's a clear violation.

One last thing, while I'm here: my general assertion is "it all adds up to normality", or more specifically, "most of the common concepts that describe human interactions are useful, and need editing rather than discarding". My best example of this principle came when I realized at age 23 that my religion was almost surely untrue. I thought at the time that my morals were all based on the religion, so I felt like I was without ethical guidance. In order to avoid doing things I might regret, though, I resolved to abide by each of my basic moral principles until I felt I'd thought them through well enough to change or abandon them. Then I started reading some moral philosophy books.

Some of those precepts were indeed mirages (my religion had a silly thing against the kinds of hedonism that hurt no-one), but most of the basic moral principles, including honesty and altruism, turned out to be based in things I still found myself caring about. And I'm glad that my past self didn't foolishly binge on violating those norms before he figured out how they actually fit into a non-religious worldview.

So if you're annoyed by the naivety of the discourse on some topic, I suggest that it's better to look into how the concept is being used and what it's useful for, and maybe try and argue for a reshaping, rather than abandoning it wholesale immediately. You are not actually as much a planet as Mars is, after all.

On Better Mental Representations Part I: Adopting 'Thinking Tools'

If you don't mind an introduction using some poppy philosophical thinking, please indulge me as I channel Alan Watts. Correct me if I am wrong but if I remember Watts once spoke or written about how we misconstrue our representations of things for the actual things. In extension of that, we can also say (as we measure) that we also misconstrue our 'valuation' of things to be the things themselves.

Watts gave out a quite clever example of this which involves wood (natural resource) and inches (measurement). Let me roughly paraphrase through this anecdote (adding some creative flair or more honestly winging it without watering down the gist).

There's a foreman and his men building a house. They had all the natural resources that need be there to build an adequate house for a small family of four. But for some "price shock" in the "Platonic" realm, the currency of measurement by inches ran out of circulation.

The foreman told his workers. "Sorry, fellas. We can't finish building the house anymore. We have ran out of inches."

"What do you mean? We still have planks, nails, tools like hammers, and enough man power to finish building the house.", answered a carpenter amongst the team.

The foreman replied, "Sorry, Edward. You don't understand. The world has ran out of inches."

Ludicrous isn't it? This is also how many of us think about the "value" of things. We think of them in monetary value. Inches, like money, is a representation of something: the length of some *thing*. It is not the thing itself. It is a measure.

It's the same thing as money. It's something that has no intrinsic value. We only give value to it. We internally represent in our minds that it has value.

I am not saying money is bad, useless, or the root of all evil, Jesus forbid. I'm just saying that it is similar to what the "inches" there serves in the anecdote. It serves, almost, as the end all be all of valuation. We, often times, confuse our representations of things to be the things themselves; i.e. we mistake concepts of things to be the things themselves.

But this is to be expected. We do think of things as how they are in our representations. This is all we can ever do. It is only from these representations that we can glean into reality. They both enable and limit our capacity to "capture" things and their features from our environment. This goes both for our body and our minds (forgive the dualism, it's just a matter of getting the point across easier).

BIO-APPROXIMATORS AND CONCEPTUAL APPROXIMATORS

Firstly, evolution has "gifted" us with a biology that can approximate reality. We now know, through science, that what we call visible light is just a *portion* of the electromagnetic spectrum that our human eyes can detect. It is but a part of the whole picture. What we detect is just an approximation of which there is there.

Our hearing works the same way. There's a certain range from about 20 to 20,000 hertz. Loudness is just our perception of the pressure exerted by sound waves to our tympanic

membrane. It's *just* air molecules going through different physical processes that convert it to "sound"--our inner representations of it. Without having a working approximator, no "sound". This works the same way with our other senses like taste, smell, and proprioception. Approximation is the name of the game. These, I term, biological approximators or bio-approximators. These are our sense and other organs that we use to navigate ourselves in the world. These can be tricked. In 2014, humans were capable of sending "smells" from New York to Paris. These more or less same humans (including co-inventor David Edwards) embarked on a mission to commercialize the scents. One "cool" thing about it is that they created a device that can be connected to a Bluetooth speaker. It can create an olfactory playlist for long commutes. One, for example, can play "Thai Beach Vacation" and the device will "play" the scents of coconut, suntan lotion, and sea breeze in a loop.

Of course, if exposed to it we don't really smell coconuts, suntan lotion, and sea breeze. We smell *highly similar* smells. We can manipulate "signals" to act "dishonestly".

In nature (or the wild, to be more precise), animals, mostly with no comprehension of doing them, emit such dishonest signals. A fiddler crab is known for having one much larger claw. If it should lose it in a battle with another male for a female, another weaker one grows. It's lighter and less effective than the original. But this 'weaker one' still looks the part and can scare other males away before they engage in combat. This is considered to be a dishonest signal. It's the same thing with the famous cuckoo bird and reed warbler dynamics. In the latter's case, it's one species against another.

But one good thing about being human is that we have conceptual approximators. These extends the "bare" capacities of our bio-approximators to understand things that evolution didn't specifically selected for us to understand. These are what Dan Dennett called 'thinking tools' . They are imagination extenders and focus-holders.

Examples of them are calculus and Bayesian inference. There are those that are mathematical in nature (abstract). There are those that are explanatory and predictive

'narratives'. Many are used by the people of science. These are quite powerful.

Like the radio telescope helps us "see" (or perceive) things our eyes normally wouldn't see (or just a plain telescope), these conceptual tools of thought let's our "mind's eye" see things that we, without them, couldn't "see".

Urbain Le Verrier just used mathematics to predict the location of an undiscovered accurately. It extended his imagination. That planet is Neptune (it didn't work for "Vulcan" though). Einstein was famous for using thought experiments for his great achievements. Extremists use their morals derived from "reason" to cause havoc and spread misinformation.

Wait what? Yes. Thinking tools are not created equal. Some may be better than others. Some are just plain dumb and dangerous. The problem is that ideologues or carriers of these faulty thinking tools mix effective "signalling" with intellectual hogwash.

Like how an "upside down goggles" can trick our eyes to see the world upside down (some people can adjust rather quickly), concepts and ideologies can also trick our "mind's eye" to see the world upside down (and still some people can adjust rather quickly, Flat Earth).

What's going on here is that ideologues use signals that would make them look legitimate, scientific, or deep. They try to look the part to play the part. This shows in their works. In the way they use facts and "facts" together. It's hard to develop a trained "mind's eye" to see such things. We can be tricked.

This is why constant testing of our selves and the tools we wield is a good exercise. Where do we test these? We send them to academic journals (the good ones who review well). We share them in a community like this. We test them in conferences or meetups to be judged and questioned by peers and other experts. We get chastised and we get helped.

While it doesn't feel good to be wrong naturally, but there is overwhelming joy to not being wrong anymore. Or, in many cases, to be less wrong.

Be Less Wrong. It's a good thing to find this rational community where people are encouraged to be "less wrong". This is exactly the direction that science has been going for. Sometimes science an be exact but at other times, all we can do is to be less wrong about things and be happy with it. Well, this is before more clever folks open the doors. For now, I think many will not be opened ever.

The adoption of tools (physical and conceptual), how their use become robust, and how they are past over from a generation to generations does not just depend upon the efficiency and honesty of such tools. It depends upon the communities that wield them. Just like in biological evolution, a gene couldn't spread without a population. Good thinking tools couldn't make their mark in the evolutionary cultural timeline without agents to pass them on, to pass through, and to be in.

As clever as hairless apes can be, humans managed with considerable success to do influence the frequency of genes through engineering. Gene drive is a technology that can help stifle the population of the mosquitoes that carry malaria. It is by cutting out the ability of a mosquitoes to carry malaria or cutting the ability to pass on its genes by making it infertile. This technology of driving a particular combination of genes in a population has even gotten funding from the Bill and Melinda Gates Foundation.

Scientists execute the gene drive through introducing genetically-altered agents carrying *harmless* and *long-lasting* genes into a 'harmful' population to infiltrate it, propagate their "specially prepared genes" in it to lessen the frequency of harm-causing genes endemic in the population in the first place. I argue that it's the same thing with ideas and thinking tools.

Causes have their lobbyist, ideologues, and propagandists. They infiltrate institutions like the government, education, and the media with the goal of spreading their way of thinking. It is through 'tapping' in many communities that they can likely spread these concepts.

For the people who hold rational thinking high up in the pedestal (I think I do belong in this demographic), they don't have many of these specialists. I even say that it is such a loose community as many in it are too busy within their own areas of research. There is no bloc. Many would like to do away with the politics of it and just focus on gazing into the glorious light at the edge of human understanding.

But there is no time to waste. These tools for the "mind's eye" need to be cultivated in communities. It is time for people in rational communities like this to act as agents of change; not to force feed reason and science into others but to cultivate the very same interest that got themselves hooked in the first place; not to make enemies but peers.

Of course, this is the ideal. It doesn't necessarily hold in reality. I'm sure many of us have bumps along the way. It's not easy. But as UFC featherweight champion Jerome Max Halloway says "It is what it is."

I'm not high on the "meme" concept but I think it works fine here with not much of a bad intellectual repercussion. Just like the *gene drive* it's high time for intense *meme drives* to let other people out there with the germ for the love of wisdom to get high on clearer thinking; to adopt thinking tools in their attempts to do thinking better.

As Bo Dahlbom, a computer scientist, philosopher, and former student of Dan Dennett, states:

> "*Just as you cannot do much carpentry with your bare hands, there is not much thinking you can do with your bare brain.*"

For Part II. In the second part of this series, I'll delve more on the concept of the degrees of accuracy and rationality.

References:

Abrahams, M. (2017, February 22). Experiments show we quickly adjust to seeing everything upside-down. Retrieved from https://www.theguardian.com/education/2012/nov/12/improbable-research-seeing-upside-down

Bereznak, A. (2014, June 17). Harvard Scientists Send the First Transatlantic Smell via iPhone. Retrieved from https://finance.yahoo.com/news/harvard-scientists-send-the-first-transatlantic-smell-89078729859.html

Backwell, P. R., Christy, J. H., Telford, S. R., Jennions, M. D., & Passmore, J. (2000). Dishonest signalling in a fiddler crab. *Proceedings of the Royal Society of London B: Biological Sciences*, *267*(1444), 719-724.

Conner-Simmons, A. (2015, April 14). How three MIT students fooled the world of scientific journals. Retrieved from https://news.mit.edu/2015/how-three-mit-students-fooled-scientific-journals-0414

Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. WW Norton & Company.

Dyer, H. T. (2018, May 2). I watched an entire Flat Earth Convention for my research ? here's what I learnt. Retrieved from https://theconversation.com/i-watched-an-entire-flat-earth-convention-for-my-research-heres-what-i-learnt-95887

Encylopaedia Brittanica. (2017, July 21). Human ear - The physiology of hearing. Retrieved from https://www.britannica.com/science/ear/The-physiology-of-hearing

Kelland, K. (2018, April 18). Gates backs gene technologies in fight to end malaria. Retrieved from https://www.reuters.com/article/us-health-malaria-gates/gates-backs-gene-technologies-in-fight-to-end-malaria-idUSKBN1HP2QF

Twilley, N. (2016, April 27). Will Smell Ever Come to Smartphones? Retrieved from https://www.newyorker.com/tech/elements/is-digital-smell-doomed

Watts, A. W. (2010). *Does It Matter?: Essays on Man s Relation to Materiality*. New World Library.

Worral, S. (2015, November 4). The Hunt for Vulcan, the Planet That Wasn't There. Retrieved from https://news.nationalgeographic.com/2015/11/151104-newton-einstein-gravity-vulcan-planets-mercury-astronomy-theory-of-relativity-ngbooktalk/

Wyss Institute. (2017, August 18). CRISPR-Cas9: Gene Drives. Retrieved from https://wyss.harvard.edu/media-post/crispr-cas9-gene-drives/

Everything I ever needed to know, I learned from World of Warcraft: Goodhart's law

This is a linkpost for https://blog.obormot.net/Everything-I-ever-needed-to-know-I-learned-from-World-of-Warcraft-Goodharts-law

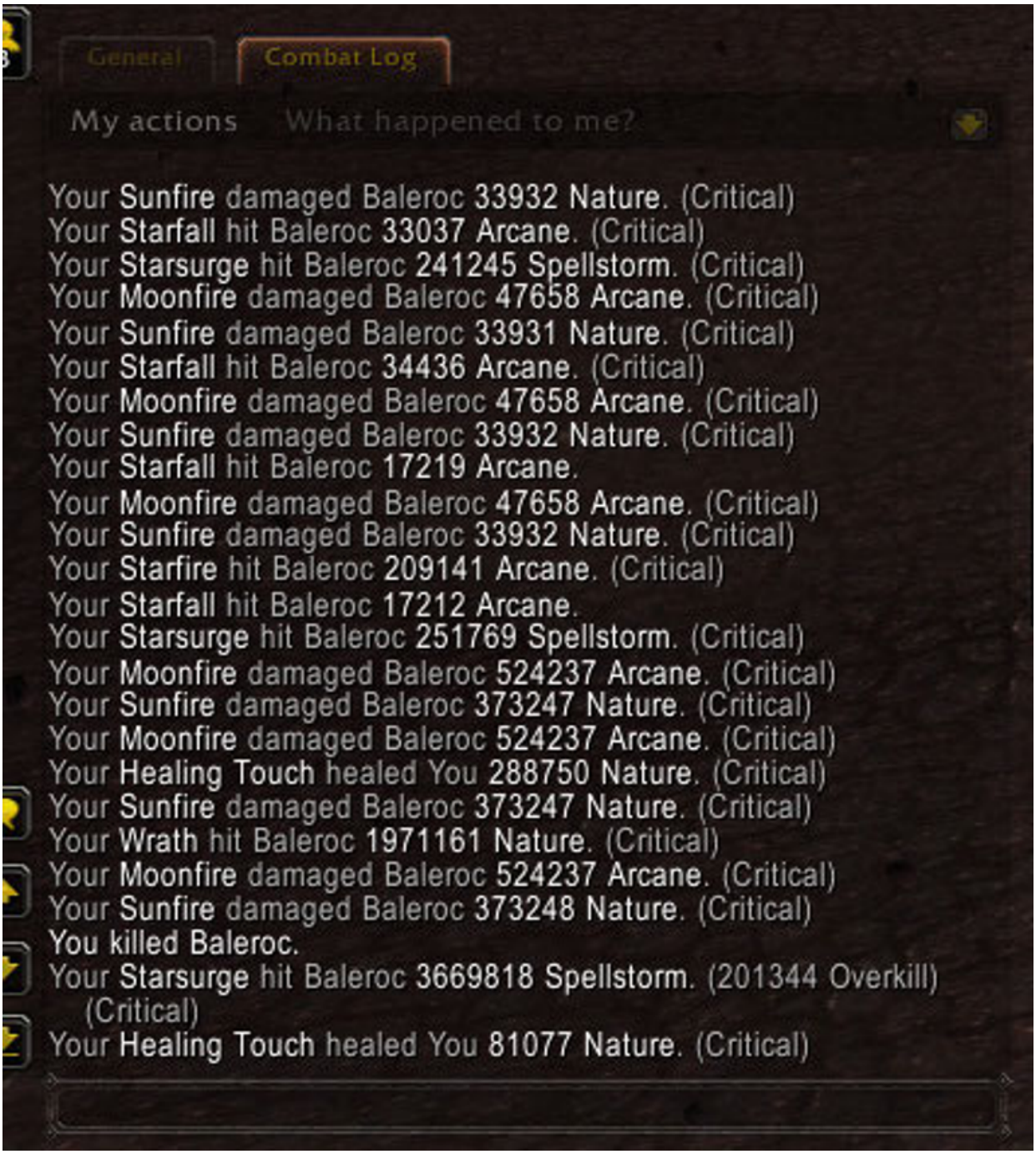
This is the first in a series of posts about lessons from my experiences in World of Warcraft. I've been talking about this stuff for a long time—in forum comments, in IRC conversations, etc.—and this series is my attempt to make it all a bit more legible. I've added footnotes to explain some of the jargon, but if anything remains incomprehensible, let me know in the comments.

World of Warcraft, especially *WoW* raiding[1], is very much a game of numbers and details.

At first, in the very early days of *WoW*, people didn't necessarily appreciate this very well, nor did they have any good way to use that fact even if they did appreciate it. (And—this bit is a tangent, but an interesting one—a lot of superstitions arose about how game mechanics worked, which abilities had which effects, what caused bosses[2] to do this or that, etc.—all the usual human responses to complex phenomena where discerning causation is hard.) And, more importantly and on-topic, there was no really good way to sift the good players from the bad; nor to improve one's own performance.

This hampered progression. ("Progression" is a *WoW* term of art for "getting a boss down, getting better at doing so, and advancing to the next challenge; rinse, repeat". Hence "progression raiding" meant "working on defeating the currently-not-yet-beaten challenges".)

The combat log



One crucial feature of *WoW* is the **combat log**. This is a little window that appears at the bottom of your screen; into it, the game outputs lines that report everything that happens to or around your character. All damage done or taken, all hits taken or avoided, abilities used, etc., etc.—*everything*. This information is output in a specific format; and it can be parsed by the add-on system[3].

Naturally, then, people soon began writing add-ons that did parse it—parse it, and organize it, and present various statistical and aggregative transformations of that data in an easy-to-view form—which, importantly, could be viewed *live*, as one played.

Thus arose the category of add-ons known as “damage meters”.

The damage meters

Of course the “damage meters” showed other things as well—but viewing damage output was the most popular and exciting use. (What more exciting set of data is there, but one that shows how much you’re hurting the monsters, with your fireballs and the strikes of your sword?) The better class of damage-meter add-ons not only recorded this data, but also synchronized and verified it, by communicating between instances of themselves running on the clients of all the people in the raid.

Which meant that **now** you could have a centralized display of just what exactly everyone in the raid was doing, and how, and how well.

This was a great boon to raid leaders and raid guilds everywhere! You have a raid of 40 people, one of the DPSers[4] is incompetent, can’t DPS to save his life, or he’s AFK[5] half the time, or he's just messing around—who can tell?

With damage meters—everyone can tell.

Now, you could sift the bad from the good, the conscientious from the moochers and slackers, and so on. And more: someone’s not performing well but seems to be trying, but failing? Well, now you look at his ability breakdown[6], you compare it to that of the top DPSers, you see what the difference is and you say—no, Bob, don't use ability X in this situation, use ability Y, it does more damage.

The problem

All of this is fantastic. But… it immediately and predictably began to be subverted by Goodhart’s law.

To wit: if you are looking at the DPS meters but “maximize DPS” is not perfectly correlated with “kill the boss” (that being, of course, your goal)… then you have a problem.

This may be obvious enough; but it is also instructive to consider the *specific ways* that those things can come uncoupled. So, let me try and enumerate them.

The Thing is valuable, but it’s not the only valuable thing

There are other things that must be done, that are less glamorous, and may detract from doing the Thing, but each of which is a *sine qua non* of success. (In *WoW*, this might manifest as: the boss must be damaged, but also, adds must be kited—never mind what this means, know only that while a DPSer is doing **that**, he can’t be DPSing!)

And yet more insidious elaborations on that possibility:

We can’t afford to specialize

What if, yes, this other thing must be done, but the maximally competent raid member must **both** do that thing and **also** the main thing? He won’t DPS as well as he could, but he also can't just *not* DPS, because then you fail and die; you can’t say “ok, **just** do the other thing and forget DPSing”. In other words, what if the secondary task isn’t just something you can put someone full-time on?

Outside of WoW, you might encounter this in, e.g., a software development context: suppose you’re measuring commits, but also documentation must be written—but you don’t have (nor can you afford to hire) a dedicated docs writer! (Similar examples abound.)

Then other possibilities:

Tunnel vision kills

The Thing is valuable, but tunnel-visioning on The Thing means that you will forget to focus on certain other things, the result being that you are horribly doomed somehow—this is an *individual* failing, but given rise to by the incentives of the singular metric (i.e., DPS maximization).

(The *WoW* example is: you have to DPS as hard as possible, *but* you also have to move out the way when the boss does his “everyone in a 10 foot radius dies to horrible fire” ability.)

And yet more insidious versions of this one:

Tunnel vision kills… other people

Yes, if this tunnel-vision dooms **you**, personally, in a predictable and unavoidable fashion, then it is easy enough to say “do this other thing or else you will predictably **also** suffer on the singular metric” (the dead throw no fireballs).

But the *real* problem comes in when neglecting such a secondary duty creates *externalities*; or when the destructive effect of the neglect can be pushed off on someone else.

(In WoW: “I won’t run out of the fire and the healers can just heal me and I won’t die and I’ll do more DPS than those who don’t run out"; in another context, perhaps “I will neglect to comment my code, or to test it, or to do other maintenance tasks; these may be done for me by others, and meanwhile I will maximize my singular metric [commits]”.)

It's almost *always* the case that **you** have the comparative advantage in doing the secondary thing that avoids the doom; if others have to pick up your slack there, it'll be way less efficient, overall.

Optimization has a price

The Thing is valuable, yes; and it may be that there are ways to *in fact* increase your level of the Thing, really do increase it, **but** at a non-obvious cost that is borne by *others*. Yes, you are improving *your* effectiveness, but the price is that others, doing other things, now have to work harder, or waste effort on the consequences, etc.

(Many examples of this in WoW, such as "start DPSing before you're supposed to, and risk the boss getting away from the tank and killing the raid". In a general context, this is "taking risks, the consequences of which are dire, and the mitigation of which is a cost borne by others, not you".)

Then this one is particularly subtle and may be hard to spot:

Everyone wants the chance to show off their skill

The Thing is valuable, and doing it well brings judgment of competence, and therefore status. There are *roles within the project's task allocation* that naturally give greater opportunities to maximize your performance of the Thing, and **therefore** people seek out those roles preferentially—even when an optimal allocation of roles, by relative skill or appropriateness to task, would lead them to be placed in roles that do not let them do the most of the Thing.

(In WoW: if the most skilled hunter is needed to kite the add, but there are no "who kited the add best" meters, only damage meters… well, then maybe that most skilled hunter, when called upon to kite the add, says "Bob over there can kite the add better"—and as a result, because Bob actually is *worse* at that, the raid fails. In other contexts… well, many examples, of course; glory-seeking in project participation, etc.)

Of course there is also:

A good excuse for incompetence

This is the converse of the first scenario: if the Thing is valuable but you are bad at it, you might deliberately seek out roles in which there is an *excuse* for not performing it well (because the role's *primary* purpose is something else)—despite the fact that, actually, the ideal person in your role **also** does the Thing (even if not *as much* as in a Thing-centered role).

---

1. "Raid dungeons" were the most difficult challenges in the game—difficult enough to require up to 40 players to band together and cooperate, and cooperate *effectively*, in order to overcome them. "Raiding" refers to the work of defeating these challenges. Most of what I have to say involves raiding, because it was this part of *WoW* that—due to the requirement for effective group effort (and for other, related, reasons)—gave rise to the most interesting social patterns, the most illuminating group dynamics, etc. ↩

2. “Boss monsters” or “bosses” are the powerful computer-controlled opponents which players must defeat in order to receive the in-game rewards which are required to improve their characters’ capabilities. The most powerful and difficult-to-defeat bosses were, of course, raid bosses (see previous footnote). ↩

3. *WoW* allows players to create add-ons—programs that enhance the game’s user interface, add features, and so on. Many of these were very popular—downloaded and used by many other players—and some came to be considered necessary tools for successful raiding. ↩

4. “Damage Per Second”, i.e. doing damage to the boss, in order to kill it (this being the goal). Along with “tank” and “healer”, “DPS” is one of the three roles that a character might fulfill in a group or raid. A raid needed a certain number of people in each role, and all were critical to success. ↩

5. “Away From Keyboard”, i.e., not actually at the computer—which means, obviously, that his character is standing motionless, and not contributing to the raid’s efforts in the slightest. ↩

6. In other words: which of his character’s abilities he was using, in what proportion, etc. Is the mage casting Fireball, or Frostbolt, or Arcane Missile? Is the hunter using Arcane Shot, and if so, how often? By examining the record—recorded and shown by the damage meters—of a character’s ability usage, it was often very easy to determine who was playing optimally, and who was making mistakes. ↩

[LINK] How to write a dominant assurance contract on the Ethereum blockchain

This is a linkpost for https://programtheblockchain.com/posts/2018/05/01/writing-a-dominant-assurance-contract/

April links

This is a linkpost for https://www.gwern.net/newsletter/2018/04

Mini-review: The Book of Why

Someone should probably write a real book review, but to make a brief recommendation: *The Book of Why* by Judea Pearl and Dana Mackenzie is probably the most interesting general-science book I've read since *Thinking Fast and Slow.*

Pearl's goal is to explain and promote *causal inference,* which you might think of as (allegedly) the next big thing after frequentist and Bayesian statistics. The introduction is probably skippable, since the authors make some rather grand claims that aren't backed up until later. I found myself thinking, "okay, maybe it's great, but explain what it is already".

Chapter 1 introduces the *Ladder of Causation*, the authors' way of distinguishing the correlations found via a model-free statistical summary of data (which is level 1) from deductions that require a causal model (levels 2 and 3).

Chapters 2 and 3 give a partial, "whiggish" history of statistics from a causal perspective, covering frequentist and Bayesian statistics and Pearl's AI work, when he invented Bayesian networks. At the end, he talks about the possible junctions in a Bayesian network: the *chain*, *fork*, and *collider,* and how they can easily cause confusion.

Chapter 4 uses causal reasoning to explain the logic behind randomized controlled trials and other ways of controlling for confounding variables.

Chapter 5 covers the scientific debate over cigarette smoking, and how lack of clarity about causation resulted in this debate taking years longer than it needed to.

Chapter 6 is a fun chapter showing how to use causal diagrams to shed new light on the Monty Hall problem and Simpson's paradox.

And that's as far as I've read, but it's enough to make a strong recommendation.

I did a quick search on Less Wrong and causality has been covered before, though not as clearly. In particular, see Yudkowsky's Causal Diagrams and Causal Models.

(I was confused about one bit, though: Yudkowsky writes that "Causal *models* (with specific probabilities attached) are sometimes known as 'Bayesian networks' or 'Bayes nets'." But in the book, the authors make a clear distinction: "Unlike the causal diagrams we will deal with throughout the book, a Bayesian network carries no assumption that the arrow has any causal meaning." Though later, they write, "These three junctions [...] are like keyholes through the door that separates the first and second levels of the Ladder of Causation.")

Brief comment on frontpage/personal distinction

I recently moved a post back from frontpage to personal blog, a decision which was followed by confusion and had some pushback. Rather than have discussion on the post about the site (i.e. about not-the-post), I've copied my comment here instead.

A short background on this particular guideline (not having meta/community discussion on frontpage) and its purpose:

- The purpose of the rule is to make sure we have a clear, well-defined space that users can go to and not expect to discover internal discussion or political discussion.
    - To contrast, my facebook wall often has interesting discussion followed by AGH ANGSTY TRIBAL CONFLICT, and if you think of Frontpage as analogous to how academic physics has journals where the best work goes in, it would be bad to also have internal / political discussion in those.
- And to clarify, this isn't at all a question of quality. Resolving tribal conflict *is* an important part of rationality. But there are many many many many excellent posts that do not belong in Frontpage.

With that common knowledge, I read the structure of this post as being

- Observation about the community
- Laying out model of what's going on
- Suggesting a different solution for the community

Which is a fine and good post. But I think it addresses a problem that many communities don't have, and so is fairly this-community-specific.

To give further information: a solution that would be totally fine-and-dandy would be to take the model of communities, status and reinforcement and write it as its own post on the Frontpage, then separately write a personal blog post on using to analyse this community. That way it could contribute to the ongoing building of knowledge while not being very context heavy (e.g. someone else could build on the theory alone), and also discuss the important community aspects.

People will naturally tend to read the social stuff a lot. I think this rule pushes against the direction of entropy where this is everything on LessWrong, rather than a thing that happens on LessWrong but not the point of LessWrong.

Trivial inconveniences as an antidote to akrasia

1.

I was taking an five hour bus back to LA from Vegas. We were stopped at a rest stop about midway through, and I got out to stretch my legs. I figured it would be a good time to call my mom and catch up.

While chatting with her, she started to tell me about a business idea she has. She started off telling me how she loves chocolate, but has zero willpower to not overindulge. She could just not buy it at the store, but she really doesn't want to give it up outright. So her idea is to have a safe with some sort of timer on it that feed her a piece of chocolate once per day, but that's *it*.

My very first impression was to role my eyes. There is this small part of me that still instinctively just wants to disagree with whatever my mom says. But that only lasted a brief moment. After thinking about it some more, I kinda love the idea! The problem is real, and it's important to solve. The solution should be pretty foolproof, and shouldn't be too expensive.

I'm an entrepreneur, and I think about startup ideas all the time. I end up forgetting about most of them. This ten minute chat with my mom happened almost three years ago, and it still holds a place in my mind.

2.

I think her original idea is fine, but I always like to think about how ideas can be improved on. One problem I see with her idea is that it might be too strict. What if you truly do have a good reason for deviating and want to be fed chocolate now?

What if, instead of feeding you a piece of chocolate once every 24 hours (or whatever you configure it to), what if it just required you to stand there for five minutes before feeding you the chocolate? And you can't just step away and watch TV while you wait out the five minutes - what if the safe made you actually stand there and do something boring for five minutes (eg. by making you type in the letter that appears on the screen every ten seconds)?

I feel like that approach would still be pretty *damn* effective. Who is going to actually stand there for five whole minutes typing letters in to a screen while they wait for a piece of chocolate?

3.

But why, exactly, does that second approach actually work? I may be guilty of explaining by labeling here, but... "trivial inconveniences" is my answer.

Trivial inconveniences are a thing. Even though standing for five minutes (or even 60 seconds) is a relatively trivial inconvenience (think about how much time you invest to be able to eat other foods), I suspect that it would be powerful enough to prevent people from overeating chocolate. There are many other times where a trivial inconvenience is similarly powerful.

I googled around for 30 minutes or so, looking for a real answer to the question of why trivial inconveniences are powerful. I didn't really find anything. So then, I'd like to take a stab at thinking it through myself. How could a trivial inconvenience be so powerful? What is going on here?

To answer that question, I'd like to start off trying to introspect and put myself inside my mind when I am akratic. Eg. when I start snacking on chocolate when I feel that I shouldn't be.

a) In that situation, my mind feels like it's on autopilot. Sure, there are some intervening thoughts like, "wait, you aren't supposed to be doing this", but they're immediately shot down with thoughts like, "it's not that big a deal, you'll only have a few pieces, it'll make you feel good, and then you'll get right back to work". More importantly, all of these thoughts are pretty faint, and I mostly just feel like I'm on autopilot.

b) I feel a pretty strong impulse to eat chocolate. "It would taste so good. It would be so satisfying. It's right there. I *want* it!"

I sense that trivial inconveniences reduce both of those problems. I think they snap you out of the autopilot mode, and I think they provide you with a sort of buffer time period for the initial impulse to go away.

Epistemic status = not very strong, just me musing.

4.

Perhaps a more useful question than how trivial inconveniences work is how we can make them work for us. I sense that ideas similar to the safe that feeds you chocolate bars could be used to fight akrasia. I'm going to spend one Yoda Timer coming up with as many as I can. I encourage you to give it a shot also and post what you've got in the comments!

1. If you're trying not to be tempted to eat chocolate on impulse, keep it in a difficult to reach place, like the back of your cabinet, or in a top drawer that requires you to get a ladder.

2. If you live in an apartment complex and have a mail room, and you don't want to use your phone too much, lock your phone in your mail box. If you really need it, you can take five minutes to walk downstairs, open your mailbox and use it. But otherwise that trivial inconvenience of walking downstairs and opening your mailbox will probably keep you from mindlessly browsing Facebook. You can try the same thing with your internet modem, or anything else.

3. If you want to spend more time walking and biking and less time driving, park your car a few blocks away from your house/apartment. Or keep your car keys somewhere inaccessible, like your attic. (Even something like your attic is only minutes away from being accessed, not too big a deal in an objective sense.)

4. Delete apps from your phone, and when you actually need them, re-install them. Taking the time to reinstall them seems like a trivial inconvenience that should be pretty powerful.