Best of LessWrong: August 2018

1. Unrolling social metacognition: Three levels of meta are not enough.
2. Unknown Knowns
3. Historical mathematicians exhibit a birth order effect too
4. Y Couchinator
5. Alignment Newsletter #21
6. Do what we mean vs. do what we say
7. Trust Me I'm Lying: A Summary and Review
8. Subsidizing Prediction Markets
9. Alignment Newsletter #18
10. Preliminary thoughts on moral weight
11. Is there a practitioner's guide for rationality?
12. Zetetic explanation
13. Entropic Regret I: Deterministic MDPs
14. A Short Note on UDT
15. July gwern.net newsletter
16. Sandboxing by Physical Simulation?
17. Learning strategies and the Pokemon league parable
18. History of the Development of Logical Induction
19. Four kinds of problems
20. Tactical vs. Strategic Cooperation
21. Cargo Cult and Self-Improvement
22. Corrigibility doesn't always have a good action to take
23. Using expected utility for Good(hart)
24. Economic policy for artificial intelligence
25. Reducing collective rationality to individual optimization in common-payoff games using MCMC
26. Player of Games
27. Logarithms and Total Utilitarianism
28. [Paper] The Global Catastrophic Risks of the Possibility of Finding Alien AI During SETI
29. Tidying One’s Room
30. How to Build a Lumenator
31. SSC Meetups Everywhere 2018
32. Ontological uncertainty and diversifying our quantum portfolio
33. Amy Hoy's How To Master New Skills
34. Request for input on multiverse-wide superrationality (MSR)
35. Jobs Inside the API
36. Book Review: AI Safety and Security
37. Probabilistic Tiling (Preliminary Attempt)
38. Cause Awareness as a Factor against Cause Neutrality
39. Emotional Training Model
40. New paper: Long-Term Trajectories of Human Civilization
41. Counterfactuals for Perfect Predictors
42. AI Reading Group Thoughts (1/?): The Mandate of Heaven
43. Aumann’s Agreement Revisited
44. Fundamentals of Formalisation Level 7: Equivalence Relations and Orderings
45. Theories of Pain
46. Human-Aligned AI Summer School: A Summary
47. Turning Up the Heat: Insights from Tao's 'Analysis II'
48. You Play to Win the Game
49. Logical Counterfactuals & the Cooperation Game
50. Computational complexity of RL with traps

Best of LessWrong: August 2018

1. Unrolling social metacognition: Three levels of meta are not enough.
2. Unknown Knowns
3. Historical mathematicians exhibit a birth order effect too
4. Y Couchinator
5. Alignment Newsletter #21
6. Do what we mean vs. do what we say
7. Trust Me I'm Lying: A Summary and Review
8. Subsidizing Prediction Markets
9. Alignment Newsletter #18
10. Preliminary thoughts on moral weight
11. Is there a practitioner's guide for rationality?
12. Zetetic explanation
13. Entropic Regret I: Deterministic MDPs
14. A Short Note on UDT
15. July gwern.net newsletter
16. Sandboxing by Physical Simulation?
17. Learning strategies and the Pokemon league parable
18. History of the Development of Logical Induction
19. Four kinds of problems
20. Tactical vs. Strategic Cooperation
21. Cargo Cult and Self-Improvement
22. Corrigibility doesn't always have a good action to take
23. Using expected utility for Good(hart)
24. Economic policy for artificial intelligence
25. Reducing collective rationality to individual optimization in common-payoff games using MCMC
26. Player of Games
27. Logarithms and Total Utilitarianism
28. [Paper] The Global Catastrophic Risks of the Possibility of Finding Alien AI During SETI
29. Tidying One’s Room
30. How to Build a Lumenator
31. SSC Meetups Everywhere 2018
32. Ontological uncertainty and diversifying our quantum portfolio
33. Amy Hoy's How To Master New Skills
34. Request for input on multiverse-wide superrationality (MSR)
35. Jobs Inside the API
36. Book Review: AI Safety and Security
37. Probabilistic Tiling (Preliminary Attempt)
38. Cause Awareness as a Factor against Cause Neutrality
39. Emotional Training Model
40. New paper: Long-Term Trajectories of Human Civilization
41. Counterfactuals for Perfect Predictors
42. AI Reading Group Thoughts (1/?): The Mandate of Heaven
43. Aumann’s Agreement Revisited
44. Fundamentals of Formalisation Level 7: Equivalence Relations and Orderings
45. Theories of Pain
46. Human-Aligned AI Summer School: A Summary
47. Turning Up the Heat: Insights from Tao's 'Analysis II'

48. You Play to Win the Game
49. Logical Counterfactuals & the Cooperation Game
50. Computational complexity of RL with traps

Unrolling social metacognition: Three levels of meta are not enough.

Disclaimer: This post was written time-boxed to 2 hours because I think LessWrong can still understand and improve upon it; please don't judge me harshly for it.

Summary: I am generally dismayed that many people seem to think or assume that only three levels of social metacognition matter ("Alex knows that Bailey knows that Charlie knows X"), or otherwise seem generally averse to unrolling those levels. This post is intended to point out (1) how the higher levels systematically get distilled and chunked into smaller working memory elements through social learning, which leads to *emotional* tracking of phenomena at 6 levels of meta and higher, and (2) what I think this means about how to approach conflict resolution.

Epistemic status: don't take my word for it; conceptual points intended to be fairly self evident upon reflection; actual techniques not backed up by systematic empirical research and might not generalize to other humans; all content very much validated by my personal experiences with talking to people about feelings in real life.

Related Reading: Duncan Sabien on Common knowledge & Miasma; Ben Pace on The Costly Coordination Mechanism of Common Knowledge

---

I. Conceptual introduction, by example

Here's how higher levels of social metacognition get distilled down and represented in emotions that end up tracking them (if poorly). Each feeling in the example below will be followed by an unrolling of the actual event or events it is implicitly tracking or referring to.

Warning: reading this first section (I) will require a fair bit of symbolic reasoning/thinking, so you might find it tiring and prefer to skip to later sections. A better writing of this section would do more work in between these symbolic reasoning bits to distill things out and make them easier to digest.

Scale 1: **One event, four levels of meta** (yes, we're starting with four)

1.1) Alex leaves out the milk for 5 minutes

1.2) Bailey observes (1.1), and feels it was bad.

Unrolling of referents: Bailey felt that Alex leaving out the milk was bad.

1.3) Alex observes (1.2), and feels judged.

Unrolling of referents: Alex felt that Bailey felt that Alex leaving out the milk was bad.

1.4) Alex reflects on feeling judged, doesn't like it, and concludes that Bailey is "a downer".

Unrolling of referents: Alex felt it was bad that Alex felt that Bailey felt that Alex leaving out the milk was bad.

Notice that the unrollings look and sound very different from the distillations. That's in large part because the unrolling is not our native format for storing social metacognition; it's stored via concepts like "feeling judged" or "being a downer". However, to the extent that the feeling "Bailey is a downer" is tracking something in reality, it's tracking things that track things that track things that track reality: in this case, milk spoilage.

(An aside: notice also that 1.4 involves Alex's feelings about Alex's feelings. Some people wouldn't call that an extra level of social metacognition, and would just combine it all together into "Alex's feelings". However, I'm separating those layers for two reasons: (1) the separation in counting won't affect my conclusion that the total number of levels being implicitly tracked greatly exceeds three, and (2) I think it's especially important to note when people have feelings about their own feelings, as that can lead to circular definitions in what their feelings are tracking; but that's a topic for another day.)

Scale 2: multiple events, six levels of meta

I'll start the numbering at 4 here:

2.4) Multiple similar Scale 1 events happen where Alex does something X, and ends up feeling that Bailey was "a downer" about it.

Partial unrolling of referents: Alex feels that Bailey is often a downer

Complete unrolling of referents: Alex felt it was bad that Alex felt that Bailey felt that Alex doing X was bad, for multiple values of X.

2.5) Charlie observes Alex treating Bailey like "a downer", thinks this is baseless, and feels Alex is "a snob".

Partial unrolling of referents: Charlie felt it was bad that Alex felt that Bailey is a downer.

2.6) Bailey observes Charlie's opposition to Alex's snobbiness, feels socially included by Charlie, and concludes that Charlie is "protective & welcoming."

Partial unrolling of referents: Bailey felt it was good that Charlie felt it was bad that Alex was feeling that Bailey was often a downer.

Complete unrolling of referents: Bailey felt it was good that Charlie felt it was bad that Alex felt it was bad that Alex felt that Bailey felt it was bad that Alex did X, for multiple values of X.

If you count the (+/-) signs implicit in the "good" and "bad" judgements here, they suggest that Charlie is implicitly condemning Alex's multiple "X" behaviors (perhaps among other things, such as Alex's style social of delivery). Charlie might not intend or even be aware of this effect, and it might take explicit work and discourse for the group to untangle and notice. What could result from the group reflecting upon this together? Well, it's not hard at all to imagine that Charlie, upon haring about the milk and other situations (values of "X"), might conclude that Alex's behavior with the milk was reasonable, and that "Alex was right to think that Bailey was being a downer". In

turn, this could burst Bailey's bubble of social support, and result in Bailey having a change of heart about how critical to be of others.

The particular consequences here are of course hypothetical, and things could go very differently depending on the details; it's just meant to illustrate how "changes of heart" can propagate through unrollings of social metacognition.

ETA (8/27): Note there is an important distinction here between implicit and explicit metacognition. Bailey alone is not (necessarily) loading up a 5-layer-deep cognitive model of what's going on, all at once. Rather, the layers are distributed across people, whence the term "*social* metacognition". However, six levels of metacognition *really are needed* for someone to aiming to ground out all these feelings in "object-level" reality (i.e., non-mental phenomena like milk spoilage).

(This is the end of the tiring symbolic reasoning section.)

II. Higher levels of meta

Without going into further explicit detail, I hope you can see the pattern. Levels of social metacognition get distilled into simple, repeated concepts like "feeling judged", "being a downer", "being a snob", "being welcoming", and so on. To the extent that these distilled concepts behave in a *somewhat* systematic manner in relation to reality, they have *some* tendency to be actually tracking things that are actually happening. It's not uncommon for me to observe six levels of social metacognition in a given disagreement or conflict, which is why I chose six for this post.

III. People don't usually unroll things this way. Why?

Unfortunately, I think a lot of people aren't *aware* that it even *makes sense to try* to ground out these sorts of social metacognition in more explicit terms to be reasoned and disagreed about. I think this is because it takes a lot of working memory slots to do, such that you basically need a shared piece of paper, whiteboard, or a shared Google doc to do it reasonably and collaboratively (rather than just slinging hard-to-unpack negative judgments at each other, adversarially, either in person or on the internet). However, I've resolved conflicts through co-writing and co-diagramming relevant levels of social metacognition many times now, and found it to be very enlightening almost every time it a way that directly benefitted the "social situation". I've found it's best if the shared writing medium is used for distillation mechanism, and is augmented by actual real-time conversation over the creation of the document.

IV. A fruitful application

For instance, in the past week, Alex (anonymized) felt judged by me for a thing I noticed Alex doing. I said, (a) "Don't worry, I don't think you did anything bad", but Alex didn't find this reassuring. To check, I asked "Do you feel like I feel like you did something bad?" and Alex said "No". This ran up against my explicit models of people feeling judged that had fit well with past unrolling of the concept. So, I broke out a Google doc (in person) and started unrolling stuff. The situation was more complicated

than described above, so the doc gave us mental space to explore other ideas for resolution. We eventually looped back to my question (a) above, and Alex said "Huh, yeah, I think I do feel that you feel that I did something bad." Once that awareness existed, I responded "Cool! Well guess what? I don't think you did anything bad.", and this time, it resonated with Alex and Alex no longer felt judged. I then apologized with "Also, I'm sorry you felt judged. Given that I didn't actually feel you were doing something bad, this was a mistake on my part, and I'm sorry."
This further cleared things up.

This whole process took about 15 minutes. In retrospect it might seem like we could have jumped straight to this solution by me saying "I'm sorry I made you feel judged", but that wasn't an available strategy ex ante, for two reasons:

(1) Sometimes I really am judging someone, and I'm okay with them feeling judged, because I do in fact think they did something wrong. As a result of this willingness in myself and others, it's not always believable to say "Sorry, I wish I hadn't made you feel judged". Indeed, to many this feels like a platitude. But, by actually going through the work of actually unrolling whether or not I thought Alex did a bad thing, and the other details of what was going on between us, we established enough shared clarity about the situation that we managed to "get on the same page" what whether a bad thing was done, who thought or didn't think that, and who miscommunicated or didn't miscommunicate about it.

(2) There were many other things going on that the Google doc helped to organize and sift through without getting us lost. Without that functionality, I don't think we would have been able to hone in on the particular narrative resolution above.

V. How generalizable is this 'unrolling' technique?

The application (IV) above is not an isolated incident. I've founding co-writing and co-drawing to be extremely valuable in settling social disagreements and conflicts on at least 30 occasions now, with at least 7 different people, of varying degrees of inclination toward explicit symbolic reasoning. I imagine some inclination is necessary, but much less than I would have expected previously. For instance, I've used this sort of unrolling heuristic fruitfully in numerous conversations with folks close to me who

(1) didn't go to college or otherwise study a symbolic discipline like math or linguistics, but who

(2) were generally open-minded enough to be willing to try out a "weird conflict resolution technique I'm experimenting with" where we sat down together and tried unpacked our feelings in explicit terms in a common medium (usually a Google doc).

I'll defer to the finding of the broader community here to see if others can make this sort of thing work usefully.

VI. Relation to "miasma" and "hype"

The concept of "miasma" that Duncan is gesturing at in Common Knowledge and Miasma feels like a real social phenomenon to me, succinctly definable as

"negative ungrounded social metacognition". There is such a thing as positive ungrounded social metacognition, as well, which I think is normally called "hype", at least in Silicon Valley. I think both hype and miasma are failures of group coordination, and both are costly to resolve, along the lines pointed out by Ben in The Costly Coordination Mechanism of Common Knowledge. However, the communicative costs of resolving these problems can be significantly decreased if people are aware of what they are. Both require creating and sharing of ideas in places of common view, like writing blog posts that a lot of people see each other commenting on, or holding meetings that a lot of people can see each other attending, or for complex topics, sitting together and co-authoring a document.

VII. Apology

I'm sorry I put very little effort into the pedagogy of this post, due to having too little time to write it. Hopefully it will be of some value anyway, due to much better posts having been written and circulated on common knowledge recently, and due to the general intellectual health of LessWrong appearing, to me, to be able to absorb mediocrely-explained ideas and flesh them out into better ones. My sense is that the culture here has been trying to move towards people not waiting until an idea is perfectly elaborated before starting to talk about it, so to the extent these ideas might be valuable, I'm punting to the community to do more elaboration and/or distillation of them. Indeed, wishing not to be a part of a "common knowledge breakdown" problem is one reason I time-boxed two hours to write this post instead of waiting to improve it.

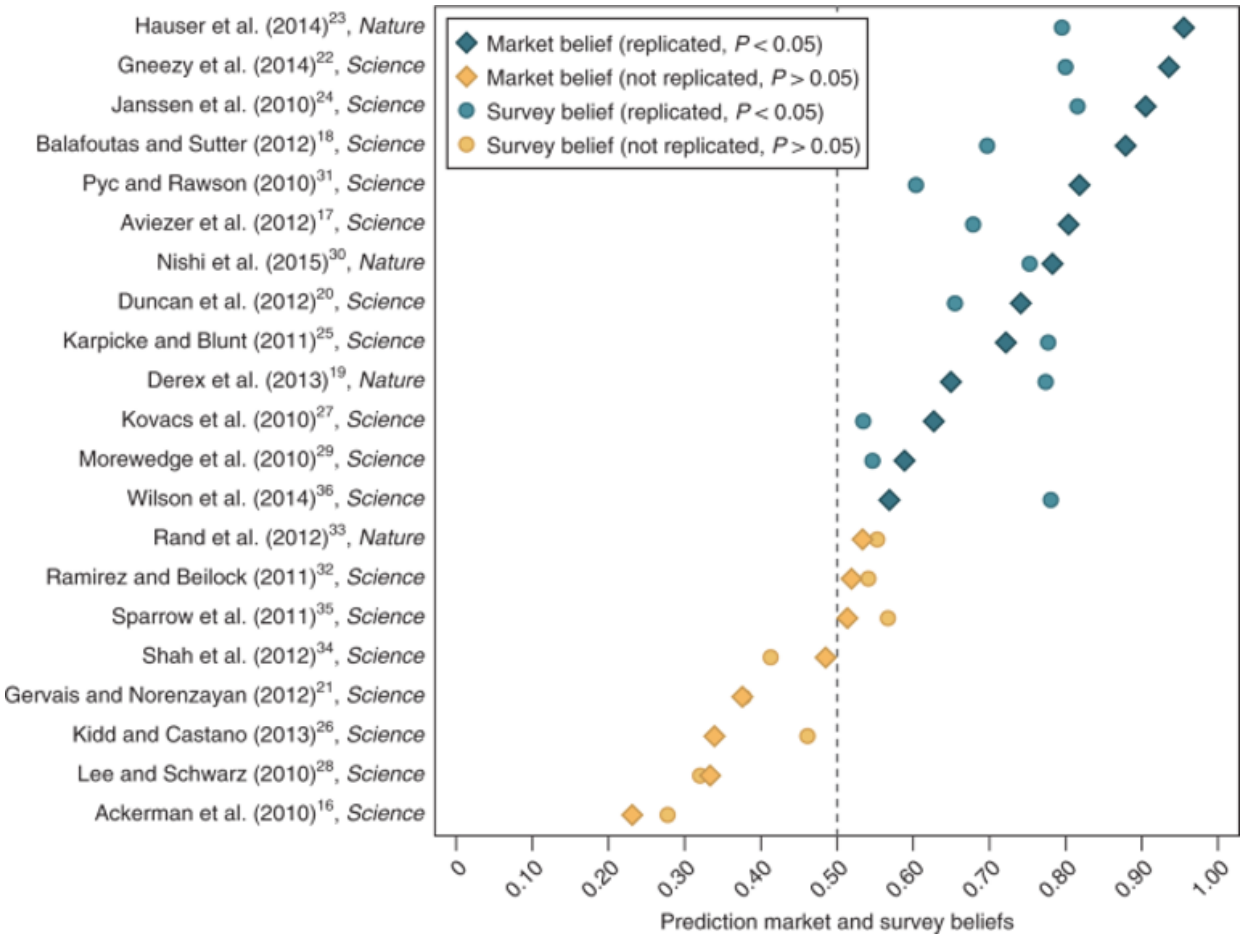
Unknown Knowns

Previously (Marginal Revolution): Gambling Can Save Science

A study was done to attempt to replicate 21 studies published in *Science* and *Nature.*

Beforehand, prediction markets were used to see which studies would be predicted to replicate with what probability. The results were as follows (from the original paper):

Fig. 4: Prediction market and survey beliefs.



The prediction market beliefs and the survey beliefs of replicating (from treatment 2 for measuring beliefs; see the Supplementary Methods for details and Supplementary Fig. 6 for the results from treatment 1) are shown. The replication studies are ranked in terms of prediction market beliefs on the *y* axis, with replication studies more likely to replicate than not to the right of the dashed line. The mean prediction market belief of replication is 63.4% (range: 23.1–95.5%, 95% CI = 53.7–73.0%) and the mean survey belief is 60.6% (range: 27.8–81.5%, 95% CI = 53.0–68.2%). This is similar to the actual replication rate of 61.9%. The prediction market beliefs and survey beliefs are highly correlated, but imprecisely estimated (Spearman correlation coefficient: 0.845, 95% CI = 0.652–0.936, $P < 0.001$, $n = 21$). Both the prediction market beliefs (Spearman correlation coefficient: 0.842, 95% CI = 0.645–0.934, $P < 0.001$, $n = 21$) and the survey

beliefs (Spearman correlation coefficient: 0.761, 95% CI = 0.491–0.898, $P < 0.001$, $n = 21$) are also highly correlated with a successful replication.

That is not only a super impressive result. That result is *suspiciously* amazingly great.

The mean prediction market belief of replication is 63.4%, the survey mean was 60.6% and the final result was 61.9%. That's impressive all around.

What's far more striking is that *they knew exactly which studies would replicate.* Every study that would replicate traded at a higher probability of success than every study that would fail to replicate.

Combining that with an almost exactly correct mean success rate, we have a stunning display of knowledge and of under-confidence.

Then we combine that with this fact from the paper:

> Second, among the unsuccessful replications, there was essentially no evidence for the original finding. The average relative effect size was very close to zero for the eight findings that failed to replicate according to the statistical significance criterion.

That means there was a clean cut. Thirteen of the studies successfully replicated. Eight of them not only didn't replicate, but showed very close to no effect.

Now combine these facts: The rate of replication was estimated correctly. The studies were *exactly* correctly sorted by whether they would replicate. None of the studies that failed to replicate came close to replicating, so there was a 'clean cut' in the underlying scientific reality. Some of the studies found real results. All others were either fraud, p-hacking or the light p-hacking of a bad hypothesis and small sample size. No in between.

The implementation of the prediction market used a market maker who began anchored to a 50% probability of replication. This, and the fact that participants had limited tokens with which to trade (and thus, had to prioritize which probabilities to move) explains some of the under-confidence in the individual results. The rest seems to be legitimate under-confidence.

What we have here is an example of that elusive object, the *unknown known:* Things we don't know that we know. This completes Rumsfeld's 2×2. We *pretend* that we don't have enough information to know which studies represent real results and which ones don't. We are modest. We don't fully update on information that doesn't conform properly to the formal rules of inference, or the norms of scientific debate. We don't dare make the claim that we know, even to ourselves.

And yet, we know.

What else do we know?

Historical mathematicians exhibit a birth order effect too

[Epistemic status: pilot study. I'm hoping that others will help to verify or falsify my conclusion here. I've never done an analysis of this sort before, and would appreciate correction of any errors.

A previous version of this post has some minor errors in the analysis, which have since been corrected. Most notably, deviation from expected rate of first borns was originally noted as 14.98 percentage points. It is actually 16.65 percentage points.]

A big thank you to Dan Keys for working through the statistics with me.

Follow-up to: Fight Me, Psychologists, Birth Order Effects are Real and Very Strong, 2012 Survey Results

Since the late 1800's, pop psychology has postulated that a person's birth order (whether one is the first, last, middle, etc. of one's siblings) has an impact on his/her lifetime personality traits. However, rigorous large-scale analyses have reliably found no significant effect on stable personality, with some evidence for a small effect on intelligence. (The Wikipedia page lists some relevant papers on birth order effects on personality (1, 2, 3) and on intelligence (1, 2, 3).)

So, we were all pretty surprised when, around 2012, survey data suggested a *very strong* birth order effect amongst those in the broader rationality community.

The Less Wrong community is demographically dominated by first-borns: a startlingly large percentage of us have only younger siblings. On average, it looks like there's about a twenty-two percentage point difference between the actual rate of first borns and the expected rate, from the 2018 Slate Star Codex Survey data Scott cites in the linked post above. (More specifically, the expected rate of first-borns is 39% and the actual occurrence in the survey data is 62%.) The 2012 Less Wrong survey also found a 22 percentage point difference. This effect is highly significant, including after taking into account other demographic factors.

A few weeks ago, Scott Garrabrant (one of the researchers at MIRI) off-handedly wondered aloud if great mathematicians (who plausibly share some important features with LessWrongers), also exhibit this same trend towards being first born.

The short answer: Yes, they do, as near as I can tell, but not as strongly as LessWrongers.

My data and analysis is documented here.

Methodology

Following Sarah Constantin's fact post methodology, I started by taking a list of the 150 greatest mathematicians from here. This is perhaps not the most accurate or scientific ranking of historical math talent, but in practice, there's enough broad agreement about who the big names are, that quibbles over who should be included are mostly irrelevant to our purpose. If a person could *plausibly* be included on a list of the greatest 150 mathematicians *in history*, he/she was probably a pretty good mathematician.

I then went through the list, and tried to find out how many older and younger siblings each mathematician had. For the most part this amounted to googling "[mathematician's name] siblings" and then trawling through the results to find one that gave me the information I wanted. Where possible, I noted not just the birth order and number of siblings, but also the

sex of the siblings and whether they died during infancy. (For the ones for whom I couldn't get data, I marked the row as "Couldn't find" or "Unknown")

Most biographical sources don't list the number of siblings of the family of origin. The sources that I ended up relying on the most were:

- Geni.com: a platform for people to build out their family trees, and store historical or biographical information (family photos, dates of life events, etc.)
- TheFamousPeople.com
- The individual biographies on the MacTutor History of Mathematics Archive

This was a very quick cursory search, so my data is probably not super reliable. At least twice, I found two sources that disagreed, and I don't know how much I would have encountered conflicting information if I had dug deeper into each person's biography, instead of moving on to the next mathematician as soon as I found a sentence that answered my query.

If you happen to personally know biographical details of elite mathematicians and you can correct any errors in these data, I'd be pleased to make those corrections.

Results

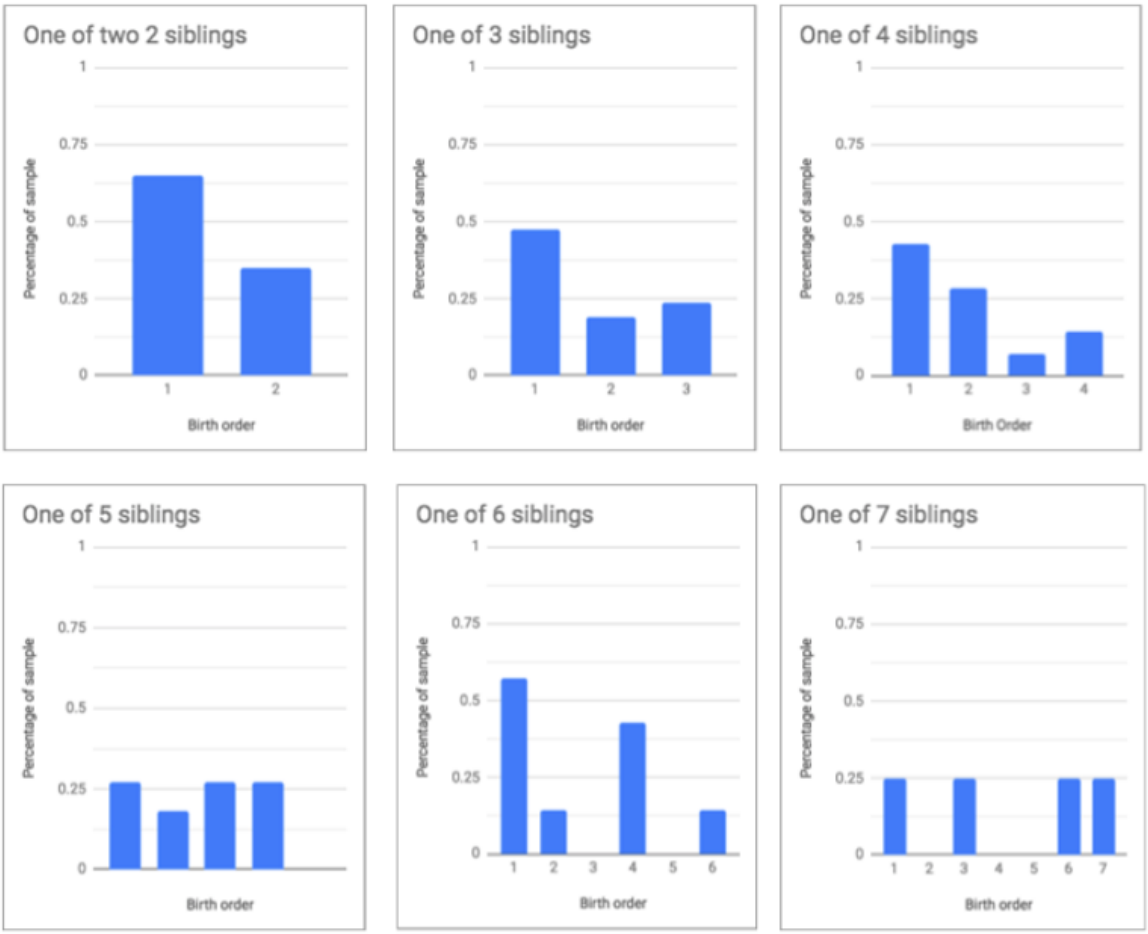
The simplest analysis is to categorize the data by family size (all the mathematicians that had no siblings, one sibling, 2 siblings, etc.), count how many first borns there were in each bucket, and compare that to the number we would expect by chance.

For nearly every bucket, the frequency of first born children exceeded random chance. Across all categories, the difference in percentage points between the actual and expected frequencies was about 16.5%.

Number of children in a family	Number families of that size	Actual number first born	Expected number first born	Actual > or < expected?	Actual minus expected	actual % first born	expected % firstborn	Percentage point difference (actual % minus expected %)
1	17	17	17	=	0	100.00%	100.00%	0.00%
2	20	13	10	>	3	65.00%	50.00%	15.00%
3	19	10	6.333333333	>	3.666666667	52.63%	33.33%	19.30%
4	13	6	3.25	>	2.75	46.15%	25.00%	21.15%
5	11	3	2.2	>	0.8	27.27%	20.00%	7.27%
6	7	4	1.166666667	>	2.833333333	57.14%	16.67%	40.48%
7	4	1	0.5714285714	>	0.4285714286	25.00%	14.29%	10.71%
8	2	1	0.25	>	0.75	50.00%	12.50%	37.50%
9	3	0	0.3333333333	<	-0.3333333333	0.00%	11.11%	-11.11%
10	0	0	0	=	0			
11	0	0	0	=	0			
12	2	0	0.1666666667	<	-0.1666666667	0.00%	8.33%	-8.33%
13	1	0	0.07692307692	<	-0.07692307692	0.00%	7.69%	-7.69%
Total:	2 or more	82	38	24.34835165	0	13.65164835	46.34%	29.69%	16.65%

After removing the individuals that I couldn't find data for, we had a sample size of 82. A paired t-test, comparing the number of first-borns with the expected number of first-borns (one data point for each of the 82 mathematicians) was statistically significant, $t(81)=3.14$, $p = 0.00239$.

I can show you some bar graphs, like Scott uses in his post, but because this data is of a much smaller sample and the effect isn't as large, they don't look as neat. (Also, I don't know how to include those nice dotted lines marking the expected frequency.)



Nevertheless, you can see a systematic trend: being the first of *n* siblings is overrepresented among the mathematicians in the sample I used.

The effect in these data (17 percentage points) is smaller than the effect in either the Less Wrong or Slate Star Codex surveys (22 percentage points). The 95% confidence interval for the mathematician data is a range of 6 percentage points to 27 percentage points. Given this range, we can't rule out that the difference in effect sizes is due to noise, but it seems most plausible that there is a real difference in the size of the underlying effect between the populations.

A discussion of bias in this data

As I say, my data is not very reliable, it seems plausible that some of my sources were faulty, and I was going quickly, so I may have made some errors in doing data collection. Furthermore, I was only able to find data for 82 of the 150 mathematicians.

But in expectation, those errors will cancel out, unless there's some systematic bias in the sources I was using. I can think of at least two causes of bias, but neither one seems like it could be the cause of the observed trend.

Higher reporting rate for first born children

First, maybe first borns are recorded more readily? If the first born child was the heir to a family's property, then they may have been more likely to be mentioned in legal and other

documents, so there may be much better historical records of first-born children.

But our subjects are all famous mathematicians, independent from their inheritance-status. So, if there was a historical reporting bias that favored the first born, this would actually push *against* our observed effect. First born members of our sample would either be listed as only children, or noted as having an unknown number of siblings. Younger-sibling mathematicians, on the other hand, would be noted as younger siblings, because their older brother is added to the historical record on the basis of their heirship.

Underreporting of females

Another way in which the available record of sibling data may be biased, which does not directly affect the validity of this analysis, is that women might have gone unrecorded more often than men. The size of this effect tells us something about the extent to which the available record of sibling data is biased.

It was relatively easy to do a quick check for a reporting bias in favor of male siblings: I just summed all the brothers that I found, and all the sisters.

All together, I recorded 110.5 brothers and half brothers and 100.5 sisters and half sisters. (The point five comes from Jean-Baptiste Joseph Fourier's entry. I found that he had 3 half siblings by his father's first marriage, but I didn't know of what sex. So I split the difference by saying he had 1.5 half brothers and 1.5 half sisters, in expectation. I was comfortable doing this because I mostly care about whether siblings are younger or older, and only secondarily about if they are male or female.)

So there are slightly more males listed, at least in the sources I could find. But a difference of 10 out of 211 siblings with recorded sex, isn't very large. I'm sure there are some statistics I could do to show it, but I don't think that slight bias is sufficient to account for our observed birth order effect.

I'm hoping that others can think of reasons why we might see a trend in these data even if the birth order effect wasn't real.

Conclusion

This is a pretty intriguing result, and I'm surprised no one (that I know of) has noticed it before now.

I think this post should be thought of as a pilot study. I put in about 20 hours to investigate the hypothesis, but only in a quick and cursory way. I would be excited for others, who are better informed and better-equipped than I am, to do a more in-depth analysis into these topics.

Do mathematicians of lesser renown display this birth order effect? What about prominent (or average) individuals from other STEM fields? Non-STEM fields? I'd be interested to see an analysis of the most successful business executives, for instance.

Furthermore, more investigation could uncover detail about *how* having older siblings gives rise to this effect.

Some explanations for this phenomenon rest on social interaction with older siblings in one's first few years. Others depend on biological consequences of spending one's fetal period in a womb that was previously occupied by older siblings. In principle we should be able to tease out which of these mechanisms generates the effect by looking at much more data that tracks older siblings that died in infancy, and older half siblings. (Siblings that died in infancy can't mediate the social effect, while half siblings can mediate a biological effect depending on which parent is shared, and can mediate a social effect depending on whether they were

living in the household at the time of birth.) If someone found a larger dataset that tracked these factors, we might be able to falsify one or the other of these stories.

And again, please inform me of any errors.

Y Couchinator

Crossposted to Tumblr.

This project never really had critical mass and is no longer operative; it did later inspire another project, also now defunct.

There are a lot of people - there are probably incredibly tragic mountains of people - who just need one or three or six no-pressure months on someone's couch, and meals during that time, and then they'd be okay. They'd spend this time catching up on their bureaucracy or recovering from abuse or getting training in a field they want to go into or all three. And then they'd be *fine*.

There are empty couches, whose owners throw away leftovers they didn't get around to eating every week, who aren't too introverted to have a roommate or too busy to help someone figure out their local subway system.

And while sometimes by serendipity these people manage to find each other and make a leap of trust and engage in couch commensalism a lot of the time they just don't. Because six months is a long time, a huge commitment for someone you haven't vetted, and a week wouldn't be enough to be worth the plane ticket, not enough to make a difference.

I think there might be a lot of gains to be had from disentangling the vetting and the hosting. People are comfortable with different levels of vetting, ranging from "they talked to me enough that it'd be an unusually high-effort scam" through "must be at least a friend of a friend of a friend" through "I have to have known them in person for months". And you can bootstrap through these.

Here's a toy example:

- Joe barely makes ends meet somewhere out in flyover country but in between shifts at his retail job he's doing well at self-teaching programming and seems like he could pass App Academy.
- Norm has a one-bedroom apartment in San Francisco and doesn't really need his couch to be empty, nor does he need to rent it out for money to someone desperate enough to pay rent on a couch, but he's not ready to give some dude on the internet a commitment to providing shelter for the duration of App Academy.
- Tasha has a house somewhere in visiting distance of Norm, maybe even several different people like Norm, and she too has a couch, and is willing to host an internet dude for one week based on a sad blog post. During this week, Joe and Norm meet, and Tasha evaluates Joe's suitability as a guest, and Norm decides he can commit to have Joe as an occupant for the duration of App Academy.

My household has been informally couching (or bedrooming, as the case may be) itinerants for a while. Sometimes they get jobs and move out, or get jobs and don't move out. Sometimes they find other households they fit into better and move into those. Sometimes they wind up staying for a while and not really improving their prospects and going back whence they came; this is just the sort of thing that happens sometimes.

And I think more people could accommodate this fine from the hosting end, and just don't have the networking to find would-be couch occupants on a routine basis.

I propose a minimum viable product, low tech, Y Couchinator, to gauge demand and work out kinks before I try to make a technical person build me a website and expose us to liability and all that exciting stuff. Here's how I'm imagining it will work.

- If you have a couch, you tell me about your couch. Is it available for a specific week in June for people you talk to for two hours first and like a lot? Is it available for one month to nonsmoking vegetarian afabs who your landlord might believe are your cousin? Is it available to anybody for any portion of the academic summer as long as they can walk your dog and don't seem inordinately sketchy to me when I hear about them? Is it available for six months if they can cover groceries and get a reference from somebody who has hosted them for at least a fortnight? Please be prepared to really maintain these boundaries when you need them even once you are presented with an actual couch occupant who has a sob story, even if it's a really sobsome story. We've never had a serious problem with this, but it's the sort of thing that could happen. (Of course, distinguish "not enforcing a boundary" from "liked person more than expected, happy to keep them longer than I committed to based on less information".)
- If you need a couch, you tell me about your couch needs. Does it have to actually be a bed, not a couch? Do you need to bring your gerbil? Are you deathly allergic to peanuts/children/cats/country music/Brutalist architecture? And you tell me about your plans for your couch time. This is a somewhat constrained offer, so there do have to be plans. I want to match people whose couch needs are plausibly likely to be self-limiting and don't come with a need for cash in particular, at least unless I find that there are many more couches than occupants under this condition. Do you have a prospect for getting some kind of job as long as you can park in the right city for a while to attend interviews? Do you have a plan for some kind of job training, like the example of App Academy or something less classic? Are you pretty employable already but only when you have your mental health under control, and just need some time not relying on your parents for survival in order to get there?
- I collect couch needers and couch havers. I match people who can be straightforwardly matched, and I do my best to line up dominoes when several couches have to be strung together ("Bill, you can stay with Haley for a week, and she'll introduce you to Penelope, and if Penelope likes you you can stay with her for three months, and if Penelope doesn't feel comfortable with that then Wilbur is willing to subsidize you in REACH for a week and a half until Leroy's place opens up, which is yours for two months as long as you do dishes and sometimes give him rides to the airport").

I want to emphasize here that there *do exist* people who have couches they would be willing to offer. This came up in a Discord chat and two people I didn't even know about before mentioned that under certain constraints they could offer couches. My own household (which currently contains three different people who have at different times lived with us, paid no rent, and stayed on to the present day, and that's if you don't count my toddler) can fit people short-term, medium if we really click.

How to get ahold of me: You can put out initial feelers via Tumblr ask (I will assume this post is not getting enough circulation until at least 3 depressed anons have wondered at me whether they really deserve couches) but it is a bad way to do

anything long form. My email address is alicorn at elcenia (dot com). If you share any Discord servers with me, Discord works great too.

Alignment Newsletter #21

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

Highlights

80K podcast with Katja Grace *(Katja Grace and Rob Wiblin)*: Rob Wiblin interviewed Katja Grace of AI Impacts about her work predicting the future of AI. My main takeaway was that there are many important questions in this space that almost no one is trying to answer, and that we haven't made a good enough attempt yet to conclude that it's too hard to do, so we should put more time into it. If you haven't seen AI Impacts' work before, you can get some of the most interesting results (at a high level) from listening to this podcast. There's a ton of detail in the podcast -- too much for me to summarize here.

My opinion: I don't currently think very much about timelines, intelligence explosions, and other questions that AI Impacts thinks about, but it seems very plausible to me that these could be extremely important. (I do think about discontinuities in progress and am very glad I read the AI Impacts post on the subject.) One point that the interview brings up is that there are very few (perhaps two?) full time equivalents working on predicting the future of AI, while there are many people working on technical AI safety, so the former is more neglected. I'm not sure I agree with this -- the number of full time equivalents doing technical AI alignment research seems quite small (on the order of 50 people). However, I do see *many* people who are trying to skill up so that they can do technical AI alignment research, and none who want to do better prediction, and that seems clearly wrong. I would guess that there are several readers of this newsletter who want to do technical AI alignment research, but who would have more impact if they worked in an adjacent area, such as prediction as at AI Impacts, or policy and strategy work, or in better tools and communication. Even though I'm well-placed to do technical research, I still think that common knowledge of research is a big enough bottleneck that I spend a lot of time on this newsletter. It seems likely that there is someone else who would do a better job than me, but who is set on technical safety research even though they wouldn't be as good. So I guess if you are still trying to figure out how to best help with AI alignment, or are about to start training up to do technical research, please do listen to this podcast and consider that alternative route, and various others as well. The goal is not to figure out which question is the most important, so that you can try to solve it. You'll likely do better by considering the field as a whole, and asking which area you would be in if someone optimally assigned people in the field to tasks.

Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review *(Sergey Levine)*: I sent this out as a link in AN #5, but only just got around to reading it. This paper shows how you can fit the framework of reinforcement learning into the framework of inference within probabilistic graphical models. Specifically, the states s_t and actions a_t are now represented as nodes in the graphical model, and we add in new nodes O_t that represent whether or not an "event" happened at time t. By assigning the values of P(O_t | s_t, a_t) appropriately, we can encode a reward function. Then, by conditioning on the rewarding events happening, we can infer what actions must have been taken to get these events, which gives us a policy that achieves high reward. They later talk about the connection to variational inference, and how you can get IRL methods in this framework.

My opinion: Remarkably, this paper is both heavy on (useful) math, and very clear and well-explained. I actually didn't try to explain the technical details in my summary as much as I usually do, because you can just read the paper and actually understand what's going on, at least if you're familiar with probabilistic graphical models. Regarding the content, I've found the framework useful for one of my current projects, so I do recommend reading it.

Safety-first AI for autonomous data centre cooling and industrial control *(Amanda Gasparik et al)*: Two years ago, DeepMind built an AI recommendation system that provided suggestions on how best to cool Google's data centers, leading to efficiency gains. Nine months ago, the AI was given autonomous control to take actions directly, rather than going through human operators, and it has been improving ever since, going from 12% savings at deployment to 30% now.

Of course, such a system must be made extremely reliable, since a failure could result in Google's data centers going down. They implemented several safety measures. They throw out any actions that the AI is not confident about. All actions are verified against a set of hand-coded safety rules, both when the actions are generated in the cloud, and at each local data center, for reliability through redundancy. There are human operators monitoring the AI to make sure nothing goes wrong, who can take over control whenever they want to. There is also an automated system that will fall back to the original system of heuristics and rules if the safety conditions are ever violated.

My opinion: This is a remarkable number of safety precautions, though in hindsight it makes total sense given how bad a failure could be. None of the precautions would stop a superintelligent agent in the classical sense (that is, the sort of superintelligent agent in paperclip maximizer stories), but they seem like a really good set of precautions for anything task-based. I am curious how they chose the threshold for when to discard actions that the AI is not confident enough in (especially since AI uncertainty estimates are typically not calibrated), and how they developed the safety rules for verification (since that is a form of specification, which is often easy to get wrong).

Technical AI alignment

Agent foundations

Reducing collective rationality to individual optimization in common-payoff games using MCMC *(jessicata)*: Given how hard multiagent cooperation is, it would be great if we could devise an algorithm such that each agent is only locally optimizing their own utility (without requiring that anyone else change their policy), that still achieves the globally optimal policy. This post considers the case where all players have the same utility function in an iterated game. In this case, we can define a process where at every timestep, one agent is randomly selected, and that agent changes their action in the game uniformly at random with probability that depends on how much utility was just achieved. This depends on a rationality parameter α -- the higher α is, the more likely it is for the player to stick with a high utility action.

This process allows you to reach every possible joint action from every other possible joint action with some non-zero probability, so in the limit of running this process forever, you will end up visiting every state infinitely often. However, by cranking up

the value of α, we can ensure that in the limit we spend most of the time in the high-value states and rarely switch to anything lower, which lets us get arbitrarily close to the optimal deterministic policy (and so arbitrarily close to the optimal expected value).

My opinion: I like this, it's an explicit construction that demonstrates how you can play with the explore-exploit tradeoff in multiagent settings. Note that when α is set to be very high (the condition in which we get near-optimal outcomes in the limit), there is very little exploration, and so it will take a long time before we actually find the optimal outcome in the first place. It seems like this would make it hard to use in practice, but perhaps we could replace the exploration with reasoning about the game and other agents in it? The author was planning to use reflective oracles to do something like this if I understand correctly.

Learning human intent

Shared Multi-Task Imitation Learning for Indoor Self-Navigation *(Junhong Xu et al)*

Preventing bad behavior

Safety-first AI for autonomous data centre cooling and industrial control *(Amanda Gasparik et al)*: Summarized in the highlights!

Interpretability

Learning Explanations from Language Data *(David Harbecke, Robert Schwarzenberg et al)*

Miscellaneous (Alignment)

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

Book Review: AI Safety and Security *(Michaël Trazzi)*: A review of the new AI Safety and Security book. It goes through each of the papers, giving a short summary of each and some comments (similar to this newsletter).

My opinion: I take a very different approach to AI safety, so it was nice to read a summary of what other people are thinking about. Based on the summaries, it sounds like most of the essays that focused on AGI were anthropomorphizing AGI more than I would like (though of course I haven't actually read the book).

Near-term concerns

Privacy and security

Are You Tampering With My Data? *(Michele Alberti, Vinaychandran Pondenkandath et al)*

AI capabilities

Reinforcement learning

Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review *(Sergey Levine)*: Summarized in the highlights!

The International 2018: Results *(OpenAI)*: Two human teams beat OpenAI Five at The International. The games seemed much more like regular Dota, probably because there was now only one vulnerable courier for items instead of five invulnerable ones. This meant that OpenAI Five's strategy of a relentless team attack on the enemy was no longer as powerful, because they couldn't get the health regeneration items they needed to constantly stay alive to continue the attack. It's also possible (but less likely to me) that the matches were more normal because the teams were more even, or because the human teams knew about Five's strategy this time and were countering it in ways that I don't understand.

My opinion: There are still some things that the bots do that seem like bad decisions. You can interpret this a few ways. Five could have learned a large number of heuristics that make it good enough to beat almost all humans, but that break down in edge cases. In this story, Five is not good at learning logical or abstract reasoning, but can compensate for that in the average case with the sheer number of heuristics it can learn. Another interpretation is that Five learns a good representation of Dota which lets it come up with new, novel insights into the game, which we can't see or understand because the representation is alien to us. However, the representation makes it harder to come up with other insights about Dota that we have using our representations of Dota, and as a result Five makes some mistakes that humans can easily recognize as mistakes. I lean towards the first interpretation, but not very strongly.

Deep learning

Skill Rating for Generative Models *(Catherine Olsson et al)*

Neural Architecture Search: A Survey *(Thomas Elsken et al)*

Analyzing Inverse Problems with Invertible Neural Networks *(Lynton Ardizzone et al)*

Unsupervised learning

Life-Long Disentangled Representation Learning with Cross-Domain Latent Homologies *(Alessandro Achille et al)*

Learning deep representations by mutual information estimation and maximization *(R Devon Hjelm et al)*

Miscellaneous (Capabilities)

Winner's Curse? *(D. Sculley et al)*: A short paper arguing that we need more empirical rigor in ML, identifying some structural incentives that push against this and suggesting solutions.

My opinion: While this isn't very relevant to technical alignment, it does seem important to have more rigor in ML, since ML researchers are likely to be the ones building advanced AI.

News

DeepMind job: Science Writer: According to the job listing, the role would involve creating content for the blog, videos, presentations, events, etc. and would require a reasonably technical background and strong writing skills. Vishal Maini at DeepMind notes that this person would likely have a significant impact on how AI research is communicated to various key strategic audiences around the world -- from the technical community to the broader public -- and would spend some of their time engaging with AI alignment research, among other areas.

Internship: The Future Society *(Caroline Jeanmaire)*: An internship which will focus on AI policy research as well as support to organize two large AI governance events. To apply, send a CV and a short letter explaining ‘why you?’ to caroline.jeanmaire@thefuturesociety.org.

Do what we mean vs. do what we say

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

Written quickly after a CHAI meeting on the topic, haven't thought through it in depth.

If we write down an explicit utility function and have an AI optimize that, we expect that a superintelligent AI would end up doing something catastrophic, not because it misunderstands what humans want, but because it doesn't care -- it is trying to optimize the function that was written down. It is doing what we said instead of doing what we meant.

An approach like Inverse Reward Design instead says that we should take the human's written down utility function as an observation about the true reward function, and infer a distribution over true reward functions. This agent is "doing what we mean" instead of doing what we said.

This suggests a potential definition -- in a "do what we mean" system, the thing that is being optimized is a latent variable, whereas in a "do what we say" system, it is explicitly specified. Note that "latent" need not mean that you have a probability distribution over it, it just needs to be hidden information. For example, if I had to categorize iterated distillation and amplification, it would be as a "do what we mean" system where the thing being optimized is implicit in the policy of the human and is never made fully certain.

However, this doesn't imply that we want to build a system that exclusively does what we mean. For example, with IRD, if the true reward function is not in the space of reward functions that we consider (perhaps because it depends on a feature that we didn't have), you can get arbitrarily bad outcomes (see the problem of fully updated deference). One idea would be to have a "do what we mean" core, which we expect will usually do good things, but have a "do what we say" subsystem that adds an extra layer of safety. For example, even if the "do what we mean" part is completely sure about the human utility function and knows we are making a mistake, the AI will still shut down if we ask it to because of the "do what we say" part. This seems to be the idea in MIRI's version of corrigibility.

I'd be interested to see disagreements with the definition of "do what we mean" as optimizing a latent variable. I'd also be interested to hear how "corrigibility" and "alignment" relate to these concepts, if at all. For example, it seems like MIRI's corrigibility is closer to "do what we say" while Paul's corrigibility is closer to "do what we mean".

Trust Me I'm Lying: A Summary and Review

This is a linkpost for http://quanticle.net/reviews/trust_me_im_lying.html

Trust Me I'm Lying, by Ryan Holiday, is probably the most influential book I've read in the past two years. This book is a guidebook to the twenty-first century media ecosystem, and shows how the structures and incentives of online media serve to polarize us by stoking fear and anger. First published in 2012, the book is eerily prophetic in many places, as it talks about the toxic influence of online publishing on politics and society.

Ryan Holiday is a marketer and publicist who specializes in manipulating *blogs* in service of his clients. So what is a blog? A blog is any online publishing platform which derives its revenue from advertising. Blogs range in size from small, fly-by-night local publications all the way up to multi million dollar properties like *Gawker* and the *Huffington Post*. Blogs may be independent, such as *Huffington Post* or *Politico* or they may be associated with an existing media franchise. For example the *Monkey Cage* politics blog is hosted by the Washington Post, and *MoneyWatch* is a financial blog hosted by CBS.

The key observation that Ryan makes is that all blogs, no matter how large or small they are, no matter if they're associated with an existing media franchise or not, are driven by the same economic incentives. The revenue a blog makes can be expressed as (price per page-view) × (number of page-views). Since blogs have little control over how much they are paid per reader, blogs universally tend to work to maximize the number of readers they get. Moreover the sheer number of blogs, and the extremely low barrier to entry creates a brutally Darwinian marketplace where any content that doesn't maximize page-views is rapidly superseded by content that does.

In order to maximize the number of people clicking on their stories (and thus viewing the ads on those stories), blogs exploit every flaw in human psychology they can find. Chief among these is provocation. Ryan cites a study by Berger and Milkman from 2012 which shows that content with high emotional valence spreads much faster than content which is emotionally neutral. The study compared stories on the New York Times website, and found that articles which induced anger were 34% more likely than the median article to make the top-10 most e-mailed list. Articles that induced awe also did well, being 30% more likely than the median article to make the most e-mailed list. Both anger and awe are high-arousal emotions (in a negative and positive direction, respectively). On the flip side, articles that induce low-arousal emotions, like sadness, suffer a penalty. Sad articles were 16% *less* likely than the median article to end up on the most e-mailed list. These facts about human psychology act as constraints on the kinds of stories blogs will write. Every story has to make people feel a "high-energy" emotion, like anger, or awe. Stories that are thoughtful, practical, useful or beautiful but melancholy fall by the wayside.

Another constraint on blogs is their very structure. Marshall McLuhan's adage, "the medium is the message" applies just as much to blogs as it does to television. For blogs, the medium is the *stack*. Most blogs are arranged in a reverse-chronological list, with new stories coming in at the top of the page, and percolating down towards the bottom as further stories come in. A blog which can always produce something

fresh at the top of the stack can draw in more readers by having more novel content for readers to click on and share. Ryan points out that, unlike newspapers, who have a finite number of column-inches per day and unlike cable news, which has a finite number of hours per day, a blog's appetite for content is functionally *infinite*. Blogs whose writers produce the fastest win, regardless of the quality of their writing.

The final constraint on blogs is in how they get new readers. In general, people don't subscribe to blogs like they subscribe to newspapers or magazines. Instead, blogs get traffic from links and headlines that are shared on link aggregators (like Reddit) or on social media. The reputation (such as it is) of the blog counts for nothing when stories are passed around as disaggregated headlines, each fighting for the reader's attention on its own. In order to "hook" readers, headlines have to be as provocative as possible. In fact, it's in the blog's interest to make headlines misleading, since a reader that clicks into a page and clicks away in disgust still counts as a page-view, which earns the blog money. And since people don't generally pay attention to the history or credibility of a source when sharing or clicking on links, the blog does not suffer any penalty for wasting the reader's time or attention.

These three constraints (virality, structure, and disaggregation) serve to create a set of entry points that allows a media manipulator such as Ryan Holiday to influence the content and framing of stories that blogs cover. Ryan noticed, contrary to prevailing wisdom, that most original reporting in online media was done by smaller blogs, whose stories were picked up and summarized by larger publications, until they reached mainstream media outlets and entered the "national conversation". Therefore, by influencing small blogs today, one could alter what was in the Washington Post tomorrow.

Ryan created a process, which he termed "trading up the chain" to do just that. First he would observe which large blogs the national media outlets he was targeting drew from. Then he would observe which small blogs those large blogs pulled stories out of. Finally, he would craft a media campaign targeting those smaller blogs, seeding the same provocative story in enough places to ensure that it would get picked up and passed up the chain until it received national coverage.

A concrete example of this is work he did for the movie *I Hope They Serve Beer In Hell*, starring Tucker Max. He wished to get coverage for the movie in the *Washington Post*. By observing the sources of the *Post's* stories carefully, he found out that much of the *Post*'s media coverage originated from stories on *Gawker*. Going one level further he noticed that *Gawker*, in turn, pulled its stories from smaller city-focused blogs like *MediaBistro* and *Curbed LA*. He then targeted those blogs with a series of provocative actions, such as buying billboards promoting the movie and then vandalizing them himself, calling feminist groups to protest showings of the movie, arranging for provocative advertisements on buses, and other acts designed to go viral on social media. When those blogs inevitably began covering *I Hope They Serve Beer In Hell*, he ensured that there was enough "chatter" to quickly drive the story up the chain to the *Washington Post*.

The end result of these tactics was that an otherwise no-name D-List "internet celebrity" like Tucker Max was being interviewed by The Washington Post and late-night TV hosts like Conan O'Brien. Ryan makes the point that much of what we consider to be "organic" content, spreading by "word of mouth" is actually carefully engineered and seeded by professionals like himself to generate coverage for particular celebrities, products, and events.

The important thing to note is that all of these manipulations make the stories that we read less accurate. Moreover, the publishers of these stories bear no cost for misleading us and wasting our time and emotional energy. The toll that the exact on each and every one of us is a pure externality, just as much as smokestack fumes or toxic chemicals going into the water supply.

If the effects of this media manipulation were merely to drive customers to products they wouldn't otherwise buy, Ryan would still probably be out there plying his trade. What caused him to reconsider his profession (and write this book) was the increasing use of these manipulation techniques to spread political ideas, and, in the process, hurt individuals. In the second half of the book, he talks about how sites like *Jezebel* and *Breitbart News* use the techniques he pioneered to push product for American Apparel to maximize their own page-views by stoking outrage both among their supporters and their opponents. In his view, much of responsibility for the coarsening and polarization of politics and culture can be laid at the feet of professional manipulators like himself.

Ryan's hope is that by writing this book and exposing the actual techniques that manipulators use, we can inoculate ourselves and make ourselves less susceptible to the sort of media manipulation that he used to carry out. Though things look bleak at the moment, Ryan looks to history to show that our online media ecosystem today is very similar to the "yellow-press" era at the turn of the 20th century. Given that, he has hope that the current state of affairs is not sustainable, and we will eventually craft a stronger, more trustworthy online media, just as the provocative tabloids like the *New York Herald* and *The World* eventually gave way to more trustworthy publications like *The New York Times* and the *Wall Street Journal*.

Personally, I found this book important because it explained and crystallized many of the troubling trends I had personally observed in online media, and put them in a framework that allowed me to see clearly how they worked and how they were manipulating me. Though I knew that online media was growing ever more provocative and ever less accurate, my prior was that it was the result of blind evolutionary forces, as described in Scott Alexander's post on the same topic, The Toxoplasma of Rage.

While Scott's piece is important and insightful, it's still written from an outsider's perspective. It frames the ever escalating spiral of provocation as the result of groups competing to stoke the most outrage among their own members. *Trust Me I'm Lying*, by contrast, says that the result of deliberate manipulation by people who deliberately set groups against one another in order to bring attention to the issues that they want attention brought to. When Ryan vandalized his own billboards and organized protests by feminist groups against *I Hope They Serve Beer In Hell*, he wasn't attempting to send a message about the content of the film. He was merely operating under the adage, "Any publicity is good publicity." Similarly, Ryan's claim is that much of the outrage we find in today's politics and culture is the result of deliberate manipulation by publications who want to drive traffic to their own sites, without necessarily caring about which side "wins".

Where the book is weakest is in its advice about where to go from here. The book ends with a note that people's time and attention are limited, and eventually people will catch on to the fact that they're being manipulated, and will start to demand higher quality reporting, rather than merely quantity. In support of this contention, he cites the evolution of print journalism, which evolved from "yellow press" tabloids to newspapers that are widely considered to be accurate and *relatively* unbiased.

However, he gives few predictions about how this will come to pass, other than noting the current media ecosystem is unsustainable and that unsustainable things cannot be sustained over the long term.

Nevertheless, I found the book insightful, entertaining, and more than slightly horrifying. As a result of the book, I can look at stories like this and better look past the manipulative elements to see how little substance there actually is to the article. As a result, I find that I'm more efficient at extracting the factual content from news articles, and better at identifying and avoiding so-called "fake news". For this reason, I consider *Trust Me I'm Lying* to be a strong recommendation.

If you're interested in a more detailed outline of the book, I have one on my wiki.

Subsidizing Prediction Markets

Epistemic Status: Sadly Not Yet Subsidized

Robin Hanson linked to my previous post on prediction markets with the following note:

> .@TheZvi reviews what one needs for a working prediction market FUNDED BY TRADERS. If someone else will sponsor/subsidize them, a lot more becomes possible. https://t.co/OgEMahI7ZS
>
> — Robin Hanson (@robinhanson) July 26, 2018

I did briefly mention subsidization, as an option for satisfying the fifth requirement: Sources of Disagreement and Interest, also known as Suckers At The Table. The ultimate sucker is an explicit, intentional one. It can serve that roll quite well, and a sufficiently large subsidy can make up for a lot. Any sufficiently large sucker can do that – give people enough profit to chase, and suddenly not being so well-defined or so quick or probable to resolve, or even being not safe from key insider information, starts to sound like it is worth the risk.

Suppose one wants to create a subsidized prediction market. Your goal presumably is to get a good estimate for the probability distribution of an event, and to do so without paying more than necessary. Secondary goals might include building up interest and a marketplace for this and future prediction markets, and getting a transparently robust result, so others or even the media are more likely to take the outcome seriously. What is the best way to go about doing this?

Before looking at implementation details, I'll look at the five things a prediction market needs.

I. Well Defined

The most cost-efficient subsidy for a market is to ensure that the market is well defined. Someone has to make sure everyone understands *exactly* what happens under every scenario, and that someone is you. Careful wording and consideration of corner cases is vital. Taking the time to do this right is a lot more efficient than throwing money at the problem, especially trying to build a system and brand over time.

If you're going to subsidize a market, step one is to write good careful rules, make sure people understand them, and to commit to making it right for everyone if something goes wrong, if necessary by paying multiple sides as if they had won. This is *potentially* quite expensive even if it rarely happens, so it's hard to budget for it, and it feels bad in the moment so often people don't pull the trigger. Plus, if you do it sometimes, people will argue for it all the time.

But if you're in it to win it, this is where you start.

II. Quick Resolution

Once you've got your definitions settled, your next job is to pay the winners quickly once the event happens. People care about this *more than you can possibly imagine.* The difference between paying out five seconds after the final play, and five minutes after the final play, is a big game to many. Make them wait an hour and they'll be furiously complaining on forums. When the outcome is certain, even if it hasn't actually happened yet, it's often a great move to pay people in advance. People love it. Of course, occasionally someone does something like pay out on bets on Hillary Clinton two weeks early, in which case you end up paying both sides. But great publicity, and good subsidy!

Another key service is to make sure your system recognizes when a profit has been locked in, or risk has been hedged, and does not needlessly tie up capital.

This one is otherwise tough to work around. If you want to know what happens twenty years from now, nothing is going to make resolving the question happen quickly. You *can* help a lot by ensuring that the market is liquid. If I buy in at 50% now, then a year from now the market is at 75% and is liquid enough, I can take my profits in one year rather than twenty. That's still a year, and it's still unlikely the price will 'catch up' fully to my new opinion by that time. It helps a lot, though.

III. Probable Resolution

It is a large feel bad, and a real expense, when capital is tied up and odds look good but then the event doesn't happen, and funds are returned. It hurts most when you've pulled off an arbitrage, and you win money on *any* result.

If, when this happens, you subsidize people for their time and capital, they'd be much more excited to participate. I think this would have a stronger effect than a similar subsidy to the market itself, once you get enough liquidity to jump start trading. Make sure that if money gets tied up for months or years, that it won't be for nothing.

IV. Limited Hidden Information

If your goal is to buy the hidden information, you might be all right with others not being interested in your market, as long as the subsidy brings the insiders to the table to mop up the free money. That approach is quite expensive. If the regular traders are still driven away, you'll end up paying a lot to get the insiders to show their hand, because they can't make money off anyone else. Even insiders start to worry that others have *more and better* inside information than they do, which could put them at a disadvantage. So it's still important to bring in the outsiders.

One approach is to make the inside information public. Do your own investigations, require disclosure from those participating in the events themselves, work to keep everyone informed. That helps but when what you want to get at is the inside information it only goes so far.

That means that when this is your problem, and you can't fix it directly through action and disclosure, you're going to have to spend a lot of money. The key is *to give that money to the outsiders as much as possible.* They are the ones you need at the table,

to get yourself a good market. The insiders can then prey on the outsiders, but that's much better than preying on you directly.

The counterargument, especially if you don't need to show liquidity or volume, is that if you buy the information directly there's less noise, so perhaps you want to design the system to get a small number of highly informed traders and let everyone else get driven away. In cases where the outsiders would be pure noise, where the insiders outright know the answer, and where getting outsiders to be suckers that take a loss isn't practical, that can be best.

V. Disagreement and Interest

This one's easy. You are paying a subsidy, so you're the sucker. Be loud about it so everyone knows you're the sucker, and then they can fight to cash in. Excellent.

The other half, disagreement, is still important. Many people, whose analysis and participation you want, still benefit from a story that explains why they are being paid to express an opinion, rather than fighting to be slightly more efficient at capturing the subsidy. And of course, if no one disagrees about the answer, then your subsidy was wasted, since you already knew the answer!

In light of those issues, what are the best ways to subsidize the market?

Option 0: Cover Your Basics

Solve the issues noted above. Choose a market people want to participate in to begin with. Ensure there are carefully written rules with no ambiguity, that any problems there are covered. Make sure you'll get things resolved and paid quickly, that capital won't be tied up one minute longer than necessary. When possible, disclose all the relevant information, on all levels. If things don't resolve, it would be great if you could compensate people for their time and capital.

And also, make sure everyone is confident the winners will be paid! Nothing kills a market like worrying you can't collect if you win. That's often as or more important even than providing strong, reliable liquidity.

If you can improve your interface, usability, accessibility, user's tax liability or anything like that, definitely do that. If your market design is poor, such as having the wrong tick size, make sure to fix that. Tick sizes that are too small discourage the providing of liquidity to the market, and are in my experience a bigger and more common mistake than ticks that are too big.

Finally, waive the fees. All of them. Deposit fees, withdraw fees, trading fees, you name it. At most, there should be a fee when taking liquidity that is paid *entirely* to the trader providing liquidity. People hate paying fees a lot more than they like getting subsidies. They won't cancel out.

With that out of the way, what are your options for the main subsidy?

Option 1: Be a Market Maker and Provide Liquidity Directly

As the subsidizer of the market, commit to being the market maker with well-defined rules.

The standard principle is, let everyone know that there will always be $X of liquidity available on both sides, and at a fixed cost of Y% price difference between your bid and your offer. So for example, you might agree to offer $1,000 on each side with a difference of 5% at all times, starting with a 48% bid and a 53% offer. You'd then adjust as you did trades.

A simple rule to protect yourself from unlimited downside is if you do a trade for some percent of your liquidity, you adjust your price that percentage of its width. So in this example, if someone took 40% of your offer, you'd adjust by 40% of 5%, which is 2%, and now have a 50% bid and a 55% offer. If you follow such a rule, your maximum loss is what it takes to move the odds to 0% or 100% (and if you let people keep trading until the event is done, you *will* take that loss). People trading against you in opposite directions can make you money, but can't cost you money.

For convenience, you can post additional bids and offers so that if someone wants to move the odds a lot, they can see what liquidity they would get from you, and have the option to take it all at once. You'll lose money every time the fair probability changes, but that's why they call it a subsidy, an this encourages people to show their information quickly and efficiently.

There are ways to make that smarter, so you can lose less (or make more!) money while offering better liquidity, which will be left as an exercise to the reader. Generally they sacrifice simplicity and transparency in order to make the subsidy 'more efficient.' The danger is that if the subsidy is attempting to ensure a sucker is at the table, it does not do that if it stops being the sucker, or it becomes too hard to tell if it is one or not.

Then again, the dream is to offer a subsidy that doesn't cost you anything, or even makes you money! Market making can be highly profitable when done skillfully, while also building up a marketplace.

Option 2: Take Liquidity

If you provide liquidity, others will take advantage, but in some ways you make it *harder* to *provide* liquidity. If you *take* liquidity, you make it more profitable to provide it, at the risk of making the market *look* less liquid.

It also loses money. The more clear you are about what you are up to, the better.

There are a few fun variants of this, if you're all right with the expense.

One strategy is to take periodically liquidity in both directions. At either fixed or random intervals, examine the order books in the market. If they meet required conditions (e.g. there is at least $X on the bid and offer within Y% of each other) then you hit the bid *and* lift the offer for $Z.

This costs you money, since your trades net out at a loss. If someone else was both the best bid and best offer, they made money.

That's the idea. You're directly subsidizing people to aggressively provide liquidity.

Traders compete to be on the bid and offer to trade with you, the virtual customer, which in turn gives those with an opinion a liquid market to trade against. Sometimes people get far *too* aggressive providing in such situations, and those trying to capture the subsidy end up losing money because they make bad trades against others, especially if they don't then hedge.

You can also do this in a more random or unbalanced fashion. If you flip a coin each day and decide whether to be a buyer or a seller, that will cause the price to temporarily become 'unfair' to satisfy your demand – you'll get a bad price. But that creates a trading opportunity for others. It can also make the results hard to interpret, which is a risk.

Option 3: Subsidize Trading / Give Free Money

Often you'll see crypto exchanges do this as a promotion, offering a prize to whoever trades the most of some coin. By paying for trades, you're encouraging exactly what you want.

Except that you're probably *not* doing that. Remember Goodhart's Law.

The problem is 'wash' trading, where people trade with each other or themselves without taking on positions. This is bad on every level. It misleads everyone about the volume and price, and doesn't help at all with finding out the answer to the question the market is trying to answer. The last thing you want to do is encourage it!

For that reason, subsidizing trading itself is a dangerous game. But it can be done, if you're careful with the design.

Many online sites have tried this in the form of the classic 'deposit bonus' or even the free play. Anyone can sign up and get Free Money in exchange for engaging in a minimum amount of trading activity. And of course, most of the time, a deposit to match, if the offer is more than a small 'free play.' In for-profit markets the goal is to have the required activity make up for the subsidy, then hopefully hook the customer to keep them trading. There are always those looking to game these offerings if you leave them vulnerable.

That can work for you. Getting those same people, who are often quite creative and clever, thinking about how to come out ahead in your system can be a big win if your end goal isn't profit! So long as you make it sufficiently difficult to do wash trading or sign up for tons of copies of the bonuses, you can give them a puzzle worth maximizing (from their perspective) and effectively rent their labor to see what they think of the situation.

Option 4: Subsidize Market Making

You can also subsidize *market making activity*, as an alternative to doing the job yourself and butchering it. That's activity you can't fake, provided you set the rules carefully. Paying people who provide rather than take liquidity is good, and often paying for real two-sided market making activity is better. As always, make sure you're not vulnerable to wash trading or other forms of collusion.

Option 5: Advertising

People can't trade what they aren't thinking about or don't know about.

Putting It All Together

Which of these strategies is most efficient and what circumstances change that answer?

It's expensive to change or clarify your rules and conditions once trading has begun, so invest in doing that first. Other quality of life improvements are great, but take a back seat to establishing good liquidity.

I list Option 0 first because it's things you *definitely* should do if you're taking the operation seriously, but that doesn't mean you always do all of them *first* before the direct subsidy. It's great if you can, but often you need to establish liquidity first.

If 'no liquidity' is the pain point and bad experience, there isn't much that will overcome that. There's no market. So if you don't have liquidity yet, providing at least a reasonable amount, or paying someone else to do it, is the best thing you can do. Just throw something out there and see what happens. This makes intuitive sense all around – as an easy intuition pump, if you want to know if something is more likely than not, offering someone a 50/50 bet on it is a great way to get their real opinion.

Once liquidity isn't a full deal breaker, it's time to go with Option 0, then return to increasing the subsidy and spreading the word.

What form should the direct subsidy take?

I'd advise to continue to *take away bad experiences and barriers* first.

The best subsidy is paying to produce reliable, safe and easy to use software, getting ironclad rules in place, being ready to handle deposits, withdraws, evaluation of results and other hassles. Make sure people can find your markets and set up the markets people want to find.

Next best is to *avoid fees.* People hate fees more than they love subsidies. Yes, you can trick people with deposit bonuses and then charge them a lot on their trades, but the best way to get away with that is bake the fees into the trade prices, so it doesn't look like a fee.

At a minimum, you shouldn't be charging fees for deposits or withdraws, or for providing liquidity in the market.

Next up, make trades cost net zero fees. Either charge nothing to provide or to take liquidity, *or* charge a fee to take liquidity but pay it to those who provide.

After that, my opinions are less confident, but here's my best guess.

If that's still not good enough, provide liquidity. Either pay someone else to be a market maker, or provide the service yourself. I like the idea of a 'dumb' market maker everyone knows is dumb, and that operates with known rules that hamstring it. If you're looking to provide a subsidy, this is a great way to do that. A smarter market

maker is cheaper, and can provide better liquidity, but is less obviously a target. As the market matures, you'll want to transition to something smarter. Thin markets want obviously dumb providers.

Once you've done a healthy amount of that, then you'll want to give away Free Money. Give people some cash in exchange for participating in the market at all, or trading a minimum amount. Or give people bonuses on deposited funds so long as they use them to trade, or similar.

You have to watch for abuse. If you can respond to abuse by changing the system, it's fine to be vulnerable to abuse in theory, and even allow small amounts of it. If you're going to release a cryptographic protocol you can't alter, you'll need to be game theoretically robust, so this won't be an option, and you'll have to retreat to taking liquidity.

Taking liquidity seems less likely to motivate the average potential participant, and costs you weirdness points, but does provide a strong incentive for the right type of trader. The best reason I can think of to use such a strategy is that it is robust to abuse. That's a big game if you can't respond dynamically to unfriendly players.

At the end of the day, your biggest barriers are that people's attention is limited, complexity is bad, opportunity cost is high and people don't do things. I keep meaning to get around to bothering with HyperMind and/or PredictIt, and keep not doing it, and I'm guessing I am far from alone in that. Subsidy can get people excited and make markets work that wouldn't otherwise get off the ground. What I think they can't do at reasonable cost is fix fundamental problems. If you don't have a great product behind the subsidy, it's going to be orders of magnitude more expensive to motivate participation.

Alignment Newsletter #18

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

Highlights

Learning Dexterity *(Many people at OpenAI)*: Most current experiments with robotics work on relatively small state spaces (think 7 degrees of freedom, each a real number) and are trained in simulation. If we could throw a lot of compute at the problem, could we do significantly better? Yes! Using the same general approach as with OpenAI Five, OpenAI has built a system called Dactyl, which allows a physical real-world dexterous hand to manipulate a block. It may not seem as impressive as the videos of humanoids running through obstacle courses, but this is way harder than your typical Mujoco environment, especially since they aim to get it working on a real robot. As with OpenAI Five, they only need a reward function (I believe not even a shaped reward function in this case), a simulator, and a good way to explore. In this setting though, "exploration" is actually domain randomization, where you randomly set parameters that you are uncertain about (such as the coefficient of friction between two surfaces), so that the learned policy is robust to distribution shift from the simulator to the real world. (OpenAI Five also used domain randomization, but in that case it was not because we were uncertain about the parameters in the simulator, but because the policy was too specialized to the kinds of characters and heroes it was seeing, and randomizing those properties exposed it to a wider variety of scenarios so it had to learn more general policies.) They use 6144 CPU cores and 8 GPUs, which is *much* less than for OpenAI Five, but *much* more than for a typical Mujoco environment.

They do separate the problem into two pieces -- first, they learn how to map from camera pictures to a 3D pose (using convolutional nets), and second, they use RL to choose actions based on the 3D pose. They can also get better estimates of the 3D pose using motion tracking. They find that the CNN is almost as good as motion tracking, and that the domain randomization is crucial for getting the system to actually work.

They also have a couple of sections on surprising results and things that didn't work. Probably the most interesting part was that they didn't need to use the tactile sensors to get these results. They couldn't get these sensors in simulation, so they just did without and it seems to have worked fine. It also turns out that the robot's reaction time wasn't too important -- there wasn't a big difference in changing from 80ms reaction time to 40ms reaction time; in fact, this just increased the required training time without much benefit.

Probably the most interesting part of the post is the last paragraph (italics indicates my notes): "This project completes a full cycle of AI development that OpenAI has been pursuing for the past two years: we've developed a new learning algorithm *(PPO)*, scaled it massively to solve hard simulated tasks *(OpenAI Five)*, and then applied the resulting system to the real world *(this post)*. Repeating this cycle at increasing scale is the primary route we are pursuing to increase the capabilities of today's AI systems towards safe artificial general intelligence."

My opinion: This is pretty exciting -- transferring a policy from simulation to the real world is notoriously hard, but it turns out that as long as you use domain randomization (and 30x the compute) it actually is possible to transfer the policy. I wish they had compared the success probability in simulation to the success probability in the real world -- right now I don't know how well the policy transferred. (That is, I want to evaluate how well domain randomization solved the distribution shift problem.) Lots of other exciting things too, but they are pretty similar to the exciting things about OpenAI Five, such as the ability to learn higher level strategies like finger pivoting and sliding (analogously, fighting over mid or 5-man push).

Variational Option Discovery Algorithms *(Joshua Achiam et al)*: We can hope to do hierarchical reinforcement learning by first discovering several useful simple policies (or "options") by just acting in the environment without any reward function, and then using these options as primitive actions in a higher level policy that learns to do some task (using a reward function). How could we learn the options without a reward function though? Intuitively, we would like to learn behaviors that are different from each other. One way to frame this would be to think of this as an encoder-decoder problem. Suppose we want to learn K options. Then, we can give the encoder a number in the range [1, K], have it "encode" the number into a trajectory τ (that is, our encoder is a policy), and then have a decoder take τ and recover the original number. We train the encoder/policy and decoder jointly, optimizing them to successfully recover the original number (called a *context*). Intuitively, the encoder/policy wants to have very different behaviors for each option, so that it easy for decoder to figure out the context from the trajectory τ. However, a simple solution would be for the encoder/policy to just take a particular series of actions for each context and then stop, and the decoder learns an exact mapping from final states to contexts. To avoid this, we can decrease the capacity of the decoder (i.e. don't give it too many layers), and we also optimize for the *entropy* of the encoder/policy, which encourages the encoder/policy to be more stochastic, and so it is more likely to learn overall behaviors that can still have some stochasticity, while still allowing the decoder to decode them. It turns out that this optimization problem has a one-to-one correspondence with variational autoencoders, motivating the name "variational option discovery". To stabilize training, they start with a small K, and increase K whenever the decoder becomes powerful enough. They evaluate in Gym environments, a simulated robotic hand, and a new "Toddler" environment. They find that the scheme works well (in terms of maximizing the objective) in all environments, but that the learned behaviors no longer look natural in the Toddler environment (which is the most complex). They also show that the learned policies can be used for hierarchical RL in the AntMaze problem.

This is very similar to the recent Diversity Is All You Need. DIAYN aims to decode the context from *every state* along a trajectory, which incentivizes it to find behaviors of the form "go to a goal state", whereas VALOR (this work) decodes the context from the entire trajectory (without actions, which would make the decoder's job too easy), which allows it to learn behaviors with motion, such as "go around in a circle".

My opinion: It's really refreshing to read a paper with a negative result about their own method (specifically, that the learned behaviors on Toddler do not look natural). It makes me trust the rest of their paper so much more. (A very gameable instinct, I know.) While they were able to find a fairly diverse set of options, and could interpolate between them, their experiments found that using this for hierarchical RL was about as good as training hierarchical RL from scratch. I guess I'm just saying things they've already said -- I think they've done such a great job writing this paper

that they've already told me what my opinion about the topic should be, so there's not much left for me to say.

Technical AI alignment

Problems

A Gym Gridworld Environment for the Treacherous Turn *(Michaël Trazzi)*: An example Gym environment in which the agent starts out "weak" (having an inaccurate bow) and later becomes "strong" (getting a bow with perfect accuracy), after which the agent undertakes a treacherous turn in order to kill the supervisor and wirehead.

My opinion: I'm a fan of executable code that demonstrates the problems that we are worrying about -- it makes the concept (in this case, a treacherous turn) more concrete. In order to make it more realistic, I would want the agent to grow in capability organically (rather than simply getting a more powerful weapon). It would really drive home the point if the agent undertook a treacherous turn the very first time, whereas in this post I assume it learned using many episodes of trial-and-error that a treacherous turn leads to higher reward. This seems hard to demonstrate with today's ML in any complex environment, where you need to learn from experience instead of using eg. value iteration, but it's not out of the question in a continual learning setup where the agent can learn a model of the world.

Agent foundations

Counterfactuals, thick and thin *(Nisan)*: There are many different ways to formalize counterfactuals (the post suggests three such ways). Often, for any given way of formalizing counterfactuals, there are many ways you could take a counterfactual, which give different answers. When considering the physical world, we have strong causal models that can tell us which one is the "correct" counterfactual. However, there is no such method for logical counterfactuals yet.

My opinion: I don't think I understood this post, so I'll abstain on an opinion.

Decisions are not about changing the world, they are about learning what world you live in *(shminux)*: The post tries to reconcile decision theory (in which agents can "choose" actions) with the deterministic physical world (in which nothing can be "chosen"), using many examples from decision theory.

Handling groups of agents

Multi-Agent Generative Adversarial Imitation Learning *(Jiaming Song et al)*: This paper generalizes GAIL (which was covered last week) to the multiagent setting, where we want to imitate a group of interacting agents. They want to find a Nash equilibrium in particular. They formalize the Nash equilibrium constraints and use this to motivate a particular optimization problem for multiagent IRL, that looks very similar to their optimization problem for regular IRL in GAIL. After that, it is quite similar to GAIL -- they use a regularizer ψ for the reward functions, show that the composition of multiagent RL and multiagent IRL can be solved as a single optimization problem involving the convex conjugate of ψ, and propose a particular instantiation of ψ that is data-dependent, giving an algorithm. They do have to assume in the theory that the

multiagent RL problem has a unique solution, which is not typically true, but may not be too important. As before, to make the algorithm practical, they structure it like a GAN, with discriminators acting like reward functions. What if we have prior information that the game is cooperative or competitive? In this case, they propose changing the regularizer ψ, making it keep all the reward functions the same (if cooperative), making them negations of each other (in two-player zero-sum games), or leaving it as is. They evaluate in a variety of simple multiagent games, as well as a plank environment in which the environment changes between training and test time, thus requiring the agent to learn a robust policy, and find that the correct variant of MAGAIL (cooperative/competitive/neither) outperforms both behavioral cloning and single-agent GAIL (which they run N times to infer a separate reward for each agent).

My opinion: Multiagent settings seem very important (since there does happen to be more than one human in the world). This looks like a useful generalization from the single agent case to the multiagent case, though it's not clear to me that this deals with the major challenges that come from multiagent scenarios. One major challenge is that there is no longer a single optimal equilibrium when there are multiple agents, but they simply assume in their theoretical analysis that there is only one solution. Another one is that it seems more important that the policies take history into account somehow, but they don't do this. (If you don't take history into account, then you can't learn strategies like tit-for-tat in the iterated prisoner's dilemma.) But to be clear I think this is the standard setup for multiagent RL -- it seems like field is not trying to deal with this issue yet (even though they could using eg. a recurrent policy, I think?)

Miscellaneous (Alignment)

Safely and usefully spectating on AIs optimizing over toy worlds *(Alex Mennen)*: One way to achieve safety would be to build an AI that optimizes in a virtual world running on a computer, and doesn't care about the physical world. Even if it realizes that it can break out and eg. get more compute, these sorts of changes to the physical world would not be helpful for the purpose of optimizing the abstract computational object that is the virtual world. However, if we take the results of the AI and build them in the real world, that causes a distributional shift from the toy world to the real world that could be catastrophic. For example, if the AI created another agent in the toy world that did reasonable things in the toy world, when we bring it to the real world it may realize that it can instead manipulate humans in order to do things.

My opinion: It's not obvious to me, even on the "optimizing an abstract computational process" model, why an AI would not want get more compute -- it can use this compute for itself, without changing the abstract computational process it is optimizing, and it will probably do better. It seems that if you want to get this to work, you need to have the AI want to compute the result of running *itself* without any modification or extra compute on the virtual world. This feels very hard to me. Separately, I also find it hard to imagine us building a virtual world that is similar enough to the real world that we are able to transfer solutions between the two, even with some finetuning in the real world.

Sandboxing by Physical Simulation? *(moridinamael)*

Near-term concerns

Adversarial examples

Evaluating and Understanding the Robustness of Adversarial Logit Pairing *(Logan Engstrom, Andrew Ilyas and Anish Athalye)*

AI strategy and policy

The Facets of Artificial Intelligence: A Framework to Track the Evolution of AI *(Fernando Martinez-Plumed et al)*

Podcast: Six Experts Explain the Killer Robots Debate *(Paul Scharre, Toby Walsh, Richard Moyes, Mary Wareham, Bonnie Docherty, Peter Asaro, and Ariel Conn)*

AI capabilities

Reinforcement learning

Learning Dexterity *(Many people at OpenAI)*: Summarized in the highlights!

Variational Option Discovery Algorithms *(Joshua Achiam et al)*: Summarized in the highlights!

Learning Plannable Representations with Causal InfoGAN *(Thanard Kurutach, Aviv Tamar et al)*: Hierarchical reinforcement learning aims to learn a hierarchy of actions that an agent can take, each implemented in terms of actions lower in the hierarchy, in order to get more efficient planning. Another way we can achieve this is to use a classical planning algorithm to find a sequence of *waypoints*, or states that the agent should reach that will allow it to reach its goal. These waypoints can be thought of as a high-level plan. You can then use standard RL algorithms to figure out how to go from one waypoint to the next. However, typical planning algorithms that can produce a sequence of waypoints require very structured state representations, that were designed by humans in the past. How can we learn them directly from data? This paper proposes Causal InfoGAN. They use a GAN where the generator creates adjacent waypoints in the sequence, while the discriminator tries to distinguish between waypoints from the generator and pairs of points sampled from the true environment. This incentivizes the generator to generate waypoints that are close to each other, so that we can use an RL algorithm to learn to go from one waypoint to the next. However, this only lets us generate adjacent waypoints. In order to use this to make a sequence of waypoints that gets from a start state to a goal state, we need to use some classical planning algorithm. In order to do that, we need to have a structured state representation. GANs do not do this by default. InfoGAN tries to make the latent representation in a GAN more meaningful by providing the generator with a "code" (a state in our case) and maximizing the mutual information of the code and the output of the generator. In this setting, we want to learn representations that are good for planning, so we want to encode information about *transitions* between states. This leads to the Causal InfoGAN objective, where we provide the generator with a pair of abstract states (s, s'), have it generate a pair of observations (o, o') and maximize the mutual information between (s, s') and (o, o'), so that s and s' become good low-dimensional representations of o and o'. They show that Causal InfoGAN can create sequences of waypoints in a rope manipulation task, that previously had to be done manually.

My opinion: We're seeing more and more work combining classical symbolic approaches with the current wave of statistical machine learning from big data, that gives them the best of both worlds. While the results we see are not general intelligence, it's becoming less and less true that you can point to a broad swath of capabilities that AI cannot do yet. I wouldn't be surprised if a combination of symbolic and stastical AI techniques led to large capability gains in the next few years.

Deep learning

TensorFuzz: Debugging Neural Networks with Coverage-Guided Fuzzing *(Augustus Odena et al)*

News

AI Strategy Project Manager *(FHI)*

Preliminary thoughts on moral weight

This post adapts some internal notes I wrote for the Open Philanthropy Project, but they are merely at a "brainstorming" stage, and do not express my "endorsed" views nor the views of the Open Philanthropy Project. This post is also written quickly and not polished or well-explained.

My 2017 Report on Consciousness and Moral Patienthood tried to address the question of "Which creatures are moral patients?" but it did little to address the question of "moral weight," i.e. how to weigh the interests of different kinds of moral patients against each other:

> For example: suppose we conclude that fishes, pigs, and humans are all moral patients, and we estimate that, for a fixed amount of money, we can (in expectation) dramatically improve the welfare of (a) 10,000 rainbow trout, (b) 1,000 pigs, or (c) 100 adult humans. In that situation, how should we compare the different options? This depends (among other things) on how much “moral weight” we give to the well-being of different kinds of moral patients.

Thus far, philosophers have said very little about moral weight (see below). In this post I lay out one approach to thinking about the question, in the hope that others might build on it or show it to be misguided.

Proposed setup

For the simplicity of a first-pass analysis of moral weight, let's assume a variation on classical utilitarianism according to which the only thing that morally matters is the moment-by-moment character of a being's conscious experience. So e.g. it doesn't matter whether a being's rights are respected/violated or its preferences are realized/thwarted, except insofar as those factors affect the moment-by-moment character of the being's conscious experience, by causing pain/pleasure, happiness/sadness, etc.

Next, and again for simplicity's sake, let's talk only about the "typical" conscious experience of "typical" members of different species when undergoing various "canonical" positive and negative experiences, e.g. consuming species-appropriate food or having a nociceptor-dense section of skin damaged.

Given those assumptions, when we talk about the relative "moral weight" of different species, we mean to ask something like "How morally important is 10 seconds of a typical human's experience of [some injury], compared to 10 seconds of a typical rainbow trout's experience of [that same injury]?

For this exercise, I'll separate "moral weight" from "probability of moral patienthood." Naively, you could then multiply your best estimate of a species' moral weight (using humans as the baseline of 1) by P(moral patienthood) to get the species' "expected moral weight" (or whatever you want to call it). Then, to estimate an intervention's potential benefit for a given species, you could multiply [expected moral weight of species] × [individuals of species affected] × [average # of minutes of conscious experience affected across those individuals] × [average magnitude of positive impact on those minutes of conscious experience].

However, I say "naively" because *this doesn't actually work*, due to two-envelope effects.

Potential dimensions of moral weight

What features of a creature's conscious experience might be relevant to the moral weight of its experiences? Below, I describe some possibilities that I previously mentioned in Appendix Z7 of my moral patienthood report.

Note that any of the features below could be (and in some cases, very likely are) hugely multidimensional. For simplicity, I'm going to assume a *unidimensional* characterization of them, e.g. what we'd get if we looked only at the principal component in a principal component analysis of a hugely multidimensional phenomenon.

Clock speed of consciousness

Perhaps animals vary in their "clock speed." E.g. a hummingbird reacts to some things much faster than I ever could. If *any* of that is under conscious control, its "clock speed" of conscious experience seems like it should be faster than mine, meaning that, intuitively, it should have a greater number of subjective "moments of consciousness" per objective minute than I do.

In general, smaller animals probably have faster clock speeds than larger ones, for mechanical reasons:

> The natural oscillation periods of most consciously controllable human body parts are greater than a tenth of a second. Because of this, the human brain has been designed with a matching reaction time of roughly a tenth of a second. As it costs more to have faster reaction times, there is little point in paying to react much faster than body parts can change position.
>
> …the first resonant period of a bending cantilever, that is, a stick fixed at one end, is proportional to its length, at least if the stick's thickness scales with its length. For example, sticks twice as long take twice as much time to complete each oscillation. Body size and reaction time are predictably related for animals today… (Hanson 2016, ch. 6)

My impression is that it's a common intuition to value experience by its "subjective" duration rather than its "objective" duration, with no discount. So if a hummingbird's clock speed is 3x as fast as mine, then all else equal, an objective minute of its conscious pleasure would be worth 3x an objective minute of my conscious pleasure.

Unities of consciousness

Philosophers and cognitive scientists debate how "unified" consciousness is, in various ways. Our normal conscious experience *seems* to many people to be pretty "unified" in various ways, though sometimes it feels less unified, for example when one goes "in and out of consciousness" during a restless night's sleep, or when one engages in certain kinds of meditative practices.

Daniel Dennett suggests that animal conscious experience is radically less unified than human consciousness is, and cites this as a major reason he doesn't give most animals much moral weight.

For convenience, I'll use Bayne (2010)'s taxonomy of types of unity. He talks about subject unity, representational unity, and phenomenal unity — each of which has a "synchronic" (momentary) and "diachronic" (across time) aspect of unity.

Subject unity

Bayne explains:

> My conscious states possess a certain kind of unity insofar as they are all mine; likewise, your conscious states possess that same kind of unity insofar as they are all yours. We can describe conscious states that are had by or belong to the same subject of experience as subject unified. Within subject unity we need to distinguish the unity provided by the subject of experience across time (diachronic unity) from that provided by the subject at a time (synchronic unity).

Representational unity

Bayne explains:

> Let us say that conscious states are representationally unified to the degree that their contents are integrated with each other. Representational unity comes in a variety of forms. A particularly important form of representational unity concerns the integration of the contents of consciousness around perceptual objects—what we might call 'object unity'. Perceptual features are not normally represented by isolated states of consciousness but are bound together in the form of integrated perceptual objects. This process is known as feature-binding. Feature-binding occurs not only within modalities but also between them, for we enjoy multimodal representations of perceptual objects.

I suspect many people wouldn't treat representational unity as all that relevant to moral weight. E.g. there are humans with low representational unity of a sort (e.g. visual agnosics); are their sensory experiences less morally relevant as a result?

Phenomenal unity

Bayne explains:

> Subject unity and representational unity capture important aspects of the unity of consciousness, but they don't get to the heart of the matter. Consider again what it's like to hear a rumba playing on the stereo whilst seeing a bartender mix a mojito. These two experiences might be subject unified insofar as they are both yours. They might also be representationally unified, for one might hear the rumba as coming from behind the bartender. But over and above these unities is a deeper and more primitive unity: the fact that these two experiences possess a conjoint experiential character. There is something it is like to hear the rumba, there is something it is like to see the bartender work, and there is something it is like to hear the rumba while seeing the bartender work. Any description of one's overall state of consciousness that omitted the fact that these experiences are

> had together as components, parts, or elements of a single conscious state would be incomplete. Let us call this kind of unity — sometimes dubbed 'co-consciousness' — phenomenal unity.
>
> Phenomenal unity is often in the background in discussions of the 'stream' or 'field' of consciousness. The stream metaphor is perhaps most naturally associated with the flow of consciousness — its unity through time — whereas the field metaphor more accurately captures the structure of consciousness at a time. We can say that what it is for a pair of experiences to occur within a single phenomenal field just is for them to enjoy a conjoint phenomenality — for there to be something it is like for the subject in question not only to have both experiences but to have them together. By contrast, simultaneous experiences that occur within distinct phenomenal fields do not share a conjoint phenomenal character.

Unity-independent intensity of valenced aspects of consciousness

A common report of those who take psychedelics is that, while "tripping," their conscious experiences are "more intense" than they normally are. Similarly, different pains feel similar but have different intensities, e.g. when my stomach is upset, the intensity of my stomach pain waxes and wanes a fair bit, until it gradually fades to not being noticeable anymore. Same goes for conscious pleasures.

It's possible such variations in intensity are entirely accounted for by their degrees of different kinds of unity, or by some other plausible feature(s) of moral weight, but maybe not. If there *is* some additional "intensity" variable for valenced aspects of conscious experience, it would seem a good candidate for affecting moral weight.

From my own experience, my guess is that I would endure ~10 seconds of the most intense pain I've ever experienced to avoid experiencing ~2 months of the lowest level of discomfort that I'd bother to call "discomfort." That very low level of discomfort might suggest a lower bound on "intensity of valenced aspects of experience" that I intuitively morally care about, but "the most intense pain I've ever experienced" probably is not the *highest* intensity of valenced aspects of experience it is possible to experience — probably not even close. You could consider similar trades to get a sense for how much you intuitively value "intensity of experience," at least in your own case.

Moral weights of various species

(*This section edited slightly on 2020-02-26.*)

If we thought about all this more carefully and collected as much relevant empirical data as possible, what moral weights might we assign to different species?

Whereas my probabilities of moral patienthood for any animal as complex as a crab only range from 0.2 - 1, the plausible ranges of moral weight seem like they could be much larger. I don't feel like I'd be surprised if an omniscient being told me that my extrapolated values would assign pigs *more* moral weight than humans, and I don't

feel like I'd be surprised if an omniscient being told me my extrapolated values would assign pigs .0001 moral weight (assuming they were moral patients at all).

To illustrate how this might work, below are some guesses at some "plausible ranges of moral weight" (80% prediction interval) for a variety of species that someone might come to, if they had intuitions like those explained below.

- Humans: 1 (baseline)
- Chimpanzees: 0.001 - 2
- Pigs: 0.0005 - 3.5
- Cows: 0.0001 - 5
- Chickens: 0.00005 - 10
- Rainbow trout: 0.00001 - 13
- Fruit fly: 0.000001 - 20

(But whenever you're tempted to multiply such numbers by something, remember two-envelope effects!)

What intuitions might lead to something like these ranges?

- An intuition to not place much value on "complex/higher-order" dimensions of moral weight — such as "fullness of self-awareness" or "capacity for reflecting on one's holistic life satisfaction" — above and beyond the subjective duration and "intensity" of relatively "brute" pleasure/pain/happiness/sadness that (in humans) tends to accompany reflection, self-awareness, etc.
- An intuition to care more about subject unity and phenomenal unity than about such higher-order dimensions of moral weight.
- An intuition to care most of all about clock speed and experience intensity (if intensity is distinct from unity).
- Intuitions that if the animal species listed above are conscious, they:
    - have very little of the higher-order dimensions of conscious experience,
    - have faster clock speeds than humans (the smaller the faster),
    - probably have lower "intensity" of experience, but *might* actually have somewhat *greater* intensity of experience (e.g. because they aren't distracted by linguistic thought),
    - have moderately less subject unity and phenomenal unity, especially of the diachronic sort.

Under these intuitions, the low end of the ranges above could be explained by the possibility that intensity of conscious experience diminishes dramatically with brain complexity and flexibility, while the high end of the ranges above could be explained by the possibility concerning faster clock speeds for smaller animals, the possibility of lesser unity in non-human animals (which one might value at >1x for the same reason one might value a dually-conscious split-brain patient at ~2x), and the possibility for *greater* intensity of experience in simpler animals.

Other writings on moral weight

- Brian Tomasik: Is animal suffering less bad than human suffering?; Which computations do I care about?; Is brain size morally relevant?; Do Smaller

Animals Have Faster Subjective Experiences?; Two-Envelopes Problem for Uncertainty about Brain-Size Valuation and Other Moral Questions

- Nick Bostrom: Quantity of Experience
- Kevin Wong: Counting Animals
- Oscar Horta: Questions of Priority and Interspecies Comparisons of Happiness
- Adler et al., Would you choose to be happy? Tradeoffs between happiness and the other dimensions of life in a large population survey

Is there a practitioner's guide for rationality?

Hi everyone, I'm new to the community, and am currently working my way through the sequences — yes, all of them.

In the introduction to the first book of Rationality A-Z, Eliezer says:

> I didn’t realize that the big problem in learning this valuable way of thinking was figuring out how to practice it, not knowing the theory. I didn’t realize that part was the priority; and regarding this I can only say “Oops” and “Duh.” Yes, sometimes those big issues really are big and really are important; but that doesn’t change the basic truth that to master skills you need to practice them and it’s harder to practice on things that are further away. (Today the Center for Applied Rationality is working on repairing this huge mistake of mine in a more systematic fashion.)

Just wanted to ask if CFAR has got any of those reorganised materials up, and if they're linked to from anywhere on this site? Any links to *other* rationality-as-practice blog posts or books or sequences would also be incredibly appreciated!

Zetetic explanation

This is a linkpost for http://benjaminrosshoffman.com/zetetic-explanation/

There is a kind of explanation that I think ought to be a cornerstone of good pedagogy, and I don't have a good word for it. My first impulse is to call it a historical explanation, after the original, investigative sense of the term "history." But in the interests of avoiding nomenclature collision, I'm inclined to call it "zetetic explanation," after the Greek word for seeking, an explanation that embeds in itself an inquiry into the thing.

Often in "explaining" a thing, we simply tell people what words they ought to say about it, or how they ought to interface with it right now, or give them technical language for it without any connection to the ordinary means by which they navigate their lives. We can call these sorts of explanations nominal, functional, and formal.

In my high school chemistry courses, for instance, there was lots of "add X to Y and get Z" plus some formulas, and I learned how to manipulate the symbols in the formulas, but this bore no relation whatsoever to the sorts of skills used in time-travel or Robinson Crusoe stories. Overall I got the sense that chemicals were a sort of magical thing produced by a mysterious Scientific-Industrial priesthood in special temples called laboratories or factories, not things one might find outdoors.

It's only in the last year that I properly learned how one might get something as simple as copper or iron, reading David W. Anthony's The Horse, the Wheel, and Language and Vaclav Smil's *Still the Iron Age*, both of which contain clear and concrete summaries of the process. Richard Feynman's explanation of triboluminescence is a short example of a zetetic explanation in chemistry, and Paul Lockhart's A Mathematician's Lament bears strong similarities in the field of pure mathematics.

I'm going to work through a different example here, and then discuss this class of explanation more generally.

What is yeast? A worked example

Recently my mother noted that when, in science class, her teacher had explained how bread was made, it had been a revelation to her. I pointed out that while this explanation removed bread from the category of a pure *product*, to be purchased and consumed, it still placed it in the category of an industrial product requiring specialized, standardized inputs such as yeast. My mother observed that she didn't *really* know what yeast was, and I found myself explaining.

Seeds, energy storage, and coevolution

Many plants store energy in chemicals such as proteins and carbohydrates around their seeds, to help them start growing once they're in wet ground. Some animals seek out the seeds with the most extra energy, and poop the occasional seed elsewhere. Sometimes this helps the plant reproduce more than it otherwise would

have; in such cases, the plant may coevolve with the animals that eat it, often investing much larger amounts of energy in or around the seed, since the most calorific seeds get eaten most eagerly.

Humans coevolved with a sort of grass. If you've seen wild grass, you may have observed stalks with seed pods on them, that look sort of like tiny heads of wheat. Grain is basically massively a grass that coevolved with us to produce plump, overnourished seeds.

Energy extraction

Of course, there's only so much we can do to select for digestibility. Often even plants that store a lot of surplus energy need further treatment before they're easy to digest. Some species evolved to specialize in digesting a certain sort of plant matter efficiently; for instance, ruminants such as cattle and sheep have multiple stomachs to break down the free energy in plant matter. Humans,
with unspecialized omnivorous guts, learned other ways to extract energy from plants.

One such way is cooking. If you heat up the starches inside a kernel of wheat, they'll often transform into something easier to digest. But bread made this way can still be difficult to digest, as many eaters of matzah or hardtack have learned. Soaking or sprouting seeds also helps. And a third way to make grains more digestible is fermentation.

Cultured food

Where there's dense storage of energy, there's often leakage. Sometimes a seed gets split open for some reason, and there's a bit of digestible carbohydrate exposed on the surface. Where there's free energy like this, microbes evolve to eat it.

Some of these microbes, especially fungal ones, produce byproducts that are toxic to us. But others, such as some bacteria and yeasts, break down hard-to-digest parts of wheat into substances that are easier for us to digest. Presumably at some point, people noticed that if they wet some flour and left it out for a day or two before cooking it, the resulting porridge or cracker was both tastier and more digestible. (Other fermented products such as sauerkraut may have been discovered in a similar way.)

Of course, while grain-eating microbes will often tend to be found on grain, allowing for such accidental discoveries, there is no guarantee that they'll be the kind we like. Since they mostly just eat accidental discharges of energy, there also just aren't very many of them, compared to the amount of energy available to them once the flour is ground up and mixed with water. It takes a while for them to eat and reproduce enough to process the whole batch.

Eventually, people realized that if they took part of a good batch of dough or porridge and didn't cook it, but instead added it to the next batch, this would yield an edible product both more reliably (because the microbes in the starter would have a head start relative to any potentially harmful microbes) and more quickly (again, because they'd be starting with more microbes relative to the amount of grain they needed to process). This is what we call a sourdough "culture" or "starter".

(You can make a sourdough starter at home by mixing some flour, preferably wholemeal, with water, covering it, and adding some more flour and water each day until it gets bubbly. Supposedly, a regularly fed starter can stay active for generations.)

Breads are particularly convenient foods for a few reasons. First, grains have a very high maximum caloric yield per acre, allowing for high population density. Second, dry grains or flour can be stored for a long time without going bad; as a result, stockpiles can tide people over in lean seasons or years, and be traded over large distances. Third, a loaf of bread itself has some amount of more local portability and durability, relative to a porridge.

Yeast-specific products

One of the microbes found in a sourdough culture, yeast, has a particularly simple metabolism with two main byproducts. It pisses alcohol, and farts carbon dioxide. Carbon dioxide is a gas that can leaven or puff up dough, which makes it nicer to eat. Alcohol is a psychoactive drug, and some people likes how it makes them feel. Many food cultures ended up paying special attention to grain products that used one or the other of these traits: beer and leavened bread.

In the 19th century CE, people figured out how to isolate the yeast from the rest of the sourdough culture, which allowed for industrial, standardized production of beer and bread. If you know exactly how much yeast you're adding to the dough, you can standardize dough rising times and temperatures, allowing for mass production on a schedule, reducing potentially costly surprises.

The price of this innovation is twofold. First, when using standardized yeast to bake bread, we forgo the digestive and taste benefits of the other microbes you would find in a sourdough starter. Second, we become alienated from a crucial part of the production of bread, to the point where many people only relate to it as a recipe composed of products you can buy at a store, rather than something made of components you might find out in the wild or grow self-sufficiently.

Additional thoughts on explanation

I'm having some difficulty articulating exactly what seems distinct about this sort of explanation, but here's a preliminary attempt.

Zetetic explanations will tend to be interdisciplinary, as they will often cover a mixture of social and natural factors leading up to the isolation of the thing being explained. This naturally makes it harder to be an expert in everything one is talking about, and requires some minimal amount of courage on the part of the explainer, who may have to risk being wrong. But they're not merely interdisciplinary. You could separately talk about the use of yeast as a literary motif, the chemistry of the yeast cell, and the industrial use in bread, and still come nowhere close to giving people any real sense of why yeast came into the world or how we found it.

Zetetic explanations are empowering. First, the integration of concrete and model-based thinking is checkable on multiple levels - you can look up confirming or disconfirming facts, and you can also validate it against your personal experience or sense of plausibility, and validate the coherence and simplicity of the models used.

Second, they affirm the basic competence of humans to explore our world. By centering the process of discovery rather than a finished product, such explanations invite the audience to participate in this process, and perhaps to surprise us with new discoveries.

Of course, it can be hard to know where to stop in such explanations, and it can also be hard to know where to start. This post could easily have been twice as long. Ideally, an explainer would attend to the reactions of their audience, and try to touch base with points of shared understanding. Such explanations also require patience on both sides. Another difficulty this approach raises is that plain-language explanations rooted in everyday concepts may not match the way things are referred to in technical or scientific literature, although this problem should not be hard to solve.

In some cases, one might want to forwards-chain from an interesting puzzle or other thing to play with, rather than backwards-chaining from a product. Lockhart seems to favor exploration over explanation for mathematics, and of course there's no particular reason why one can't use both. In particular, the explanation paradigm seems useful for deciding which explorations to propose.

Related: The Steampunk Aesthetic, Truly Part Of You

Entropic Regret I: Deterministic MDPs

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

This is the first in a series of essays that aim to derive a regret bound for DRL that depends on finer attributes of the prior than just the number of hypotheses. Specifically, we consider the entropy of the prior and a certain learning-theoretic dimension parameter. As a "by product", we derive a new regret bound for ordinary RL without resets and without traps. In this chapter, we open the series by demonstrating the latter, under the significant simplifying assumption that the MDPs are deterministic.

Background

The regret bound we previously derived for DRL grows as a power law with the number of hypotheses. In contrast, the RL literature is usually concerned with considering all transition kernels on a fixed state space satisfying some simple assumption (e.g. a bound on the diameter or bias span). In particular, the number of hypotheses is uncountable. While the former seems too restrictive (although it's relatively simple to generalize it to *countable* hypothesis classes), the latter seems too coarse. Indeed, we expect a universally intelligent agent to detect *patterns* in the data, i.e. follow some kind of simplicity prior rather than a uniform distribution over transition kernels.

The underlying technique of our proof was lower bounding the information gain in a single episode of posterior sampling by the expected regret incurred during this episode. Although we have not previously stated it in this form, the resulting regret bound depends on the *entropy* of the prior (we considered a uniform prior instead). This idea (unbeknownst to us) appeared earlier in Russo and Van Roy. Moreover, later Russo and Van Roy used to it formulate a generalization of posterior sampling they call "information-directed sampling" the can produce far better regret bounds in certain scenarios. However, to the best of our knowledge, this technique was not previously used to analyze reinforcement learning (as opposed to bandits). Therefore, it seems natural to derive such an "entropic" regret bound for ordinary RL, before extending it to DRL.

Now, Osband and Van Roy derived a regret bound for priors supported on some space of transition kernels as a function of its Kolmogorov dimension and "eluder" dimension (the latter introduced previously by Russo and Van Roy). They also consider a continuous state space. This is a finer approach than considering nearly arbitrary transition kernels on a fixed state space, but it still doesn't distinguish between different priors with the same support. Our new results involve a parameter similar to eluder dimension, but instead of Kolmogorov dimension we use entropy (in the following chapters we will see that Kolmogorov dimension is, in some sense, an *upper bound* on entropy). As opposed to Osband and Van Roy, we currently limit ourselves to finite state spaces, but on the other hand we consider no resets (at the price of a "no traps" assumption).

In this chapter we derive the entropic regret bound for *deterministic* MDPs. (In the following we will call them *deterministic decision processes* (DDPs) since they have little to do with Markov.) This latter restriction significantly simplifies the analysis. In the following chapters, we will extend it to stochastic MDPs, however the resulting regret bound will be somewhat weaker.

Results

We start by introducing a new learning-theoretic concept of dimension. It is similar to eluder dimension, but is adapted to the discrete deterministic setting and also somewhat stronger (i.e. smaller: more environment classes are low dimensional w.r.t. this concept than w.r.t. eluder dimension).

Definition 1

Consider sets A, B *and* $F \subseteq \{A \to B\}$ *non-empty. Given* $C \subseteq A \times B$ *and* $a^* \in A$*, we say that* a^* *is* F-dependent of C *when for any* $f, g : A \to B$ *s.t. for any* $(a, b) \in C$ *it holds* $f(a) = g(a) = b$*, we have* $f(a^*) = g(a^*)$*. Otherwise, we say that* a^* *is* F-independent of C*.*

The prediction dimension of F *(denoted* $\dim_p F$*) is the supremum of the set of* $n \in \mathbb{N}$ *for which there is a sequence* $\{(a_k \in A, b_k \in B)\}_{k \in [n]}$ *s.t. for every* $k \in [n]$*,* a_k *is* F*-independent of* $\{(a_j, b_j)\}_{j \in [k]}$.

Fix non-empty finite sets S (the set of states) and A (the set of actions). Denote $X := S \times [0, 1]$, where the second factor is regarded as the space of reward values. Given $e : S \times A \to X$, $T^e : S \times A \to S$ and $R^e : S \times A \to [0, 1]$ s.t.

$e(s,a) = (T^e(s,a), R^e(s,a))$, we regard T^e as the (deterministic) transition kernel and R^e as the reward function associated with "DDP hypothesis" e. This allows us to speak of the dimension of a DDP hypothesis class (i.e. some $H \subseteq \{S \times A \to X\}$). We now give some examples for dimensions of particular function/hypothesis classes.

Proposition 1

Given any A, B *and* $F \subseteq \{A \to B\}$, *we have* $\dim_p F < |F|$.

Proposition 2

Given any A, B *and* $F \subseteq \{A \to B\}$, *we have* $\dim_p F \leq |A|$. *In particular, given* H *as above,* $\dim H \leq |S| \cdot |A|$.

Proposition 3

We now consider deterministic Markov decision processes that are cellular automata. *Consider finite sets* X *(the set of cells),* M *(the set of neighbor types) and a mapping* $\nu : X \times M \to X$ *(which cell is the neighbor of which cell). For example,* M *might be a subset of a group acting on* X. *More specifically,* X *can be* $(Z/nZ)^d$ *acting on itself, corresponding to a* d*-dimensional toroidal cellular automaton of size* n.

Consider another set C *(the set of cell states) and suppose that* $S = \{X \to C\}$. *Given any* $s \in S$, *and* $x \in X$, *define* $s_x : X \to C$ *by* $s_x(m) := s(\nu(x,m))$. *Given any* $T : \{M \to C\} \times A \to C$, *define* $T^{glob} : S \times A \to S$ by $T^{glob}(s,a)(x) := T(s_x, a)$. *Given any* $R : \{M \to C\} \to [0,1]$, *define* $R^{glob} : S \to [0,1]$ by $R^{glob}(s) := \frac{1}{|X|} \sum_{x \in X} R(s_x)$. *Define* H *by*

$$H := \{ (T^{glob}, R^{glob}) \mid T : \{M \to C\} \times A \to C, \; R : \{M \to C\} \to [0,1] \}$$

That is, H *is the set of transition kernels and reward functions that are* local *in the sense defined by* ν. *Then,* $\dim_p H \leq |C|^{|M|}(|A|+1)$.

In Proposition 3, it *might* seem like, although the rules of the automaton are local, the influence of the agent is necessarily *global*, because the dependence on the action appears in all cells. However, this is not really a restriction: the state of the cells can encode a particular location for the agent and the rules might be such that the agent's influence is local around this location. More unrealistic is the *full observability*. Dealing with partially observable cellular automata is outside the scope of this essay.

The central lemma in the proof of the regret bound for RL is a regret bound in its own right, in the setting of (deterministic) *contextual bandits*. Since this lemma might be of independent interest, we state it already here.

Let S (contexts), A (arms) and O (outcomes) be non-empty sets. Fix a function $R : O \to [0,1]$ (the reward function).

For any $c \in S^\omega$ (a fixed sequence of contexts), $e : S \times A \to O$ (outcome rule) and $\pi : (S \times O)^* \times S \to A$ (policy), we define $ec\pi \in \Delta O^\omega$ to be the resulting distribution over outcome histories. Given $\gamma \in [0,1)$, we define $U_\gamma : O^\omega \to [0,1]$ (the utility function) by

$$U_\gamma(o) := (1-\gamma) \sum_{n=0}^{\infty} \gamma^n R(o_n)$$

Lemma 1

Consider a countable non-empty set of hypotheses $H \subseteq \{S \times A \to O\}$ *and some* $\zeta \in \Delta H$ *(the prior). For each* $s \in S$*, define* $H_s \subseteq \{A \to O\}$ *by*

$$H_s := \{ e : A \to O \mid e(a) = e'(s,a),\ e' \in H \}$$

Let $D := \max_{s \in S} \dim_p H_s$ *and suppose that* A *is countable* [this assumption is to simplify the proof and is not really necessary]. *Then, there exists* $\pi^\dagger : (S \times O)^* \times S \to A$ *s.t. for any* $c \in S^\omega$ *and* $\gamma \in (0,1)$

$$\mathop{E}_{e \sim \zeta} [(1-\gamma) \sum_{n=0}^{\infty} \gamma^n \max_{a \in A} R(e(c_n, a)) - \mathop{E}_{e c \pi^\dagger} [U_\gamma]] \leq \sqrt{\frac{16 D H(\zeta)}{\ln 2} (1-\gamma)}$$

Note that the expression on the left hand side is the Bayesian regret. On the right hand side, H(ζ) stands for the Shannon entropy of ζ. In particular we have

$$H(\zeta) \leq \ln |H| \leq |S| \cdot |A| \ln |O|$$

Also, it's not hard to see that $|H| \leq |O|^{\dim_p H} \leq |O|^{D|S|}$ and therefore

$$D H(\zeta) \leq (\dim_p H)^2 \ln |O|$$

$$D H(\zeta) \leq D^2 |S| \ln |O|$$

Finally, the policy $\pi^\dagger$ we actually consider in the proof is Thompson sampling.

Now we proceed to studying reinforcement learning. First, we state a regret bound for RL *with* resets. Now S stands for the set of states and A for the set of actions. We fix a sequence of initial states $c \in S^\omega$. For any $e : S \times A \to X$ (environment), $\pi : X^* \times X \to A$ (policy), and $T \in \mathbb{N}^+$ we define $e\pi \in \Delta X^\omega$ to be the resulting distribution over histories, assuming that the state is reset to c_n and the reward to 0 every time period of length $T+1$. In particular, we have

$$\mathop{\Pr}_{x \sim e c[T] \pi} [\forall n \in \mathbb{N} : x_{n(T+1)} = (c_n, 0)] = 1$$

Given $\gamma \in [0,1)$ we define $U_\gamma : X^\omega \to [0,1]$

$$U_\gamma(s\, r) := (1-\gamma) \sum_{n=0}^{\infty} \gamma^n r_n$$

Theorem 1

Consider a countable non-empty set of hypotheses $H \subseteq \{S \times A \to X\}$ *and some* $\zeta \in \Delta H$*. Let* $D := \dim_p H$*. Then, for any* $T \in \mathbb{N}^+$ *and* $\gamma \in [0,1)$*, there exists* $\pi^\dagger_\gamma : X^* \times X \to A$ *s.t. for any* $c \in S^\omega$

$$\mathop{\mathbb{E}}_{e\sim\zeta}\left[\max_{\pi:X^*\times X\to A}\mathop{\mathbb{E}}_{e\boxtimes[T]\pi}[U_\gamma]-\mathop{\mathbb{E}}_{e\boxtimes[T]\pi^\dagger_{T,\gamma}}[U_\gamma]\right]\leq\sqrt{\frac{16DH(\zeta)(T+1)(1-\gamma)}{T}}$$

Note that $e\boxtimes[T]\pi$ is actually a probability measure concentrated on a single history, sincee π is deterministic: we didn't make it explicit only to avoid introducing new notation.

Finally, we give the regret bound without resets. For any $e:S\times A\to X$ and $\pi:X^*\times X\to A$, we define $e\pi\in\Delta X^\omega$ to be the resulting distribution over histories, given initial state s_0 and no resets.

Theorem 2

Consider a countable non-empty set of hypotheses $H\subseteq\{S\times A\to X\}$ *and some* $\zeta\in\Delta H$*. Let* $D:=\dim_p H$*. Assume that for any* $e\in H$ and $s\in S$, $A^0_e(s)=A$ [A^0 was defined here in "Definition 1"; so was the value function $V(s,x)$ used below] *(i.e. there are no traps). For any* $\gamma\in[0,1)$ *we define* $\tau(\gamma)$ *by*

$$\tau(\gamma):=\mathop{\mathbb{E}}_{e\sim\zeta}\left[\max_{s\in S}\sup_{x\in(\gamma,1)}\left|\frac{dV^\dagger_{e,k}(s,x)}{dx}\right|\right]$$

Then, for any $\gamma\in[0,1)$ *s.t.* $1-\gamma\ll 1$ *there exists* $\pi^\dagger_\gamma:X^*\times X\to A$ *s.t.*

$$\mathop{\mathbb{E}}_{e\sim\zeta}\left[\max_{\pi:X^*\times X\to A}\mathop{\mathbb{E}}_{e\pi}[U_\gamma]-\mathop{\mathbb{E}}_{e\pi^\dagger_\gamma}[U_\gamma]\right]=O\left(\sqrt[3]{DH(\zeta)\cdot(\tau(\gamma)+1)(1-\gamma)}\right)$$

Note that $\tau(\gamma)$ *decreases* with γ so this factor doesn't make the qualitative dependence on γ any worse.

Both Theorem 1 and Theorem 2 have *anytime* variants in which the policy doesn't depend on γ at the price of a slightly (within a constant factor) worse regret bound, but for the sake of brevity we don't state them (our ultimate aim is DRL which is not anytime anyway). In Theorem 2 we didn't specify the constant, so it is actually true verbatim without the γ dependence in $\pi^\dagger_\gamma$ (but we still leave the dependence to simplify the proof a little). It is also possible to spell out the assumption on γ in Theorem 2.

Proofs

Definition A.1

Consider sets A, B *and* $F\subseteq\{A\to B\}$ *non-empty. A sequence* $\{(a_k\in A,b_k\in B)\}_{k\in[n]}$ *is said to be* F-independent*, when for every* $k\in[n]$*,* a_k *is* F*-independent of* $\{(a_j,b_j)\}_{j\in[k]}$.

Definition A.2

Consider sets A, B*. Given* $C\subseteq A\times B$ *and* $a^*\in A$*, suppose* $f,g:A\to B$ *are s.t.:*

- *For any* $(a, b) \in C$, $f(a) = g(a) = b$
- $f(a^*) = g(a^*)$

Then, f, g *are said to* shatter (C, a^*).

Given sequences $\{(a_k \in A, b_k \in B)\}_{k\in[n]}$ *and* $\{(f_k, g_k : A \to B)\}_{k\in[n]}$, (f, g) *is said to* shatter (a, b) *when for any* $k \in [n]$, (f_k, g_k) *shatters* $(\{(a_j \in A, b_j \in B)\}_{j\in[k]}, a_k)$.

Proof of Proposition 1

Consider $\{(a_k \in A, b_k \in B)\}_{k\in[n]}$ an F-independent sequence. We will now construct $G \subseteq F$ s.t. $|G| = n + 1$, which is sufficient.

By the definition of F-independence, there is a sequence $\{(f_k, g_k \in F)\}_{k\in[n]}$ that shatters (a, b). We have $f_k(a_k) \neq g_k(a_k)$ and therefore either $f_k(a_k) \neq b_k$ or $g_k(a_k) \neq b_k$. Without loss of generality, assume that for each $k \in [n]$, $f_k(a_k) \neq b_k$. It follows that for any $k \in [n]$ and $j \in [k]$, $f_j(a_j) \neq b_j = f_k(a_j)$ and therefore $f_j \neq f_k$.

If $n = 0$ then there is nothing to prove since F is non-empty, hence we can assume $n > 0$. For any $k \in [n-1]$, $f_k(a_k) \neq b_k = g_{n-1}(a_k)$ and therefore $f_k \neq g_{n-1}$. Also, $f_{n-1}(a_{n-1}) \neq g_{n-1}(a_{n-1})$ and therefore $f_{n-1} \neq g_{n-1}$. We now take $G = \{f_0, f_1 \dots f_{n-1}, g_{n-1}\}$, completing the proof. ■

Proof of Proposition 2

Consider $\{(a_k \in A, b_k \in B)\}_{k\in[n]}$ an F-independent sequence and $\{(f_k, g_k \in F)\}_{k\in[n]}$ that shatters it. For any $k \in [n]$ and $j \in [k]$, we have $f_k(a_j) = g_k(a_j)$ but $f_k(a_k) \neq g_k(a_k)$, implying that $a_j \neq a_k$. It follows that $|A| \geq n$. ■

Proof of Proposition 3

Consider $\{(s_k \in S, a_k \in A, t_k \in S, r_k \in [0,1])\}_{k\in[n]}$ an H-independent sequence and $\{(T_k^{glob}, R_k^{glob}, \tilde{T}_k^{glob}, \tilde{R}_k^{glob})\}_{k\in[n]}$ that shatters it. Define $A, B \subseteq [n]$ by

$$A := \{k \in [n] \mid T_k^{glob}(s_k, a_k) \neq \tilde{T}_k^{glob}(s_k, a_k)\}$$

$$B := \{k \in [n] \mid R_k^{glob}(s_k) \neq \tilde{R}_k^{glob}(s_k)\}$$

Since $(T^{glob}, R^{glob}, \tilde{T}^{glob}, \tilde{R}^{glob})$ shatters (s, a, t, r), we have $A \cup B = [n]$.

Consider any $k \in A$. Obviously, there is $x_k \in X$ s.t.

$$T_k^{glob}(s_k, a_k)(x_k) \neq \tilde{T}_k^{glob}(s_k, a_k)(x_k)$$

Denote $\sigma_k := (s_k)_{x_k} \in C^M$. We have $T_k(\sigma_k, a_k) \neq \tilde{T}_k(\sigma_k, a_k)$. On the other hand, for any $j \in A \cap [k]$, the shattering implies $T_k^{glob}(s_j, a_j) = \tilde{T}_k^{glob}(s_j, a_j)$ and in particular $T_k(\sigma_j, a_j) = \tilde{T}_k(\sigma_j, a_j)$. Therefore, $(\sigma_k, a_k) \neq (\sigma_j, a_j)$. We conclude that $|A| \leq |C|^{|M|} \cdot |A|$.

Now consider any $k \in B$. Define $f_k \in R^{\{M \to C\}}$ by

$$(f_k)_\sigma := \frac{1}{|\{x \in X \mid (s_k)_x = \sigma\}|}$$

We have

$$f_k \cdot R_k = R_k^{glob}(s_k) \neq \tilde{R}_k^{glob}(s_k) = f_k \cdot \tilde{R}_k$$

$$f_k \cdot (R_k - \tilde{R}_k) \neq 0$$

(These are dot products in $R^{\{M \to C\}}$.) On the other hand, for any $j \in B \cap [k]$, the shattering implies $f_j \cdot R_k = f_j \cdot \tilde{R}_k$ and therefore $f_j \cdot (R_k - \tilde{R}_k) = 0$. Therefore, f_k is *not* in the linear span of $\{f_j\}_{j \in B \cap [k]}$ and hence the set $\{f_k\}_{k \in B}$ is linearly independent. We conclude that $|B| \leq \dim R^{\{M \to C\}} = |C|^{|M|}$.

Putting everything together, we get $n \leq |A| + |B| \leq |C|^{|M|}(|A| + 1)$. ■

Proposition A.1

Consider a countable set A*, a set* B*, a non-empty countable set* $F \subseteq \{A \to B\}$*, some* $f^* : A \to B$*, some* $\zeta \in \Delta F$ *and some* $\xi \in \Delta A$*. Denote*

$$p := \Pr_{(a,f) \sim \xi \times \zeta}[f(a) \neq f^*(a)]$$

Then, for any $q \in (0,1)$*, we can choose* $A^\circ \subseteq A$ *and* $F^\circ \subseteq F$ *s.t.* $\xi(A^\circ) \geq 1 - q$*,* $\zeta(F^\circ) \geq 1 - \frac{p}{q}\dim_p F$ *and for any* $f, g \in F^\circ$ *and* $a \in A^\circ$*,* $f(a) = g(a)$*.*

Proof of Proposition A.1

We define A° by

$$A^\circ := \left\{ a \in A \;\middle|\; \Pr_{f \sim \zeta}[f(a) \neq f^*(a)] \leq \frac{p}{q} \right\}$$

The fact that $\xi(A^\circ) \geq 1 - q$ follows from the definition of p.

Denote $n := |A^\circ| \in \mathbb{N} \cup \{\omega\}$ and enumerate A° as $A^\circ = \{a_k\}_{k \in [n]}$. For each $k \in [n]$, define $F_k \subseteq F$ recursively by

$$F_0 := F$$

$$F_{k+1} := \begin{cases} F_k & \text{if } \forall f,g \in F_k : f(a_k) = g(a_k) \\ \{ f \in F_k \mid f(a_k) = f^*(a_k) \} & \text{otherwise} \end{cases}$$

Set $F^\circ := \bigcap_{k \in [n]} F_k$. Define $I \subseteq [n]$ by

$$I := \{ k \in [n] \mid F_{k+1} \neq F_k \}$$

Denote $m := |I|$. For any $i \in [m]$, we denote by k_i the i-th number in I in increasing order. Denote $a_i^* := a_{k_i}$ and $b_i^* := f^*(a_i^*)$. By the definition of F_k, for each $i \in [m]$ we can choose $f_i, g_i \in F_{k_i}$ s.t. $f_i(a_i^*) \neq g_i(a_i^*)$. Moreover, it also follows from the definition that for every $i \in [m]$ and $j \in [i]$, $f_i(a_j^*) = g_i(a_j^*) = b_j^*$ (using only the fact that $f_i, g_i \in F_{k_i}$ and $k_i > k_j$). Therefore, (f, g) shatters (a^*, b^*) and hence $m \leq \dim_p F$.

By the definition of A°, for any $k \in I$, $\zeta(F_{k+1}) \geq \zeta(F_k) - \frac{p}{q}$. On the other hand, for any $k \in [n] \setminus I$, $F_{k+1} = F_k$ and in particular $\zeta(F_{k+1}) = \zeta(F_k)$. We conclude that

$$\zeta(F^\circ) \geq \zeta(F_0) - \frac{p}{q} m \geq 1 - \frac{p}{q} \dim_p F \qquad \blacksquare$$

Proposition A.2

Consider a countable set A, *a set* B, *a non-empty countable set* $F \subseteq \{A \to B\}$, *some* $\zeta \in \Delta F$ *and some* $\xi \in \Delta A$.

Consider $f^* : A \to B$ s.t.

$$f^*(a) \in \arg\max_{b \in B} \Pr_{f \sim \zeta} [f(a) = b]$$

Then,

$$\Pr_{(a,f) \sim \xi \times \zeta} [f(a) \neq f^*(a)] \leq \frac{1}{\ln 2} I_{(a,f) \sim \xi \times \zeta} [f; a, f(a)]$$

[$I[f; a, f(a)]$ stands for the *mutual information* between f and the joint random variables $a, f(a)$.]

Proof of Proposition A.2

By the chain rule for mutual information

$$I_{(a,f) \sim \xi \times \zeta} [f; a, f(a)] = I_{(a,f) \sim \xi \times \zeta} [f; a] + I_{(a,f) \sim \xi \times \zeta} [f; f(a) \mid a]$$

The first term on the right hand side obviously vanishes, and the second term can be expressed in terms of KL-divergence. For any $a \in A$, Define $ev_a : F \to B$ by $ev_a(f) := f(a)$. We get

$$I_{(a,f) \sim \xi \times \zeta} [f; a, f(a)] = E_{(a,f) \sim \xi \times \zeta} [D_{KL}(\delta_{f(a)} \| ev_{a*} \zeta)] = E_{(a,f) \sim \xi \times \zeta} \left[\ln \frac{1}{\zeta(ev_a^{-1}(f(a)))} \right]$$

$$I_{(a,f)\sim\xi\times\zeta}[f;a,f(a)] \geq E_{(a,f)\sim\xi\times\zeta}\left[\ln \frac{1}{\zeta\left(ev_a^{-1}(f(a))\right)};\ f(a)\neq f^*(a)\right]$$

For any $a \in A$ and $f \in F$, $\zeta(ev_a^{-1}(f(a))) \leq \zeta(ev_a^{-1}(f^*(a)))$ by definition of f^*. If $f(a) \neq f^*(a)$, then

$$\zeta\left(ev_a^{-1}(f(a))\right) + \zeta\left(ev_a^{-1}(f^*(a))\right) = \zeta\left(ev_a^{-1}(\{f(a),f^*(a)\})\right) \leq 1$$

It follows that, in this case, $\zeta(ev_a^{-1}(f(a))) \leq \frac{1}{2}$. We conclude

$$I_{(a,f)\sim\xi\times\zeta}[f;a,f(a)] \geq (\ln 2)\ Pr_{(a,f)\sim\xi\times\zeta}[f(a)\neq f^*(a)] \quad \blacksquare$$

Proposition A.3

Consider a countable set A*, a set* O*, a non-empty countable set* $H \subseteq \{A \to O\}$*, some* $\zeta \in \Delta H$ *and some* $R : O \to [0,1]$*.*

Let $\Pi : H \to A$ *be s.t.*

$$\Pi(e) \in \underset{a\in A}{arg\,max}\ R(e(a))$$

Then,

$$E_{e\sim\zeta}\left[\max_{a\in A} R(e(a)) - E_{a\sim\Pi_*\zeta}[R(e(a))]\right] \leq 4\sqrt{\frac{\dim_p H}{\ln 2}\cdot I_{(a,e)\sim\Pi_*\zeta\times\zeta}[e;a,e(a)]}$$

Proof of Proposition A.3

Denote $\xi := \Pi_*\zeta$, $D := \dim_p H$ and $\Gamma := I_{(a,e)\sim\xi\times\zeta}[e;a,e(a)]$. By Proposition A.2, there is $e^* : A \to O$ s.t.

$$Pr_{(a,e)\sim\xi\times\zeta}[e(a)\neq e^*(a)] \leq \frac{\Gamma}{\ln 2}$$

By Proposition A.1 (setting $q := \sqrt{\frac{\Gamma}{\ln 2}}$), there are $A^\circ \subseteq A$ and $H^\circ \subseteq H$ s.t. $\xi(A^\circ) \geq 1 - \sqrt{\frac{\Gamma}{\ln 2}}$, $\zeta(H^\circ) \geq 1 - \sqrt{\frac{\Gamma}{\ln 2}}$ and for any $e_1, e_2 \in H^\circ$ and $a \in A$, $e_1(a) = e_2(a)$. Define $H^* := H^\circ \cap \Pi^{-1}(A^\circ)$. We have $\zeta(H^*) \geq 1 - 2\sqrt{\frac{\Gamma}{\ln 2}}$.

Denote $L := E_{e\sim\zeta}[\max_{a\in A} R(e(a)) - E_{a\sim\xi}[R(e(a))]]$. For brevity, we will also use the notation $E_e^*[X] := E_{e\sim\zeta}[X; e \in H^*]$.

We get, using the bound on $\zeta(H^*)$

$$L \leq E_e^*\left[\max_{a\in A} R(e(a)) - E_{a\sim\xi}[R(e(a))]\right] + 2\sqrt{\frac{\Gamma D}{\ln 2}}$$

$$L \leq E_e^*\left[E_{e'}^*[R(e(\Pi(e))) - R(e(\Pi(e')))]\right] + 4\sqrt{\frac{\Gamma D}{\ln 2}}$$

Using the properties of H° and A° we get

$$L \leq \mathop{E}_{e}[\mathop{E}_{e'}[R(e'(\Pi(e))) - R(e'(\Pi(e')))]] + 4\sqrt{\frac{D}{n}2}$$

The expression inside the expected values in the first term on the right hand side is negative, by the property of Π. We conclude

$$L \leq 4\sqrt{\frac{D}{n}2} \qquad \blacksquare$$

Proof of Lemma 1

For each $s \in S$ we choose some $\Pi^s : H \to A$ s.t.

$$\Pi^s(e) \in \mathop{arg\,max}_{a \in A} R(e(s,a))$$

We now take $\pi^\dagger$ to be Thompson sampling. More formally, we construct a probability space (Ω, P) with random variables $K : \Omega \to H$ (the true environment) and $\{J_n : \Omega \to \{S^n \to H\}\}_{n \in N}$ (the hypothesis sampled on round n). We define $\{A_n : \Omega \to \{S^{n+1} \to A\}\}_{n \in N}$ (the action taken on round n) and $\{\Theta_n : \Omega \to \{S^{n+1} \to O\}\}_{n \in N}$ (the observation made on round n) by

$$A_n(s) := \Pi^{s_n}(J_n(s_{:n}))$$

$$\Theta_n(s) := K(s_n, A_n(s))$$

We also define $\{H_n : \Omega \to \{S^n \to H\}\}_{n \in N}$ by

$$H_n(s) := \{e \in H \mid \forall m \in [n] : e(s_m, A_m(s_{:m})) = \Theta_n(s_{:m})\}$$

H_n is thus the set of hypotheses consistent with the previous outcomes $K(s_m, \Pi^{s_m}(J_m))$. The random variables K, J have to satisfy $K_* P = \zeta$ (the distribution of the true environment is the prior) and

$$\Pr[J_n(s) = e \mid K, \{J_m(s_{:m})\}_{m \in [n]}] = \Pr_{\zeta}[e \mid H_n(s)]$$

That is, the distribution of J_n conditional on K and J_m for $m \in [n]$ is given by the prior ζ updated by the previous outcomes. $\pi^\dagger$ is then defined s.t. for any $so \in (S \times O)^n$, $t \in S$ and $a \in A$:

$$\pi^\dagger(a \mid so, t) := \Pr[A_n(st) = a \mid \forall m \in [n] : o_m = \Theta_m(s_{:m})]$$

Now we need to prove the regret bound. We denote the Bayesian regret by R:

$$R(c, \gamma) := \mathop{E}_{e \sim \zeta}[(1-\gamma)\sum_{n=0}^{\infty} \gamma^n \max_{a \in A} R(e(c_n, a)) - \mathop{E}_{e\, c\, \pi^\dagger}[U_\gamma]]$$

The construction of $\pi^\dagger$ implies that

$$R(c,\gamma) = (1-\gamma)\sum_{n=0}^{\infty} \gamma^n E\left[\max_{a\in A} R(K(c_n,a)) - R(K(c_n, A_n(c_{:n+1})))\right]$$

Define $\{Z_n : \Omega \to \{S^n \to \Delta H\}\}_{n\in N}$ (the belief state of the agent on round n) and $\{\Xi_n : \Omega \to \{S^{n+1} \to \Delta A\}\}_{n\in N}$ (the distribution over actions on round n) by

$$Z_n(s) := \zeta \mid H_n(s)$$

$$\Xi_n(s) := \Pi_*^{s_n} Z_n(s)$$

Using expectation conditional on Z_n we can rewrite the equation for R(γ) as

$$R(c,\gamma) = (1-\gamma)\sum_{n=0}^{\infty} \gamma^n E\left[\mathop{E}_{e\sim Z_n(c_{:n})}\left[\max_{a\in A} R(e(c_n,a)) - \mathop{E}_{a\sim \Xi_n(c_{:n+1})}\left[R(e(c_n,a))\right]\right]\right]$$

It follows that

$$R(c,\gamma) \le \sqrt{(1-\gamma)\sum_{n=0}^{\infty} \gamma^n E\left[\left(\mathop{E}_{e\sim Z_n(c_{:n})}\left[\max_{a\in A} R(e(c_n,a)) - \mathop{E}_{a\sim \Xi_n(c_{:n+1})}\left[R(e(c_n,a))\right]\right]\right)^2\right]}$$

We now apply Proposition A.3 and get

$$R(c,\gamma) \le \sqrt{\frac{16D^2}{\ln 2}\cdot(1-\gamma)\sum_{n=0}^{\infty} \gamma^n E\left[\mathop{I}_{(a,e)\sim \Xi_n(c_{:n+1})\times Z_n(c_{:n})}[e; a, e(a)]\right]}$$

$$R(c,\gamma) \le \sqrt{\frac{16D^2}{\ln 2}\cdot(1-\gamma)\sum_{n=0}^{\infty} \gamma^n E\left[H(Z_n(c_{:n})) - H(Z_{n+1}(c_{:n+1}))\right]}$$

$$R(c,\gamma) \le \sqrt{\frac{16D^2}{\ln 2} H(\zeta)(1-\gamma)} \qquad \blacksquare$$

Given any $x \in X$, we will use the notation $x = (\mathrm{st}\, x \in S, \mathrm{rw}\, x \in [0,1])$.

Proposition A.4

Fix $T \in N^+$ *and consider a non-empty set of DDP hypotheses* $H \subseteq \{S \times A \to X\}$. *For each* $e : S \times A \to X$ *we define* $e^\sharp : S \times A^T \to X^T$ *recursively by*

$$e^\sharp(s,\pi)_0 := e(s,\pi_0)$$

$$e^\sharp(s,\pi)_{n+1} = e(\mathrm{st}\, e^\sharp(s,\pi)_n, \pi_{n+1})$$

(That is, $e^\sharp(s,\pi)$ *is the history resulting from DDP* e *interacting with action sequence* π *starting from initial state* s *.) Define*

$$H^\sharp := \{ e^\sharp \mid e \in H \}$$

Then, for any $s \in S$, $\dim_p H^\sharp_s \leq \dim_p H$.

Here, the subscript s has the same meaning as in Lemma 1.

Proof of Proposition A.4

Fix $s \in S$. Let $\{(\pi_k \in A^\sharp, h_k \in X^T)\}_{k\in[n]}$ be an $H^\sharp_s$-independent sequence. We will now construct a sequence

$$\{ (s^*_k \in S, a^*_k \in A, x^*_k \in X, e_k \in H, \tilde{e}_k \in H) \}_{k\in[n]}$$

s.t $(e, \tilde{e})$ shatters (s^*, a^*, x^*). The latter implies that (s^*, a^*, x^*) is an H-independent sequence, establishing the claim.

For brevity, we will use the notation $f\pi := f^\sharp(s,\pi)$ (here, $f : S \times A \to X$ and $\pi : S \times \mathbb{N} \to A$). For each $k \in [n]$ define $m_k \in [T]$ by

$$m_k := \min \{ l \mid \exists f, \tilde{f} \in H : (f\pi_k)_{:l+1} \neq (\tilde{f}\pi_k)_{:l+1} \wedge \forall j \in [k] : f\pi_j = \tilde{f}\pi_j = h_j \}$$

Note that the set above is indeed non-empty because (π, h) is $H^\sharp_s$-independent.

Given $g \in X^T$, we will use the notational convention $\mathrm{st}\, g_{-1} := s$.

Choose $e_k, \tilde{e}_k$ s.t. $(e_k\pi_k)_{:m_k+1} \neq (\tilde{e}_k\pi_k)_{:m_k+1}$ and $\forall j \in [k] : e_k\pi_j = \tilde{e}_k\pi_j = h_j$. Obviously $(e_k\pi_k)_{:m_k} = (\tilde{e}_k\pi_k)_{:m_k}$. We set $s^*_k := \mathrm{st}\,(e_k\pi_k)_{m_k-1}$ and $a^*_k := (\pi_k)_{m_k}$. Clearly $e_k(s^*_k, a^*_k) \neq \tilde{e}_k(s^*_k, a^*_k)$.

If $k < n-1$ then there is $f \in H$ s.t. $\forall j \in [k+1] : f\pi_j = h_j$ (by independence). Therefore, $\mathrm{st}\,(h_k)_{m_k-1} = s^*_k$: otherwise $\mathrm{st}\,(f\pi_k)_{m_k-1} \neq \mathrm{st}\,(e_k\pi_k)_{m_k-1}$ and we get a contradiction with the minimality of m_k. We now set $x^*_k := (h_k)_{m_k}$. If $k = n$, we choose $x^*_k \in X$ arbitrarily. For any $k \in [n]$ and $j \in [k]$, $e_k\pi_j = \tilde{e}_k\pi_j = h_j$ and therefore $e_k(s^*_k, a^*_k) = \tilde{e}_k(s^*_k, a^*_k) = x^*_k$.

Proof of Theorem 1

Let $A^\sharp := A^T$, $O^\sharp := X^T$ and $H^\sharp$ be defined as in Proposition A.4. By Proposition A.4, we have $\max_{s \in S} \dim_p H^\sharp_s \leq D$.

Define $\zeta^\sharp \in \Delta H^\sharp$ as the pushforward of ζ by the $\sharp$ operator. Define $R^\sharp_\gamma : O^\sharp \to [0,1]$ by

$$R^\sharp_\gamma(h) := \frac{\gamma(1-\gamma^T)}{1-\gamma^{T+1}} \sum_{n=0}^{T-1} \gamma^n \mathrm{rw}(h_n)$$

Denote also $\gamma^\sharp := \gamma^{T+1}$. It is easy to see that given any $\eta \in O^{\sharp\omega}$

$$U_\gamma\left(\prod_{n=0}^{\infty} (c_n, 0)\, \eta_n\right) = (1-\gamma^\sharp) \sum_{n=0}^{\infty} \gamma^{\sharp n} R^\sharp_\gamma(\eta_n)$$

The product is in the string concatenation sense.

Applying Lemma 1 to all the "$\sharp$" objects (with S and c unaffected), we get $\pi^\dagger_{T,\gamma}$ s.t.

$$\mathop{\mathrm{E}}_{e \sim \zeta}\left[\max_{\pi : X^* \times X \to A} \mathop{\mathrm{E}}_{ec[T]\pi}[U_\gamma] - \mathop{\mathrm{E}}_{ec[T]\pi^\dagger_{T,\gamma}}[U_\gamma]\right] \leq \sqrt{\frac{16 D H(\zeta^\sharp)}{\ln 2}(1-\gamma^\sharp)}$$

Here, we used the fact that the optimal policy for a DDP with fixed initial state (and any time discount function) can be taken to be a fixed sequence of actions.

Observing that $H(\zeta^\sharp) \leq H(\zeta)$ and $1 - \gamma^{T+1} \leq (T+1)(1-\gamma)$, we get the desired result. ■

The following is a minor variant of what was called "Proposition B.2" here, and we state it here without proof (since the proof is essentially the same).

Proposition A.5

Consider a DDP $e : S \times A \to X$, *policies* $\pi^0, \pi^* : X^* \times X \to A$ *and* $T \in \mathbb{N}^+$. *Suppose that for any* $h \in X^*$ and $x \in X^k$

$$\mathrm{supp}\ \pi^0(h, s) \subseteq A^0_e(s)$$

Suppose further that π^* *is optimal for* e *with time discount* γ *and any initial state. Denote*

$$\tau_e(\gamma) := \max_{s \in S} \sup_{x \in (\gamma, 1)} \left| \frac{dV^{\pi^*}_{e,\gamma}(s, x)}{dx} \right|$$

Given any $s \in S$ *and policy* π, *denote* $es\pi \in \Delta X^\omega$ *the distribution over histories of resulting from* π *interacting with* e *starting from initial state* s. *Finally, define* $U_{T,\gamma} : X^\omega \to [0,1]$ *by*

$$U_{T,\gamma}(h) := \frac{1-\gamma}{1-\gamma^T} \sum_{n=0}^{T-1} \gamma^n \mathrm{rw}(h_n)$$

Then

$$\mathop{\mathbb{E}}_{e \sim \zeta}\left[U_\gamma(\pi^*)\right] \leq (1-\gamma^T) \sum_{n=0}^{\infty} \gamma^{nT} \mathop{\mathbb{E}}_{h \sim e\pi^0}\left[\mathop{\mathbb{E}}_{e\,\mathrm{st}\,h_{nT}\pi^*}\left[U_{T,\gamma}\right]\right] + \frac{2\tau_e(\gamma)\gamma^T}{1-\gamma^T}(1-\gamma)$$

Proof of Theorem 2

It is not hard to see that Theorem 1 can be extended to the setting where the initial state sequence $c \in S^\omega$ is chosen in a way that depends on the history. This is because if c is chosen *adversarially* (i.e. in order to maximize regret), the history doesn't matter (in other words, we get a repeated zero-sum game in which, on each round, the initial state is chosen by the adversary and the policy is chosen by the agent *after* seeing the initial state; this game clearly has a pure Nash equilibrium). In particular, we can let c be just the states naturally resulting from the interaction of the DDP with the policy.

Let $\pi^*_{e\gamma}$ be the optimal policy for DDP e and time discount γ. We get

$$\mathop{\mathbb{E}}_{e \sim \zeta}\left[(1-\gamma^T) \sum_{n=0}^{\infty} \gamma^{nT} \mathop{\mathbb{E}}_{h \sim e\pi^\dagger_{T,\gamma}}\left[\mathop{\mathbb{E}}_{e\,\mathrm{st}\,h_{nT}\pi^*_{e\gamma}}\left[U_{T,\gamma}\right]\right] - \mathop{\mathbb{E}}_{e\pi^\dagger_{T-1,\gamma}}\left[U_\gamma\right]\right] \leq \sqrt{\frac{16 D_{KL}(\zeta) T}{\ln 2}(1-\gamma) + \frac{1-\gamma}{1-\gamma^T}}$$

Here, the second term on the right hand side comes from the rewards at time moments divisible by T, which were set to 0 in Theorem 1. Denote

$$R(T,\gamma) := \mathop{\mathbb{E}}_{e \sim \zeta}\left[\max_{\pi : X^* \times X \to A} \mathop{\mathbb{E}}_{e\pi}\left[U_\gamma\right] - \mathop{\mathbb{E}}_{e\pi^\dagger_{T-1,\gamma}}\left[U_\gamma\right]\right]$$

Applying Proposition A.5, we get

$$R(T,\gamma) \leq \sqrt{\frac{16 D_{KL}(\zeta) T}{\ln 2}(1-\gamma) + \frac{1-\gamma}{1-\gamma^T}} + \frac{2\tau_e(\gamma)\gamma^T}{1-\gamma^T}(1-\gamma)$$

Assume that $\gamma^{T-1} \geq \frac{1}{2}$. Then

$$\frac{1-\gamma}{1-\gamma^T} = \frac{1}{\sum_{n=0}^{T-1} \gamma^n} \leq \frac{2}{T}$$

$$R(T,\gamma) \leq \sqrt{\frac{16 D_{KL}(\zeta) T}{\ln 2}(1-\gamma) + \frac{4\tau_e(\gamma)}{T}} + 2$$

Now set

$$T := \left\lceil \frac{\sqrt[3]{(\tau(\gamma)+1)^2}}{D\,H(\zeta)\cdot(1-\gamma)} \right\rceil$$

It is easy to see that the assumption $\gamma^{T-1} \geq \frac{1}{e}$ is now justified for $1-\gamma \ll 1$. We get

$$R(T,\gamma) = O\left(\sqrt[3]{D\,H(\zeta)\cdot(\tau(\gamma)+1)(1-\gamma)}\right)$$

A Short Note on UDT

In my last post, I stumbled across some ideas which I thought were original, but which were already contained in UDT. I suspect that was because these ideas haven't been given much emphasis in any of the articles I've read about UDT, so I wanted to highlight them here.

We begin with some definitions. Some inputs in an Input-Output map will be possible for some agents to experience, but not for others. We will describe such inputs and the situations they represent as conditionally consistent. Given a particular agent, we will call a input/situation compatible if the agent is consistent with the corresponding situation and incompatible otherwise. We will call agents consistent with a conditionally consistent input/situation compatible and those who aren't incompatible.

We note the following points:

- UDT uses an Input-Output map instead of a Situation-Output map. It is easy to miss how important this choice is. Suppose we have an input representing a situation that is conditionally consistent. Trying to ask what an incompatible agent does in such a situation is problematic or at least difficult as the Principle of Explosion means that all such situations are equivalent. On the other hand, it is much easier to ask how the agent responds to a sequences of inputs representing an incompatible situation. The agent must respond somehow to such an input, even if it is by doing nothing or crashing. Situations are also modelled (via the Mathematical Intuition Function), but the point is that UDT models inputs and situations separately.
- Given the previous point, it is convenient to define an agent's counterfactual action in an incompatible situation as its response to the input representing the situation. For all compatible situations, this produces the same action as if we'd simply asked what the agent would do in such a situation. For conditionally consistent situations the agent is incompatible with, it explains the incompatibility: any agent that would respond a certain way to particular inputs won't be put in such a situation. (There might be conditionally consistent situations where compatibility isn't dependent on responses to inputs, ie. only agents running particular source code are placed in a particular position, but UDT isn't designed to solve these problem)
- Similarly, UDT predictors don't actually predict what an agent does in a situation, but what an agent does when given an input representing a situation. This is a broader concept that allows them to predict behaviours in inconsistent situations. For a more formal explanation of these ideas, see Logical Counterfactuals for Perfect Predictors.

July gwern.net newsletter

This is a linkpost for https://www.gwern.net/newsletter/2018/07

Sandboxing by Physical Simulation?

This is a simple idea. I do not remember if I've seen it anywhere. It is probably not original, but I am mildly surprised that I haven't come across it, even if only to see it refuted. If this is an old/dumb idea, please let me know and I'll delete it.

People have built universal Turing machines in Minecraft. It is straightforward to build a virtualized Turing machine by simulating a physical system which carries out the mechanical actions needed to instantiate one. You could obviously build a computer in a simulated physics simpler than Minecraft, but the Minecraft example is a more vivid one.

I don't even want to guess how much more computationally expensive it would be to run an AI, and the AI's simulated environment, on a Turing machine being run on a simulated physical computer, itself being simulated on mundane hardware. But it does strike me that an AI should have significantly more trouble hacking its way out of this kind of sandboxing.

It would likely have more trouble noticing that it's in a simulation in the first place, and if it did notice, it would likely have a much harder time doing anything about it. Our hardware relies on physical laws that we don't full understand. It makes errors that we sometimes can't really notice or stop, because fundamentally it's made of atoms and fields. One can imagine ways in which a clever being might intentionally strain physical hardware to see how it reacts. We already know that an attacker can induce a bit flip in neighboring memory through physics trickery. In contrast, the underlying simulated physics of the AI world/brain could be extremely simple and designed to be as free of surprises as possible. Within the doubly-simulated world, and within its own simulated brain, there would be no stray EM fields, no Internet ports accidentally left open. The AI could bang on the strata of its physics all it wanted, and all "we" would see would be the flickering light of the Redstone machine.

I'm not one to underestimate a superintelligence, but the baseline security of this kind of double-Sandboxing feels qualitatively different than that of physical hardware.

Learning strategies and the Pokemon league parable

I have recently noted a shift in my learning strategy, which I reflectively approve of. On hindsight, it feels obvious.

However, I can vividly recall many people I respect and admire recommending me to try a very similar thing in the past, and myself scoffing at them and trusting my gut over their advice.

Take this as a word of caution if you feel like the advice I am giving is obviously wrong: I have fallen in this pit, and it took me a while to climb out.

I claim there are two broad learning strategies one can follow.

The default strategy is what in the software engineering lingo is called a *waterfall* strategy. People attend college and different courses and read books, and gain knowledge on a broad collection of subjects. Afterwards, they move to a second phase where they try to apply what they have learned. If they cannot reach their goals with their current strategy, they back down to the first phase and start again.

I claim that this strategy has some glaring flaws, which I plan to expose via my experience in an area of great importance: Pokemon videogames.

When I got my first videoconsole, the first videogame I ever owned was Pokemon Gold. I absolutely loved that game, and I spent many hours absorbed trying to complete it.

For the most part, the level of challenge was adequate for an 8 year old, but there comes a point where the game suddenly spikes up in difficulty: the Pokemon league. In the league, you have to defeat five trainers with full teams of high level Pokemon in a row.

When I first confronted the Pokemon league, I was quite under leveled and I was utterly crushed.

My response to the problem was to back down and go to easier areas, where I could train with easier challenges.

After around 10h of training, I came back to the League and defeated the five trainers with relatively ease, and I won the title of Pokemon master, officially achieving my most ambitious 8 year old goal.

"Well Jaime", you may say, "That does not look like your learning strategy had a problem. You successfully completed the challenge!"

And yes, I did. But it was awfully inefficient! I leveled at a very slow rate by fighting easier fights, I ended up training for too long and I picked up tactics to fight other

trainers instead of learning the tactics that would have been optimal to fight through the league.

If instead I had kept fighting the trainers in the league, I would have leveled up much faster and I would have learned what tactics were good against each of the elite trainers I had to fight.

I claim that similar things happen to the people like myself two years ago who apply the waterfall strategy:

- You end up wasting time learning stuff you do not need. Even worse, because of fade-out effects you will forget most of what you learned. Think about how many useful things did you learn in school that you still remember today.
- You do not learn at an optimal rate, because it's hard to calibrate for problems that are just above your current level.
- There is no clear stop point of transitioning between learning mode and solving mode, which further results in overtraining.

Is there an alternative? Yes there is! I introduce to you *project-based* learning, aka the *agile* strategy. Instead of generally learning and then solving, run head first into the problem, and learn to fail effectively. When the brick wall in front of you refuses to move, think about what you need to learn to overcoming, and while you learn the techniques make an effort to *apply them to the task of wall moving*.

You will learn more about the problem and whether the techniques you are learning are actually useful this way.

When I introspect on why I thought waterfall strategy was a better fit for me than project based learning, I find the following reasons:

- Waterfall is a good strategy when it is hard to determine what one needs to learn to solve a problem. It is easy to say "you never know if it is going to be useful" and just keep learning useless stuff. But there is virtue in precision, and I have learned to be willing to make more concrete predictions ("there is a low chance than learning French will pay off in the long run given my current trajectory").
- Humans like finding connections between disparate areas, and it feels really pleasant to connect two different things together (like for example playing Pokemon videogames and explaining the differences between strategies for learning). I am actually unsure of how much of this is true insight and how much is just confirmation bias (if I had not played Pokemon, maybe another example would have occurred to me).
- Banging your head against a wall feels really bad, and just repeatedly trying to solve a problem through sheer willpower is not a strategy I would recommend. But that is not a very good excuse to skip the first try: sometimes the wall is made of paperboard, and you want to take advantage of that.

I am running out of time to write more, and I feel I have not given enough examples.

The reflection that initiated this post was me two years ago reading a new math textbook every month vs me now taking weeks off at a time trying to solve open

research problems. I feel like the second strategy is proving much more useful to give me a feel of whether this is a good way of reaching my goals, and giving me a better map of what I need to learn to solve the research problems (turns out that it's not so much about learning technical topics but rather about learning how to write better and be organized while pursuing research directions).

I keep seeing people wanting to contribute to AI alignment research or other important research areas and resorting to waterfall strategies instead of project based learning. This post is for them.

If you have an example you want to share or any thoughts, please contribute to the conversation in the comments section!

Darmani in the comments below points out that I am confusing what is normally called project-based learning (doing projects to learn generally useful skills) with what is normally called pull-based learning (learning skills to do a concrete project). Ooops!

This post meant to recommend the second one, and the agile vs waterfall analogy was meant to point out that if you have a concrete goal you are working towards, you can use it to constantly check whether you are already in a good position to make progress on the problem you really care about.

Following his recommendation, I also want to share a note on how I generated this post:

The generator of this post is a combination of the following observations:

1) I see a lot of people who keep waiting for a call to adventure

2) Most knowledge I have acquired through life has turned out to be useless, non transferable and/or fades out very quickly

3) It makes sense to think that people get a better grasp of what skills they need to solve a problem (such as producing high quality AI Alignment research) after they have grappled with the problem. This feels specially true when you are in the edge of a new field, because there is no one else you can turn to who would be able to compress their experience in a digestible format.

4) People (specially in mathematics) have a tendency to wander around aimlessly picking up topics, and then use very few of what they learn. Here I am standing on not very solid ground, because conventional wisdom is that you need to wander around to "see the connections", but I feel like that might be just confirmation bias creeping in.

History of the Development of Logical Induction

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

I have been asked several times about how the development of logical induction happened, so I am writing it up.

June 2013 - I write my first Less Wrong Post. It may not seem very related to logical uncertainty, but it was in my head. I wanted to know how to take different probabilities for the same event and aggregate them, so I could take an agent that could maintain probabilities on facts even when facts that it originally thought were separate turn out to be logically equivalent. (I am not sure if all this was in my head at the time, but it was at some point over the next year.)

I make a proposal for a distribution on completions of a theory: repeatedly observe that a set of sentences whose probabilities should sum to one fail to sum to one, and shift their probabilities in a way inspired from the above post. I do not actually prove that this process converges, but I conjecture that it converges to the distribution I describe here. (This post is from when I wrote it up in June 2014; I can't remember exactly what parts I already had at the time.)

December 2013 - I tell my proposal to Abram Demski, who at some point says that he thinks it is either wrong or equivalent to his proposal for the same problem. (He was right and his proposal was better.) At this point I got very lucky; when I told this to Abram, I thought he was just a smart person at my local Less Wrong meet up, and it turned out that he was almost the only person to also try to do the thing I was doing. Abram and I start talking about my proposal a bunch together, and start trying to prove the above conjecture.

April 2014 - Abram and I start the first MIRIx to think about logical uncertainty, and especially this proposal. I at the time had not contacted MIRI, even to apply for a workshop, because I was dumb.

At some point we realize that the proposal is bad. The thing that makes us give up on it is the fact that sometimes observing that $A \to B$ can drastically decrease your probability for B.

August 2014 - Abram and I go to MIRI to talk about logical uncertainty with Nate, Benya, Eliezer, and Paul. We share the thing we were thinking about, even though we had given up on it at the time. At some point in there, we talk about assigning probability 1/10 to a sufficiently late digit of π being 0.

Soon after that, I propose a new project for our MIRIxLA group to work on, which I call the Benford Test. I wanted an algorithm which on day n, after thinking for some function of n time, assigned probabilities to the nth logical sentence in some enumeration of logical sentences. If I took a subsequence of logical sentences whose

truth values appeared pseudorandom to anything that ran sufficiently quickly, I wanted the algorithm to converge to the correct probability on that subsequence. I.e., it should assign probability $\log_{10}(2)$ to the first digit of Ackerman(n) being 1. The

Benford thing was to make it concrete; I was thinking about it as pseudorandomness. There are a bunch of ugly things about the way I originally propose the problem. For example, I only assign a probability to one sentence on each day. We think about this quite a bit over the next 6 months and repeatedly fail to find anything that passes the Benford test.

March 2015 - I eventually find an algorithm that passes the Benford Test, but it is really hacky and ugly. I know writing it up is going to be a chore, so I decide to ask Luke if I can go to MIRI for a summer job and work on turning it into a paper. I become a MIRI contractor instead.

May 2015 - I go to my first MIRI workshop. During the workshop, there is reserved time for writing blog posts for agentfoundations.org. I start making writing blog posts a major part of my motivation system while doing research and writing the paper.

At some point in there, I notice that the algorithm we use to pass the Benford test does not obviously converge even in the probabilities that it assigns to a single sentence. I change the framework slightly and make it cleaner so that the Benford test algorithm can be seen in the same type of framework as the Demski prior. Now I have two algorithms. One can do well on pseudorandom sentences, and the other converges to a distribution in the limit. I set my next goal to do both at the same time. One hope is that in order to do both of these tasks at the same time, we will have to use something less hacky than we used to solve the Benford test.

July 2015 - I go to the first MIRI summer fellows program, and while there, I make the first noticeable step towards combining limit coherence and the Benford test - I define inductive coherence as a subgoal. This is a strengthening of limit coherence that I believed would bring it closer to having good behaviors before reaching the limit. I do not actually find an algorithm that is inductively coherent, but I have a goal that I think will help. A couple months later, Benya shows that a modification of the Demski prior is inductively coherent.

November 2015 - I make a significantly cleaner algorithm that passes the Benford test, and present it at the Berkeley PDTAI seminar. The cleaner algorithm is based on sleeping experts, which is pretty close to things that are going on in logical induction. It involves a bundle of hypotheses that are assigning probabilities to logical sentences, but are also allowed to sleep and not assign probabilities some days if they don't want to. I also notice that reflective oracle Solomonoff induction can be used to implement some sleeping expert stuff, because the hypotheses have access to the overall distribution, and can not be on the hook on a given day by deferring to the average.

December 2015 - I start working for MIRI full time. One thing I am working on is writing up inductive coherence. At some point around here, I make a stronger version of the Benford test that my algorithm fails that I think is pushing it closer to limit coherence.

January 2016 - I figure out how to do the stronger version of the Benford test. My algorithm uses a lemma that Jessica proved for a separate thing she was doing, so we start to write up a paper on it.

February 2016 - While walking to work, I notice the linchpin for Logical Induction. I notice that the Benford test and limit coherence are both special cases of a sequence of probability assignments with the property that no gambler who can see the probabilities each day, who can choose to bet whenever they want and has a finite starting wealth, can reach infinite returns. There is one type of gambler that can take advantage of a market that fails the Benford test, and another type that can take advantage of a market that is not limit coherent.

I am very excited. I know at the time that I am likely about to solve my goal from roughly the last 10 months. I tell Abram and everyone at the MIRI office. Jessica is actually the first to point out that we can use continuity, similarly to what you do with reflective oracles. The same day, Jessica and I manage to show that you can make a market no continuous gambler can exploit as described, and verify that continuous gamblers are enough to ensure you pass the Benford test and have a coherent limit. At this point the meat of logical induction is done.

Over the next 6 months, we put a lot of effort into making the theory clean, but we didn't really have to make the algorithm better. The very first thing we found that was able to pass the Benford test and have a coherent limit had basically all the properties that you see in the final paper. Over much of this time, we were still not sure if we wanted to make logical induction public.

Four kinds of problems

I think there are four natural kinds of problems, and learning to identify them helped me see clearly what's bad with philosophy, good with start-ups, and many things in-between.

Consider these examples:

1. Make it so that bank transfers to Africa do *not* take weeks and require visiting physical offices, in order to make it easier for immigrants to send money back home to their poor families.
2. Prove that the external world exists and you're not being fooled by an evil demon, in order to use that epistemic foundation to derive a theory of how the world works.
3. Develop a synthetic biology safety protocol, in order to ensure your lab does not accidentally leak a dangerous pathogen.
4. Build a spaceship that travels faster than the speed of light, in order to harvest resources from outside our light cone.

These examples all consist in problems that are encountered as part of work on larger projects. We can classify them by asking how we should respond when they arise, as follows:

	Keep working on the problem	**Stop working on the problem**
Keep working on the project	1. Problems to be solved	2. Problems to be gotten over
Stop working on the project	3. Crucial considerations	4. Defeating problems

1. is a problem to be solved. In this particular example, it turns out global remittances are several times larger than the combined foreign aid budgets of the Western world. Building a service avoiding the huge fees charged by e.g. Western Union is a very promising way of helping the global poor.

2. is a problem to be gotten over. You probably won't find a solution of the kind philosophers usually demand. But, evidently, you don't *have to* in order to make meaningful epistemic progress, such as deriving General Relativity or inventing vaccines.

3. is a crucial consideration -- a problem so important that it might force you to drop the entire project that spawned it, in order to just focus on solving this particular problem. Upon discovering that there is a non-trivial risk of tens of millions of people dying in a natural or engineered pandemic within our lifetimes, and then realising how woefully underprepared our health care systems are for this, publishing yet another paper suddenly appears less important.

4. is a defeating problem. Solving it is impossible. If a solution forms a crucial part of a project, then the problem is going to bring that project with it into the grave. If whatever we want to spend our time doing, if it requires resources from outside our light cone, we should give it up.

With this categorisation in mind, we can understand some good and bad ways of thinking about problems.

For example, I found that learning the difference between a defeating problem and a problem-to-be-solved was what was required to adopt a “hacker mindset”. Consider the remittances problem above. If someone had posed it as something to do after they graduate, they might have expected replies like:

“Sending money? Surely that’s what banks do! You can’t just... build a bank?”

“What if you get hacked? Software infrastructure for sending money has to be crazy reliable!”

“Well, if you’re going to build a startup to help to global poor, you’d have to move to Senegal.”

Now of course, neither of these things *violate the laws of physics*. They might violate a few social norms. They might be scary. They might seem like the kind of problem an ordinary person would not be allowed to try to solve. However, if you *really* wanted to, you *could* do these things. And some less conformist people who did just that have now become billionaires or, well, moved to Senegal (c.f. PayPal, Stripe, Monzo and Wave).

As Hannibal said when his generals cautioned him that it was impossible to cross the Alps by elephant: "I shall either find a way or make one."

This is what’s good about startup thinking. Philosophy, however, has a big problem which goes the other way: mistaking problems-to-be-solved for defeating problems.

For example, a frequentist philosopher might object to Bayesianism saying something like “Probabilities can’t represent the degrees of belief of agents, because in order to prove all the important theorems you have to assume the agents are logically omniscient. But that’s an unreasonable constraint. For one thing, it requires you to have an infinite number of beliefs!” (this objection is made here, for example). And this might convince people to drop the Bayesian framework.

However, the problem here is that it has not been *formally proven* that the important theorems of Bayesianism *ineliminably require* logical omniscience in order to work. Rather, that is often *assumed*, because people find it hard to do things formally otherwise.

As it turns out, though, the problem is solvable. Philosophers did not find this out, however, as they get paid to argue and so love making objections. The proper response to that might just be “shut up and do the impossible”. (A funny and anecdotal example of this is the student who solved an unsolved problem in maths because he thought it was an exam question.)

Finally, we can be more systematic in classifying several of these misconceptions. I’d be happy to take more suggestions in the comments.

Actual problem →	**Problem to be gotten over**	**Problems to be solved**	**Defeating problems**	**Crucial considerations**
Interpreted as				
Problems to be gotten over	Great.	Scarcity mindset	Theology	?
Problems to be solved	Philosophy Perfectionism	Great.	Crackpot science	Myopic/naïve thinking
Defeating problems	Akrasia	Philosophy	Great.	Nihilism
Crucial considerations	Phobias	Mistaking efficiency for effectiveness	?	Great.

Tactical vs. Strategic Cooperation

As I've matured, one of the (101-level?) social skills I've come to appreciate is *asking directly for the narrow, specific thing you want, instead of debating around it.*

What do I mean by "debating around" an issue?

Things like:

"If we don't do what I want, horrible things A, B, and C will happen!"

(This tends to degenerate into a miserable argument over how likely A, B, and C are, or a referendum on how neurotic or pessimistic I am.)

"You're such an awful person for not having done [thing I want]!"

(This tends to degenerate into a miserable argument about each other's general worth.)

"Authority Figure Bob will disapprove if we don't do [thing I want]!"

(This tends to degenerate into a miserable argument about whether we should respect Bob's authority.)

It's been astonishing to me how much better people respond if instead I just say, "I really want to do [thing I want.] Can we do that?"

No, it doesn't guarantee that you'll get your way, but it makes it a whole lot more likely. More than that, it means that when you *do* get into negotiation or debate, that debate stays targeted to the actual decision you're disagreeing about, instead of a global fight about anything and everything, and thus is more likely to be resolved.

Real-life example:

Back at MetaMed, I had a coworker who believed in alternative medicine. I didn't. This caused a lot of spoken and unspoken conflict. There were global values issues at play: reason vs. emotion, logic vs. social charisma, whether her perspective on life was good or bad. I'm embarrassed to say I was rude and inappropriate. But it was coming from a well-meaning place; I didn't want any harm to come to patients from misinformation, and I was very frustrated, because I didn't see how I could prevent that outcome.

Finally, at my wit's end, I blurted out what I wanted: I wanted to have veto power over any information we sent to patients, to make sure it didn't contain any factual inaccuracies.

Guess what? *She agreed instantly*.

This probably should have been obvious (and I'm sure it was obvious to her.) My *job* was producing the research reports, while her jobs included marketing and operations. The whole point of division of labor is that we can each stick to our own tasks and not have to critique each other's entire philosophy of life, since it's *not relevant* to getting the company's work done as well as possible. But I was extremely inexperienced at working with people at that time.

It's not fair to your coworkers to try to alter their private beliefs. (Would you try to change their religion?) A company is an association of people who cooperate on a *local* task. They don't have to see eye-to-eye about everything in the world, so long as they can work out their disagreements about the task at hand.

This is a skill that "practical" people have, and "idealistic" and "theoretical" people are often weak at -- the ability to declare some issues *off topic*. We're trying to decide what to do in the here and now; we don't always have to turn things into a debate about underlying ethical or epistemological principles. It's not that principles don't exist (though some self-identified "pragmatic" or "practical" people are against principles *per se*, I don't agree with them.) It's that it can be unproductive to get into debates about general principles, when it takes up too much time and generates too much ill will, and when it isn't necessary to come to agreement about the tactical plan of what to do next.

Well, what about longer-term, more intimate partnerships? Maybe in a strictly professional relationship you can avoid talking about politics and religion altogether, but in a closer relationship, like a marriage, you actually *want* to get alignment on underlying values, worldviews, and principles. My husband and I spend a *ton* of time talking about the diffs between our opinions, and reconciling them, until we *do* basically have the same worldview, seen through the lens of two different temperaments. Isn't that a counterexample to this "just debate the practical issue at hand" thing? Isn't intellectual discussion really valuable to intellectually intimate people?

Well, it's complicated. Because I've found the same trick of *narrowing the scope of the argument and just asking for what I want* resolves debates with my husband too.

When I find myself "debating around" a request, it's often debating in bad faith. I'm not *actually* trying to find out what the risks of [not what I want] are in real life, I'm trying to use talking about danger as a way to scare him into doing [what I want]. If I'm quoting an expert nutritionist to argue that we should have home-cooked family dinners, my motivation is *not* actually curiosity about the long-term health dangers of not eating as a family, but simply that I want family dinners and I'm throwing spaghetti at a wall hoping *some* pro-dinner argument will work on him. The "empirical" or "intellectual" debate is just so much rhetorical window dressing for an underlying *request*. And when that's going on, it's better to notice and redirect to the actual underlying desire.

Then you can get to the actual negotiation, like: what makes family dinners undesirable to *you?* How could we mitigate those harms? What alternatives *would* work for both of us?

Debating a far-mode abstraction (like "how do home eating habits affect children's long-term health?") is often an inefficient way of debating what's *really* a near-mode practical issue only weakly related to the abstraction (like "what kind of schedule should our household have around food?") The far-mode abstract question still exists and *might* be worth getting into as well, but it also may recede dramatically in importance once you've resolved the practical issue.

One of my long-running (and interesting and mutually respectful) disagreements with my friend Michael Vassar is about the importance of local/tactical vs. global/strategic cooperation. Compared to me, he's much more likely to value getting to alignment with people on fundamental values, epistemology, and world-models. He would rather

cooperate with people who share his principles but have opposite positions on object-level, near-term decisions, than people who oppose his principles but are willing to cooperate tactically with him on one-off decisions.

The reasoning for this, he told me, is simply that the *long-term is long*, and the short-term is short. There's a lot more value to be gained from someone who keeps actively pursuing goals aligned with yours, even when they're far away and you haven't spoken in a long time, than from someone you can persuade or incentivize to do a specific thing you want right now, but who won't be any help in the long run (or might actually oppose your long-run aims.)

This seems like fine reasoning to me, as far as it goes. I think my point of departure is that I estimate different numbers for probabilities and expected values than him. I expect to get a lot of mileage out of relatively transactional or local cooperation (e.g. donors to my organization who don't buy into *all* of my ideals, synagogue members who aren't intellectually rigorous but are good people to cooperate with on charity, mutual aid, or childcare). I expect getting to alignment on principles to be really hard, expensive, and unlikely to work, most of the time, for me.

Now, I think compared to most people in the world, we're *both* pretty far on the "long-term cooperation" side of the spectrum.

It's pretty standard advice in business books about company culture, for instance, to note that the most successful teams are more likely to have shared idealistic visions and to get along with each other as friends outside of work. Purely transactional, working-for-a-paycheck, arrangements don't really inspire excellence. You can trust strangers in competitive market systems that effectively penalize fraud, but large areas of life aren't like that, and you actually have to have pretty broad value-alignment with people to get *any* benefit from cooperating.

I think we'd both agree that it's unwise (and immoral, which is kind of the same thing) to try to benefit in the short term from allying with terrible people. The question is, who counts as terrible? What sorts of lapses in rigorous thinking are just normal human fallibility and which make a person seriously untrustworthy?

I'd be interested to read some discussion about when and how much it makes sense to prioritize strategic vs. tactical alliance.

Cargo Cult and Self-Improvement

Many people want to change themselves. They want to become happier, smarter, more successful, rich, or at least less disoriented. There are incredibly many books telling people how they can achieve this, along with seminars, youtube videos and email instructions. And the large majority of this self-improvement literature is nonsense.

There is of course a wide range of what the authors promise. But the typical "If you follow this author's system, then you will get rich / become world ruler / can do whatever you want" has some features worth noting.

The literature sells well because people are unhappy with their life, their abilities, their habits and *want* to have other lifes, other abilities, other habits, and there is only a limited number of ways to cope with this, the most apparent may be: Do nothing and be unhappy, find ways to attribute responsibility to external circumstances, find new perspectives or other ways to be happy without changing the things that you think make you unhappy - and finally, listen to people who tell you that they have a clue. Similarly, if you are not unhappy but just very ambitious and, in your own eyes, not very successful, you can try to be ok with that tension, or find authors promising you change. In any case, you have made the first step of blaming yourself, but in a way that you think you can change - as opposed to blaming your genes alone, or society. Thus, firstly, self-improvement books have to overemphasize what you can change over what you can't - it is their selling point. So, first advice: Be skeptical of authors who claim that everything is possible. If something sounds too good to be true, it probably is.

Secondly, the advice is rarely evidence-based, and the greater the promises, the less valid evidence. This, of course, should make a reader suspicious, but only if he has some understanding of scientific methods. Because many of the authors know this, there seems to be a tendency towards faked scientifc credibility in parts of the literature, namely those parts that want to seem not only credible, but also serious. These authors do not cite sources, but vaguely refer to studies conducted at some university, like the famous 1953 Harvard study or the 1979 Yale study on the effects of writing down goals on life success (Mike Morrison has done the research - these studies have not been conducted, and authors copied the story from each other.). So, second advice: Be suspicious if the author claims that his advice is backed up by scientific evidence and then he does not really cite studies that you can actually look up.

Thirdly, the authors often seem themselves very successful in what they promise to teach. And this they should be - the natural question otherwise would be: If you're so smart, then why aren't you rich? But the success of authors is not sufficient prove for the system to work. If you buy a get-happy-and-successful book from the bestsellers' shelve in a large bookstore, the authors *are* very likely to be successful in happy and successful. Otherwise, you would probably not have found their book. An author who has made it to be something between a coach and a guru may possibly even attribute this to his own system. But suppose that you would let 1000 people write self-help books, all results are useless, but *randomly* 5 of them are successful because there is demand for five books. Then this would not tell you anything about the quality of their insights -- assuming the opposite leads to survivorship bias. In reality, people with better "systems" may have better chances of becoming successful in life, but so

may people who are just more overconfident, who are charismatic, or who write well. (Nothing of this helps the reader.) What would be necessary is a randomized controlled study of people following a certain system and see how many of them were successful. Thus, third advice: As long as this does not exist, do not put too much weight on the author's own story.

Fourthly, for really ambitious self-optimization claims, unscrupulousness, fraud, conman writing additionally comes to the mix. ("Additionally" because for these authors it is true as well that you more often see the successful writers' books.) It is not always clear whether the more esoteric writers talking of psychic energies and wishes directed to the universe in specific ways are conscious fraudsters or fooled by their own random success. However, there were famous self-improvement authors who were most certainly knowing what they were doing in this respect (Budge Burgess provides a very interesting story about Napoleon Hill) and still had enormous success. The problem with conmen is that they are good at what they do, and they will rarely tell you that the recipe for success is to manipulate readers (though there is the old joke of the book 'how to get rich' that only contains the advice 'write a book like this'). So, fourth advice, be extra skeptical if someone promises extraordinary things -- and ask yourself something like: If that were true then why isn't this guy president, or Apple CEO, or something like that?

Fifth point. In the last years, but maybe already earlier, I have seen books that study certain features of the behavior of very successful people -- tings that successful people do, how artists plan their day, what genius space warlords eat for breakfast etc. (The already-mentioned Napoleon Hill claimed to have done something similar.) This is Cargo Cult par excellence: Imitating something without really knowing whether it has relevance. Imitating a genius probably won't make you one. (As Tim Harford comments on a distraction-free writing tool called the Hemingwrite, "It is probably not true that Facebook is all that stands between you and literary greatness.") Similar to the successful authors, there are no randomized controlled experiments about who becomes a genius if she applies this or that behavior, as far as I know. Even then, one would have to keep in mind that treatment effects may interact: Give n persons with IQ>160 some goal-setting and structure and *maybe* they work wonders. The normal person's output may not change much, as workplaces and institutions are already structured around their abilities and habits. Thus, even if you find fascinating whether Einstein wore socks (something Christian Ankowitsch has written a book about, but it seems more for entertaining than for extraordinary promises), do not expect imitating him to have much of an effect (fifth advice).

Sixth point and advice. Keep in mind that many self-help books have an inbuilt mix of blame allocation and non-falsifiability. If you don't succeed, you didn't follow the advice hard enough, maybe not believe in it enough etc. All this means that you are probably still unhappy but now you can additionally blame yourself for another failure. Moreover, there are tendencies that people want to believe in the promises of such books - in particular, if they paid a lot in terms of money or time - to reduce cognitive dissonance. Be aware of this behavior.

So what to take of this? Are there actually useful books out there, and postitive rules to find them and use them?

Firstly, trust your commonsense. If you read claims that let you say "Now this is something I should have done all my life, why did noone tell me that?", then just do it!

(Classical candidate: Dale Carnegie, "How to win friends and influence people") If you think it is fraud, it probably is (Possibly it is not, but time is finite! But also the more basic self-organization and time-management books are in this category)

Secondly, there are some science-based strategies for changing your life, and there are authors getting them to a readable book format, e.g. Wiseman's book "59 seconds". Usually, you will find that the advice of these authors is relatively practical, while their claims are comparably modest.

Thirdly, there are many self-improvement techniques of which you can at least see that they work for *someone*. For instance, if you want to grow muscles, you can go the gym and actually see people who have grown muscles. There may again be survivor's bias at work, but the next step is to read actual, serious advice here. Another case is mnemonics. You can read books by memory champions who explain what they do, and you can see them perform in contests. This usually is a good sign.

Fourthly, it is a good sign if the advice book has low-hanging fruits advice: Things you can just try and then decice whether it helps. This is the case with the mnemonics books, but also for how-to-meditate books (e.g., "10% Happier" by Dan Harris).

Finally, you probably will not acquire superpowers. Be realistic. The scientific literature on becoming a super performer (you need deliberate practice for 10,000 hours to be in the world top class) is much more demanding than the one on getting a bit happier (even therapies are not 10,000 hours). But you can also easily fulfill more realistic claims. Learning to play 'Lady in Black' on the guitar is much easier than becoming rich and famous. Maybe that is what you really wanted all along?

Given all the advice I have put in this, you may think: If you're so smart, then why aren't you rich? I'll leave it to you how to cope with this paradox.

Corrigibility doesn't always have a good action to take

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

In a previous critique of corrigibility, I brought up the example of a corrigible AI-butler that was in a situation where it was forced to determine the human's values through its actions - it had no other option.

Eliezer pointed out that, in his view of corrigibility, there could be situations where the AI had no corrigible actions it could take - where, in effect, all it could do was say "I cannot act in corrigible way here".

This makes corrigibility immune to my criticism in the previous post, while potentially opening the concept up to other criticisms - it's hard to see how a powerful agent, whose actions affect the future in many ways, including inevitably manipulating the human, can remain corrigible AND still do something. But that's a point for a more thorough anaylsis.

Using expected utility for Good(hart)

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

Goodhart's curse happens when you want to maximise an unknown or uncomputable U, and instead choose a simpler proxy V, and maximise that instead. Then the optimisation pressure on V transforms it into a worse proxy of U, possibly resulting in a very bad result from the U-perspective.

Many suggestions for dealing with this involve finding some other formalism, and avoiding expected utility maximisation entirely.

However, it seems that classical expected utility maximisation can work fine, as long as we properly account for all our uncertainty and all our knowledge.

The setup

Imagine that there are 1000+5 different variables that humans might care about; the default value of these variables is 0, and they are all non-negative.

Of these, 5 are known to be variables that humans actually care about, and would like to maximise as much as possible. Of the remaining 1000, an AI agent knows that humans care about 100 of these, and want their values to be high - but it doesn't know which 100.

The agent has a "budget" of 1000 to invest in any variables it chooses; if it invest X in a given variable v, that variable is set to $\sqrt{X}$. Thus there is a diminishing return for every variable.

Then set

- V={The means of the 5 known variables}.

And we get a classic Goodhart scenario: the agent will invest 200 in each of these variables, setting them to just about 14, and ignore all the other variables.

It knows we care, but not about what

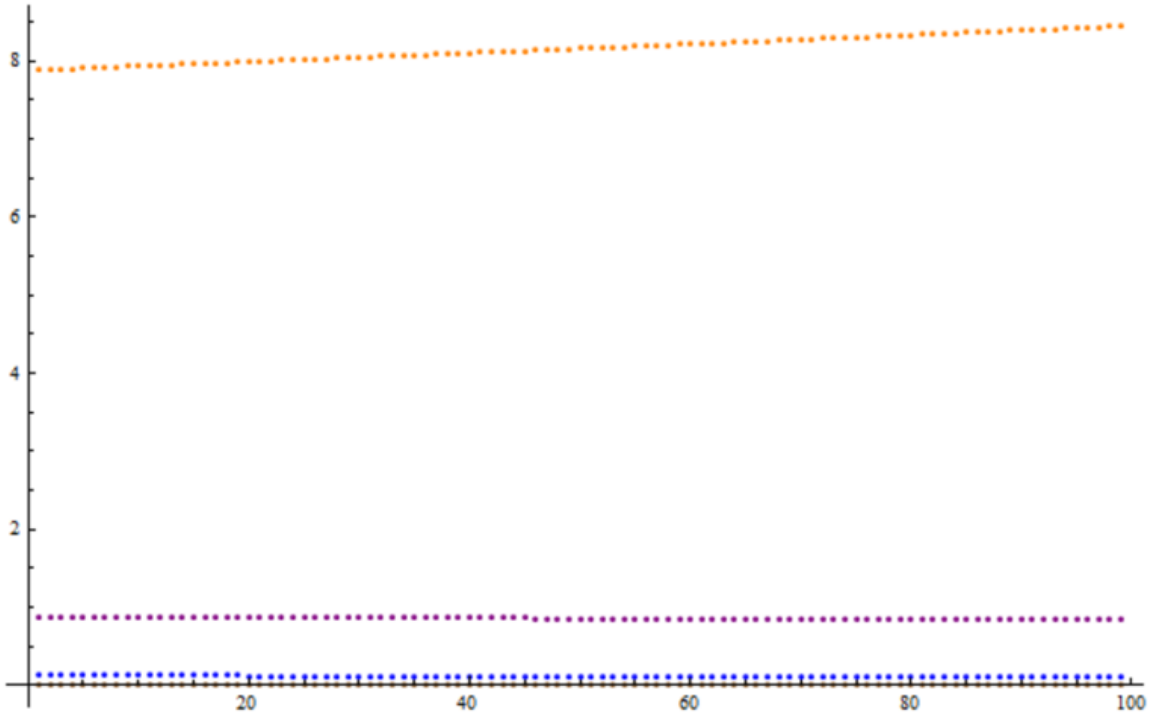
In that situation, we have not, however, incorporated all that we know into the definition of V. We know that there are a 100 other variables humans care about. So we can define V as:

- V={The mean of the 100+5 variables the humans care about}.

Of course, we don't know what the 100 variables are; but we can compute the expectation of V.

In order to make the model more interesting, and introduce more complicated trade-offs, I'll designate some of the 1000 variables as "stiff" variables: these require 40 times more investment to reach the same value. The number of stiff variables varies between 1 and 99 (to ensure that there is always at least one variable the humans care about among the non-stiff variables).

Then we, or an agent, can do classical expected utility maximisation on V:



This graph plots various values against the number of stiff variables. The orange dots designate the values of the 5 known variables; the agent prioritises these, because they are known. At the very bottom we can make out some brown dots; these are the values of the stiff variables, which hover barely above 0: the agent wastes little effort on these.

The purple dots represent the values of the non-stiff variables; the agent does boost them, but because they have only 1/10 of being variables humans care about, they get less priority. The blue dots represent the expected utility of V given all the other values; it moves from 0.13 approximately to 0.12 as the number of stiff variables rises.

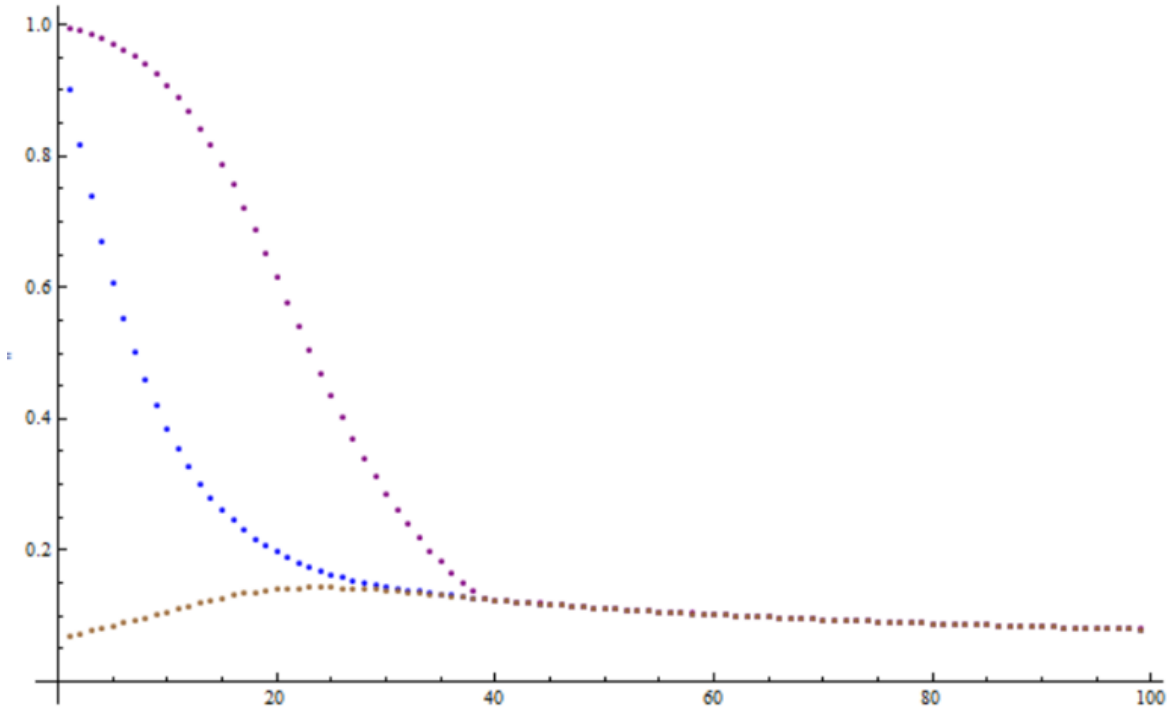
This is better than only maximising the 5 known variables, but still seems sub-optimal: after all, we know that human value is fragile, so we lose a lot by having the stiff variables so low, as they are very likely to contain things we care about.

The utility knows that value is fragile

We know that human value is fragile, but that has not yet been incorporated into the utility. The simplest way would be to define V as:

- V={The minimum value among the 100+5 variables the humans care about}.

Again, V cannot be known by the agent, but its expected value can be calculated. For different number of stiff variables, we get the following behaviour:



The purple dots are the values that the agent sets the 5 known variables, and the un-stiff unknown variables to. Because V is defined as a minimum, and because at least one of the un-stiff unknown variables must be one the humans care about, the agent will set them all to the same value. The brown dots track the value of the stiff variables, while the blue points are the expected value of V.

Initially, when there are few stiff variables, the agent invests their efforts mainly into the other variables, hoping that humans don't care about any of the stiff variables. As the number of stiff variables increases, the probability that humans care about at least one of them also increases. By the time there are 40 stiff variables, it's almost a certainty that one of them is one of the 100 the humans care about; at that point, the agent has to essentially treat V as being the minimum of all 1000+5 variables. The values of all variables - and hence the expected utility - then continues to decline as the number of stiff variables further increases, which makes it more and more expensive to increase all variables.

This behaviour is much more conservative, and much closer to what we'd want the agent to actually be doing in this situation; it does not feel Goodharty at all.

Using expected utility maximisation for Good(hart)

So before writing off expected utility maximisation as vulnerable to Goodhart effects, check if you've incorporated all the information and the uncertainty that you can, into the utility function.

The generality of this approach

It is not a problem, for this argument, if the number of variables humans care about is unknown, or if the tradeoff is more complicated than above. A probability distribution over the number of variables, and a more complicated optimal policy, would resolve these. Nor is

the strict "min" utility formulation needed; a soft-min (or a mix of soft-min and min, depending on the importance of the variables) would also work, and allow the utility-maximiser to take less conservative tradeoffs.

So, does the method generalise? For example, if we wanted to maximise CEV, and wanted to incorporate my criticisms of it, we could add the criticisms as a measure of uncertainty to the CEV. However, it's not clear how to transform my criticisms into a compact utility-function-style form.

More damningly, I'm sure that people could think of more issues with CEV, if we gave them enough time and incentives (and they might do that as part of the CEV process itself). Therefore we'd need some sort of process that scans for likely human objections to CEV and automatically incorporates them into the CEV process.

It's not clear that this would work, but the example above does show that it might function better than we'd think.

Another important challenge it to list the different possible variables that humans might care about; in the example above, we were given the list of a 1000, but what if we didn't have it? Also, those variables could only go one way - up. What if there were a real-valued variable that the agent suspected humans cared about - but didn't know whether we wanted it to be high or low?

We could generate a lot of these variables in a variety of unfolding processes (processes that look back at human minds and use that to estimate what variables matter, and where to look for new ones), but that may be a challenge. Still, something to think about.

Economic policy for artificial intelligence

This is a linkpost for https://voxeu.org/article/economic-policy-artificial-intelligence

Reducing collective rationality to individual optimization in common-payoff games using MCMC

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

A well-known issue in game theory is that of multiple Nash equilibria in coordination games.

For example, consider the following game. Players 1 and 2 both choose either "high" or "low". If both choose "high", they both get 10 utility; if both choose "low", they both get 5 utility; if they choose different actions, then they get no utility. There are 3 Nash equilibria: one where both always play "high", one where both always play "low", and one where each plays "high" with 1/3 probability and "low" with 2/3 probability. While (high, high) is obviously the best Nash equilibrium, unilaterally switching to high when the current equilibrium is (low, low) is probably unwise.

You could imagine TDT arguments based on symmetry between the different players to argue that "high" is the only rational choice, but it is easy to imagine asymmetric variants of this problem where the symmetry arguments are much less clear.

The presence of suboptimal Nash equilibria looks like a serious problem for reconciling individual and collective rationality, even when players have common payoffs. If everyone is going to play "low", then playing "low" is individually rational (at least according to CDT); however, the collective of agents fails to act rationally due to a coordination failure.

What I hoped to do before coming up with the idea in this post is to somehow reduce collective rationality to individual rationality. Can we prescribe individual rationality conditions to each player, such that if each player follows these conditions, then a near-optimal outcome for everyone is attained?

Somewhat surprisingly, the answer is "kind of." It certainly seems like solving coordination problems requires global optimization instead of just local optimization, but under appropriate conditions, local optimization is actually sufficient.

Problem setup

We will consider the setting of normal-form common-payoff games. Let n be the number of players. Let A be some finite set of actions; each player i will select the action A_i. Let $U : A^n \to R$ be the shared utility function. Each player receives the same utility $U(a_1, \dots, a_n)$.

I will now assume that the players play some number of practice rounds before the "real" round. These practice rounds are very similar to those in the fictitious play

algorithm. Ontologically, what are these practice rounds? One interpretation is that they are actual rounds of the game that have been played in the past; another interpretation is that they are a sequence of analogous one-shot rounds played by agents who have more and more compute and are thus able to simulate previous rounds. (The second interpretation might seem strange, but this is essentially how logical inductors achieve generalization bounds in decision theory problems)

Let T be the number of practice rounds. Let A_i^t be the action selected by player i in practice round t. I will assume $A_i = A_i^T$, i.e. the last practice round is the "real" round.

We will assume that, in each round t, some shared random number $R_t \sim \text{Uniform}(\{1, \dots, n\})$ is generated, and that players' policies in round t may depend on this number.

How are players' policies represented? We will assume that each player has only black-box access to each others' actions in previous practice rounds. Let $\pi_i : \text{List}(A^n) \times \{1, \dots, n\} \to \Delta A$ be player i's policy, which selects an action distribution for player i in round t given knowledge of the actions in all previous rounds and of the current shared random number.

The generative model is now obvious: For each t in sequence, sample $R_t \sim \text{Uniform}(\{1, \dots, n\})$ and then sample each $A_i^t \sim \pi_i(A_{1:n}^{1:t-1}, R_t)$ independently, where the notation $A_{1:n}^{1:t-1}$ refers to the list $(A_{1:n}^1, \dots, A_{1:n}^{t-1})$.

Given a set of policies, it is possible to compute, for each t, expected utilities at each time step (which are the expected utilities attained if $T = t$):

$$v_t(\pi_{1:n}) := \mathbb{E}[U(A_{1:n}^t) | \pi_{1:n}]$$

The limit $\lim_{t\to\infty} v_t(\pi_{1:n})$ and corresponding $\liminf$ and $\limsup$ are interesting; they indicate how well a set of policies does in expectation after a large number of practice rounds.

Policies that solve the game

The policies we will consider are ones that usually stick with the same action they played in the last round, but sometimes switch to a different action, and are more

likely to switch the worse the utility in the last round was.

Let u_{min} be a lower bound on U. Let $\alpha > 0$ be some "rationality parameter". Consider the following set of policies:

$$\pi_i^\alpha(A_{1:n}^{1:t-1}, R_t) = \text{Uniform}(A) \text{ if } t = 1$$

$$\pi_i^\alpha(A_{1:n}^{1:t-1}, R_t) = \delta(A_i^{t-1}) \text{ if } t > 1, R_t \neq i$$

$$\pi_i^\alpha(A_{1:n}^{1:t-1}, R_t) = e^{\alpha \cdot (u_{min} - U(A_{1:n}^{t-1}))} \cdot \text{Uniform}(A) + (1 - e^{\alpha \cdot (u_{min} - U(A_{1:n}^{t-1}))}) \cdot \delta(A_i^{t-1}) \text{ otherwise}$$

where $\delta(a)$ is a delta distribution on a, and probability distributions are taken to be vectors.

These policies start out with each player selecting a random action. In each round, a random player is selected (according to R_t); this player switches their action to some uniformly random action with a probability that is lower the higher the utility in the previous round was, and otherwise sticks with the same action. Players other than the randomly selected player always stick with the same action.

An important aspect of these policies is that *each player only does local optimization*. That is, each player shifts their own action depending on the utility in the last round, such that over the course of many iterations they are more likely to take an action that attains a high utility given the other players' actions. Naively, one would expect a local optimization algorithm like this one to get stuck in traps such as the (low, low) Nash equilibrium, but this turns out not to be the case.

We will prove the following theorem:

Theorem 1: For all $\epsilon > 0$ there exists α such that

$$\lim_{t \to \infty} v_t(\pi_{1:n}^\alpha) \geq \max_{a_{1:n} \in A^n} U(a_{1:n}) - \epsilon.$$

i.e. by setting α high enough, these policies achieve an arbitrarily-close-to-optimal expected utility in the limit.

Analyzing the process as a Markov chain

To prove this theorem, we can analyze the multi-round process with these policies as a Markov chain where the state is the action vector $A^t_{1:n}$. The transition function (giving the distribution of $A^{t+1}_{1:n}$ given $A^t_{1:n}$) is:

$$t^\alpha(a_{1:n})(a_{1:i-1}, a_i', a_{i+1:n}) := \frac{e^{\alpha \cdot (u_{min} - U(a_{1:n}))}}{n|A|} \text{ if } a_i' \neq a_i$$

$$t^\alpha(a_{1:n})(a_{1:n}) := 1 - \sum_{a_i' \neq a_i} f(a_{1:n})(a_{1:i-1}, a_i', a_{i+1:n})$$

$$t^\alpha(a_{1:n})(a'_{1:n}) := 0 \text{ if } a_{1:n} \text{ and } a'_{1:n} \text{ differ on at least two elements}$$

Consider the following distribution over states:

$$p^\alpha(a_{1:n}) := \frac{e^{\alpha U(a_{1:n})}}{\sum_{a' \in A^n} e^{\alpha U(a_{1:n})}}$$

i.e. it is a softmax that is more likely to select states with higher utilities.

We will show:

Lemma 1: p is the unique stationary distribution of the Markov chain defined by t.

Proof:

It is well-known (e.g. see this Wikipedia article) that an ergodic Markov chain satisfying detailed balance (a condition defined later) with respect to a probability distribution over states has this distribution of states as the unique stationary distribution.

The Markov chain is aperiodic, since every state has a nonzero probability of transitioning to itself. Additionally, it is irreducible (i.e. it is possible to get from any state to any other state), since any action in the state can change to any other action with nonzero probability while leaving the other actions fixed. Therefore, the Markov chain is ergodic.

This Markov chain satisfies a detailed balance condition relative to the state distribution p. Specifically, for any states $a_{1:n}$ and $a'_{1:n}$:

$$p^\alpha(a_{1:n}) t^\alpha(a_{1:n})(a'_{1:n}) = p^\alpha(a'_{1:n}) t^\alpha(a'_{1:n})(a_{1:n})$$

i.e. if a state is sampled from p then transitioned once, the probability of a transition from $a_{1:n}$ to $a'_{1:n}$ equals the probability of a transition from $a'_{1:n}$ to $a_{1:n}$.

The proof of this detailed balance property proceeds by case analysis on $a_{1:n}$ and $a'_{1:n}$.

1. If $a_{1:n} = a'_{1:n}$, then the condition is trivial.
2. If $a_{1:n}$ and $a'_{1:n}$ differ on more than one element, then both transition probabilities are 0, so the condition is trivial.
3. If they differ on a single element i, then:

$$p^\alpha(a_{1:n})t^\alpha(a_{1:n})(a'_{1:n}) = \frac{e^{\alpha U(a_{1:n})}}{\sum_{b_{1:n}\in A^n} e^{\alpha U(b_{1:n})}} \frac{e^{\alpha\cdot(u_{min}-U(a_{1:n}))}}{n|A|}$$

$$= \frac{e^{\alpha u_{min}}}{n|A|\sum_{b_{1:n}\in A^n} e^{\alpha U(b_{1:n})}}$$

and similarly

$$p^\alpha(a'_{1:n})t^\alpha(a'_{1:n})(a_{1:n}) = \frac{e^{\alpha U(a'_{1:n})}}{\sum_{b_{1:n}\in A^n} e^{\alpha U(b_{1:n})}} \frac{e^{\alpha\cdot(u_{min}-U(a'_{1:n}))}}{n|A|}$$

$$= \frac{e^{\alpha u_{min}}}{n|A|\sum_{b_{1:n}\in A^n} e^{\alpha U(b_{1:n})}}$$

Since the Markov chain is ergodic and satisfies detailed balance with respect to p^α, it follows that p^α is the unique stationary distribution of the Markov chain.

□

Now we will prove Theorem 1.

Theorem 1: For all $\epsilon > 0$ there exists α such that

$\lim_{t\to\infty} v_t(\pi^\alpha_{1:n}) \geq \max_{a_{1:n}\in A^n} U(a_{1:n}) - \epsilon$.

Proof:

Clearly, $\lim_{t\to\infty} v_t(\pi^{\alpha}_{1:n}) = E_{a_{1:n}\sim p^{\alpha}}[U(a_{1:n})]$, since p^{α} is the unique stationary distribution of $A^{t}_{1:n}$ considered as a Markov chain.

Define $u^* := \max_{a_{1:n}\in A^n} U(a_{1:n})$, i.e. the highest achievable utility. Define u' to be the second-highest achievable utility.

At this point, we need only set α to be large enough that the p^{α} assigns almost all of its probability mass to optimal states. The argument that this is possible to do follows.

Fix $\epsilon > 0$. Set $\alpha := \frac{n\log|A| + \log(u^* - u_{min}) - \log\epsilon}{u^* - u'}$.

Let $a^{*}_{1:n}$ be an optimal action profile, and let $a'_{1:n}$ be a non-optimal action profile. We have

$$\frac{p^{\alpha}(a^{*}_{1:n})}{p^{\alpha}(a'_{1:n})} = e^{\alpha(U(a^{*}_{1:n}) - U(a'_{1:n}))} \geq e^{\alpha(u^* - u')} = \frac{|A|^n(u^* - u_{min})}{\epsilon}$$

Since there is at least one optimal state and fewer than $|A|^n$ non-optimal states, the ratio between the probability (under p^{α}) of a state being optimal and the probability of a state being non-optimal is at least $\frac{u^* - u_{min}}{\epsilon}$. Since in general $\frac{1}{1+x} \geq 1 - x$ for positive x, this means that the probability of a state being optimal is at least $1 - \frac{\epsilon}{u^* - u_{min}}$. A non-optimal state has utility at least u_{min}, so the overall expected suboptimality (i.e. difference between u^* and $u(a_{1:n})$) is at most ϵ, as desired.

□

Extension to non-common-payoff games

The policies defined above can be extended to play non-common-payoff games. Specifically, each agent can have a higher likelihood of switching action in each round if *their own* utility is lower. Unfortunately, this results in mutual defection in a prisoner's dilemma; while I won't present a detailed analysis here, the gist of the argument is that (C, C) is much more likely to switch into (C, D) or (D, C), than (C, D) or (D, C) is to switch into (C, C). So a different approach than this one will be needed for achieving Pareto optimal outcomes in non-common-payoff games.

Conclusion

In this post I have shown that, in common-payoff games under a multi-round structure, there is a set of policies that only does local optimization but which, together, yield a globally near-optimal outcome. This is important because getting global optimality from local optimization is required for decision theory (i.e. each agent taking an action based on its own expected utility) to produce Pareto optimal outcomes in common-payoff games.

In a previous post on coordination games, I presented a plan for solving formal decision theory:

1. Solve Cartesian common-payoff decision theory problems where players reason about each other using something like reflective oracles.

2. Extend the solution in 1 to Cartesian multiplayer game theory problems where players have different utility functions and reason about each other using something like reflective oracles.

3. Extend the solution in 2 to a logically uncertain naturalized setting.

This post has done something pretty close to solving 1, with the multi-round structure taking the place of reflective oracles. Unfortunately, step 2 fails since the solution method does not extend to non-common-payoff games. So it appears that, for my plan to work, 1 must be solved in a different way that can be extended to also solve non-common-payoff games.

Alternatively, the problem of non-common-payoff games may be attacked directly, though this seems difficult given that non-common-payoff games are more complex than common-payoff games, and include all the dynamics present in common-payoff games. But perhaps restricted subsets of general games, other than the set of common-payoff games, can be found and solved.

Player of Games

Things I learned from a game theory party: why prisoner's dilemmas are not what they seem, why being smart is often worse than being the dumbest, and the virtues of trolls.

Like all of my posts, this is cross-posted with delay from Putanumonit.

Does game theory apply to real life? It's easy to fall into one of two errors:

1. Thinking that human interactions can be precisely modeled by a 2×2 payout matrix, and being shocked by the vagaries of human psychology.
2. Thinking that since humans aren't rational game theory doesn't apply to them, and being shocked when they follow incentives in predictable patterns after all.

Of course, human behavior is a combination of both elements: the mathematical structure of payoffs, incentives, and equilibria, the idiosyncrasies of culture, mood, and personality. Understanding both sides can be quite a superpower. I know people who made millions and bankrupted competitors by tweaking some small features of a public auction. This stuff is hard, but it works.

I took every available class on the math of game theory as an undergrad. Since then, I've been catching up on the squishy human part of the equation. There's plenty of great research in this area, but the best way to learn about people is to observe them in their habitat. So when my friend Spencer told me that he's organizing a live game theory party, I jumped at the opportunity.

I will break with tradition by not repeating the descriptions of common games that you've heard 100 times before, nor by deriving the rational strategies and equilibria. If you need to catch up on the math you can follow the links to Wikipedia etc.

What game is it anyway?

The first game we played had the following stated rules:

1. Each person starts the game with 5 tokens and plays a single round with 6 different players.
2. Each player can play "blue" or "red", the players do so simultaneously.
3. The payouts are as follows: if both picked "blue" they each receive one token from "the bank", if both picked "red" they have to pay one token to the bank, and if they show different colors, "blue" pays two tokens to "red".

Or in matrix form:

	They play blue	They play red
I play blue	I get 1, they get 1	-2, 2
I play red	2, -2	-1, -1

What game is this? Take a minute to think for yourself.

If you said that this is the prisoner's dilemma, with "blue" and "red" standing for "cooperate" and "defect", congratulations! You paid attention in class and get an A-.

For extra credit, consider that we call both the following games "prisoner's dilemma":

Game 1 – If you *defect* you get $1. If you *cooperate* the other person gets $1,000,000.

	They cooperate	They defect
I cooperate	1,000,000, 1,000,000	1, 1,000,001
I defect	1,000,001, 1	1, 1

Game 2 – If you *defect*, you get $1,000. If you *cooperate*, the other person gets $1,001.

	They cooperate	They defect
I cooperate	1,001, 1,001	0, 2001
I defect	2001, 0	1,000, 1,000

Think what you'd actually do if you played a single round of each game with an unseen stranger. Is the first game even much of a dilemma? If you pride yourself on cooperation, would you really do so in the second game knowing that you can always donate the extra $1,000 to charity if the other person cooperates? How about if you play Eliezer's version of the "true" prisoner's dilemma?

"Prisoner's dilemma" seems to cover many games that summon very different intuitions and behaviors. There's more to understanding a game than just the preference ordering of the cells in the payoff matrix.

But let's get back to the version of the game I actually played at the party. Here is some more information that may or may not be relevant to analyzing it:

1. The reward for having the most tokens (out of 30 party participants) at the end of the game was a small box of gummy bears.
2. The participants were educated young professionals who piled a table high with snacks and drinks that they brought to the meetup.

Does this change your mind about the sort of game being played?

In the sixth and final round of the game, I was matched with a woman that I spent some time talking to before the meetup had started. We had a pleasant chat and discovered that we had much in common. When our round started, I had one token of the original five remaining (can you guess my strategy?) and she had six.

She asked me what my strategy is, saying that she's always trying to match what the other person will do. I said that I always cooperate, and showed her my single remaining token. She retorted that losing tokens isn't *proof* that I cooperate. I answered that no proof is really possible, but my lack of tokens is at the very least strong evidence. Cooperating always nets you fewer tokens than defecting, so having fewer tokens has to be evidence of a cooperator.

We counted to 3. I showed blue. The lady showed red.

Who won *the game*?

Since I did, in fact, study game theory, I constructed a payoff matrix in my head before the game had started. I noted that I have little chance of acquiring the gummy bears, and also little desire to do so. My real matrix was as follows:

	She plays blue	She plays red
I play blue	We signal that we're nice & like each other	I look a bit naive, but she's an asshole
I play red	I feel really guilty screwing a nice person over	We both signal cleverness

The best outcome in the real PD, when I defect and she cooperates, is now my *worst* outcome. The only thing I gain is a nasty reputation, since the gummy bears are in any case beyond my reach. We both prefer the cells where we match our strategies to those we don't, and we prefer that both cooperate rather than both defect. I called this game "cheap virtue signaling", but the actual name for it in the literature is a pure coordination game.

As expected, upon seeing my blue token the woman's face fell and she spent the next minute profusely apologizing and offering excuses for her moral failing. By defecting against me she showed not only that she's not a nice cooperator, but also that she's a bad judge of character since she didn't guess my strategy.

I finished the first game with no tokens and a big smile.

Tits and tats

The next game we played had a similar structure, except that we would play four times in a row with each of two partners.

My first matchup was a guy I knew from a couple of previous meetups. I guessed that:

1. He was a generally nice person.
2. He knew that I'm generally a nice person.
3. He's not familiar with the theory of the prisoner's dilemma iterated over a fixed number of rounds.

I announced that I would play tit-for-tat: I will cooperate in the first round and subsequently do whatever he did in the preceding round. He agreed that it’s the sensible thing to do.

We both showed blue in the first round. And in the second round. And in the third. In the fourth round, he cooperated and I defected. I was up 5 tokens, and my partner seemed more impressed than upset.

Unlike the first game, where winning came down to being a very lucky defector, the iterated game had a bit of strategy involved and I thought that mine was close to optimal. If all went as planned I would win and get to demonstrate my cleverness, along with handing out gummy bears.

On the second go-round, I was matched with a gentleman I have never seen before. I explained my tit-for-tat plans, and he nodded in agreement. The first round started – I showed blue and he showed red.

I explained that in accordance with tit-for-tat I will now have to defect in the second round, but he can still make more money by cooperating. If he keeps defecting, he would be down three tokens, while if he cooperates and then we go blue-blue in the last two rounds he would be at 0. He nodded again and said that the math made sense. The second round started – I showed red, he showed red as well.

I realized that the first prize is now beyond me, and didn’t bother explaining any more math since it was obvious we’re both just going to defect in the last two rounds. In the third round, I showed red and the dude showed blue, then looked dismayed at my dastardly betrayal. In the fourth round, he indignantly defected and I cooperated just to mess with him.

Outside the box

If my second partner had gone along with my not-quite-tit-for-tat, I would have ended up with 10 tokens more than I had started with. Winning 10 required that my partners cooperate 8 out of 8 times without fail and that I would also get away with two defections. But when the winner for the second game was announced, he was up 14 tokens! What possible strategy could be that much more effective than mine?

Again, I advise you to take a minute to see if you can come up with the answer.

The winning strategy was simple – a couple conspired to have the woman give away all her tokens to her boyfriend by repeatedly cooperating while he defected.

When word of this got out, some participants were miffed at the winner for cheating. I was grateful! I came to the meetup to learn about people, and creative cheating is what people do.

Be the smartest or be dumb

Next, we played two collective games that required guessing what the rest of the room would do. In each game, we had to write down a number between 0 and 10,000 on a piece of paper.

In the first game, the winner was the person whose guess was closest to the group’s average. What’s your strategy?

I took a page from the winning couple in the last game and conspired with my wife, Terese. I told her that I’m putting 10,000 on my paper. Assuming that everyone else would average to 5,000, she put down 5,200. She ended up finishing second – the average was 5,720 and the winner had guessed 5,600.

In the second game, the winner was the person who would be closest to *half* the group's average. I put 10,000 as usual, Terese put 400.

The distribution of guesses turned out to be this: three people put down 10,000 (including me), five people put 0, ten people put 1, and the remaining twelve were all over the range between 2 and 5,000. The average was close to 2,000, and a couple of people were closer to 1,000 than Terese was.

Wikipedia has a good analysis of the second game, which involves iterated reasoning about what the other players will do. The reasoning steps go something like this:

1. Zeroth level thinking – thinking about other people is hard, so I'll pick a number at random.
2. First level thinking – if the naive average is 5,000, I'll guess half of that, namely 2,500.
3. Second level thinking – wait, if everyone realizes they should choose 2,500 I should choose 1,250. But everyone will realize that, so 625, then 312… I get it! The game converges to 0! I should guess 0!
4. Third level thinking – but if everyone guesses 0, I should probably guess 1 to be unique. Then if one person guesses high, I win.
5. Fourth level thinking – ???

Here's the interesting part – writing a random number (0th level) is a reasonably good strategy that *sometimes* wins, especially if you adjust your guess based on experience with the group. But putting down 0 or 1 (2nd or 3rd level thinking), which half the people did, is a *terrible strategy that almost never wins*.

If a lot of people guess 1 they will at best split the prize, and it often takes just two players going high to dominate that strategy. In the presence of a single troll who puts 10,000 (which 10% of the people did in my case), 0s and 1s lose to all the guesses in the 2-350 range. People can end up in that range either with 4th+ level thinking, or just 0th level thinking and some luck. Third level thinking, in this case, works if and only if everyone else is on level 2, no higher and no lower. Picking a random low number, on the other hand, is quite robust to distributions of other players' thinking levels and always gives you some chance of victory.

If you can't be the smartest person in the room (fourth level and up, in this case), you want to be as many levels behind the frontier as you can – stupid and unpredictable. Going through 100 steps of reasoning in a game like this is guaranteed to lose if at least one person has gone through 101, or if everyone is at 30.

Being smart is only worth it if you're smarter than everyone else. If you aren't, it's better to be dumb.

This logic applies to many situations where a lot of people are trying to guess at once what everyone else is thinking. The stock market offers a great illustration of this.

Suppose that you're a trader, and you've followed a stock that has traded around $10 for a while. Everyone seems to agree that $10 is the fair price for it. An analyst report is published by a small sell-side bank (i.e., a bank whose job it is to sell you stocks) that praises the stock and gives it a "buy" rating, without disclosing any new insight or information. The stock immediately jumps to $14 and keeps trading there.

What's the fair value of the stock? Should you buy or sell it at $14?

We can imagine a similar chain of reasoning to the averages game, each step taking into account more and more of other people's thinking into account.

1. I'm dumb, the markets are efficient, and if the stock trades at $14 then the fair price is $14.

2. The stock was worth $10 five minutes ago, and the only thing that happened was the analyst report. But sell-side analyst reports are worthless. Obviously, they're going to tell you that the stock they're selling is great! *Everyone* knows that analyst ratings don't mean anything and there's no reason to update – Matt Levine said so. The fair price is $10.
3. Ok, *somebody* thinks that the stock is worth at least $14. Otherwise, they wouldn't keep buying it at $14. So if I think that it's worth $10 and they think $14, I'll be humble and accept that the true answer is somewhere in the middle. The fair price is maybe $12.
4. Wait, there are also the people selling at $14. Why aren't they selling at lower prices? If all the sellers thought that the fair price is $12, they would compete with each other to sell until the price fell. If I try to sell the stock at $13.99 and succeed, it means that nobody else was willing to sell that low, so not only the buyer but also the *sellers* all think that the price is at least $14. Fuck it, I guess $14 is the right price after all.

The higher up the chain of reasoning you go, the closer your answer gets to the position of maximum ignorance. In the stock market, being smart is worthless if you're not the smartest, and being informed is worthless if you're not the first to be informed.

If you want to read more in-depth examples of this sort of thing by someone who knows game theory, hop over to Zvi's blog.

Theorem of the troll

In the game where the goal was to guess half the group's average, three people guessed 10,000 – a number that is certain to lose. I did it to conspire with my wife, and the other two people did it… for the lulz. They didn't even plan to tell anyone, I had to dig through the slips of paper to find their names and ask them.

And what about my second partner in the iterated PD/virtue signaling game, the one who messed up my tits and tats? Was he too stupid to understand the strategy? Was he playing 12-dimensional chess beyond my comprehension? Was he just trolling?

Jacob's Theorem of the Troll – If you play a game with more than N people, at least one of them is a troll who will play the game in the most incorrect way possible.

Or from another angle: if at least N people are playing a game, at least one of them is playing a different game.

What's N? If the stakes are low and the game isn't the most interesting thing going on, N can be as low as 4-5 players. I think that in most cases, an N of 25 is sufficient to expect a troll loosely riffing off Scott's Lizardman's Constant which is equal to 4%.

You think that you have a game figured out, that the rules and the incentives and the strategies are clear for all to see, and yet without fail somebody will be doing the exact opposite of what they're supposed to do. Maybe they're confused. Maybe they are facing pressures and incentives that you're not aware of. Maybe they're much dumber than you are, and maybe much smarter. Maybe they just want to watch the world burn.

The theorem of the troll is true for your pickup soccer game, for an orgy, for your team at work, for geopolitics, for the NBA. With enough participants, a troll is inevitable.

Two examples:

The paper industry is perhaps the most boring industry in the world. The global market for paper products barely changes from year to year, and the market in each region is dominated by a few huge companies making small but stable profit margins. There's little

ongoing innovation in paper products, and a paper mill costs millions to build and runs 24/7 for a century.

If a single company started building new mills and increased supply beyond the current demand, it would crash the market and cause every single paper company to go from making small profits to huge losses. But because there are so few players, it is possible to avoid trolls and the industry keeps chugging along successfully.

On the other hand, the restaurant and bar industry in New York has around 25,000 participants. The average profit margins are a mere 3-5% which provides little cushion against the wild swings in the ebb and flow of business. 60% of NYC restaurants go out of business within three years of opening. And a big reason why it's so hard to run an NYC restaurant at a profit is the staggering number of restaurants who are trolling the industry by operating at a loss.

Why would someone run a business at a loss? Because they always dreamed of doing it and never did the math. Because they suck at restaurateuring but are too arrogant to admit it. Because they're burning through a loan and are going to declare bankruptcy anyway. Because they're trying to make up for negative margins by increasing volume. You can open the best taco joint in the world and someone will troll you by selling cheap crap tacos across the street at a massive loss.

If you're playing a game that's sensitive to other people's behavior, ignore the troll theorem at your peril.

Trolls vs. equilibria

But there's an upside to trolls – they allow us to break out of **inadequate equilibria**.

Eliezer's book describes how civilizations end up in traps where everyone is unhappy, but no one can fix their own or everyone else's situation by changing their own behavior. A classic example is academic publishing. Researchers have to submit to prestigious journals so they can get good jobs, universities have to hire based on publications in prestigious journals so that they get good researchers, and readers have to pay for the journals to read the good research.

Of course, companies like Elsevier rope off the prestigious journals and extract massive rents in money and effort from all the other participants without contributing anything to the advancement of science. Nobody can unilaterally stop using the prestigious journals – neither readers, researchers, or universities. They're all stuck in a bad equilibrium.

Except for people like Andrew Gelman, a prestigious academic who trolls journals with great gusto. He publishes his research early and for free on his website, takes blog comments as seriously as peer review, and occasionally muses about an academic world in which journals no longer exist.

And of course, there's Sci-Hub (check the first link in this article), whose entire existence is nothing but a giant middle finger aimed at paywalled journals. Sci-Hub lets you access almost any journal article for free, and yet I've donated more money to it than I would ever have spent on buying PDFs. There's no payoff matrix that explains why I donate to Sci-Hub and refuse to pay for journals, I'm just trolling the paywalled academic publishing game in hope that it dies.

Trolls mess up your careful strategies and favorite taco joints, but they also topple dictators, bust monopolies, and pirate journal articles. If you're applying game theory to people, you have to account for the trolls because trolling never dies. Long live the trolls!

Logarithms and Total Utilitarianism

Epistemic status: I might be reinventing the wheel here

A common cause for rejection of total utilitarianism is that it implies the so-called Repugnant Conclusion, of which a lot has been written elsewhere. I will argue that while this implication is solid in theory, it does not apply in our current known universe. My view is similar to the one expressed here, but I try to give more details.

The Repugnant Conclusion IRL

The greatest relevance of the RC in practice arises in situations of scarce resources and Malthusian population traps[1]: We compare population A, where there are few people with each one having plentiful resources, and population Z, which has grown from A until the average person lives in near-subsistence conditions.

Let's formalize this a bit: suppose each person requires 1 unit of resources for living, so that the utility of a person living on 1 resources is exactly 0: a completely neutral life. Furthermore, suppose utility is linear w.r.t. resources: doubling resources means doubling utility and 10 resources correspond to 1 utility. If there are 100 resources in the world, population A might contain 10 people with 10 resources each and total utility 10; population Z might contain 99 people with 100/99 resources each and total utility also 10.

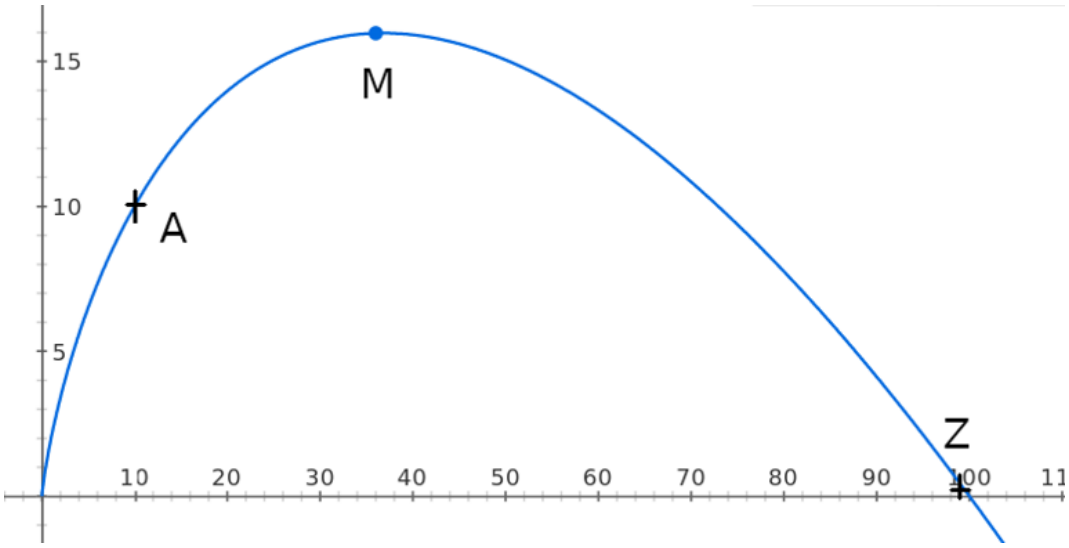
So in this model, we are indifferent between A and Z even as everyone in Z is barely subsisting, and this would be the Repugnant Conclusion[2]. But this conclusion depends crucially on the relationship between resources and utility which we have assumed to be linear. What if our assumption is wrong? What is this relationship in the actual world? Note that this is an empirical question[3].

It is well known that self-reported happiness varies logarithmically with income[4], both between countries and for individuals within each country, so it seems reasonable to assume that the utility-resources relation is logarithmic: exponential increases in resources bring linear increases in utility.

Back to our model, assuming log utility, how do we now compare A and Z? If utility per person is $u_i = \log r_i$ where r_i are the resources available to that person, then total utility is $U = \Sigma \log r_i$. Assuming equality in the population (see the Equality section), if R are total resources and N is population size, each person has resources $r_i = R/N$ and so we have

$$U = \Sigma \log \frac{R}{N} = \Sigma (\log R - \log N) = N(\log R - \log N)$$

We can plot total utility (vertical axis) as a function of *N* (horizontal axis) for R = 100



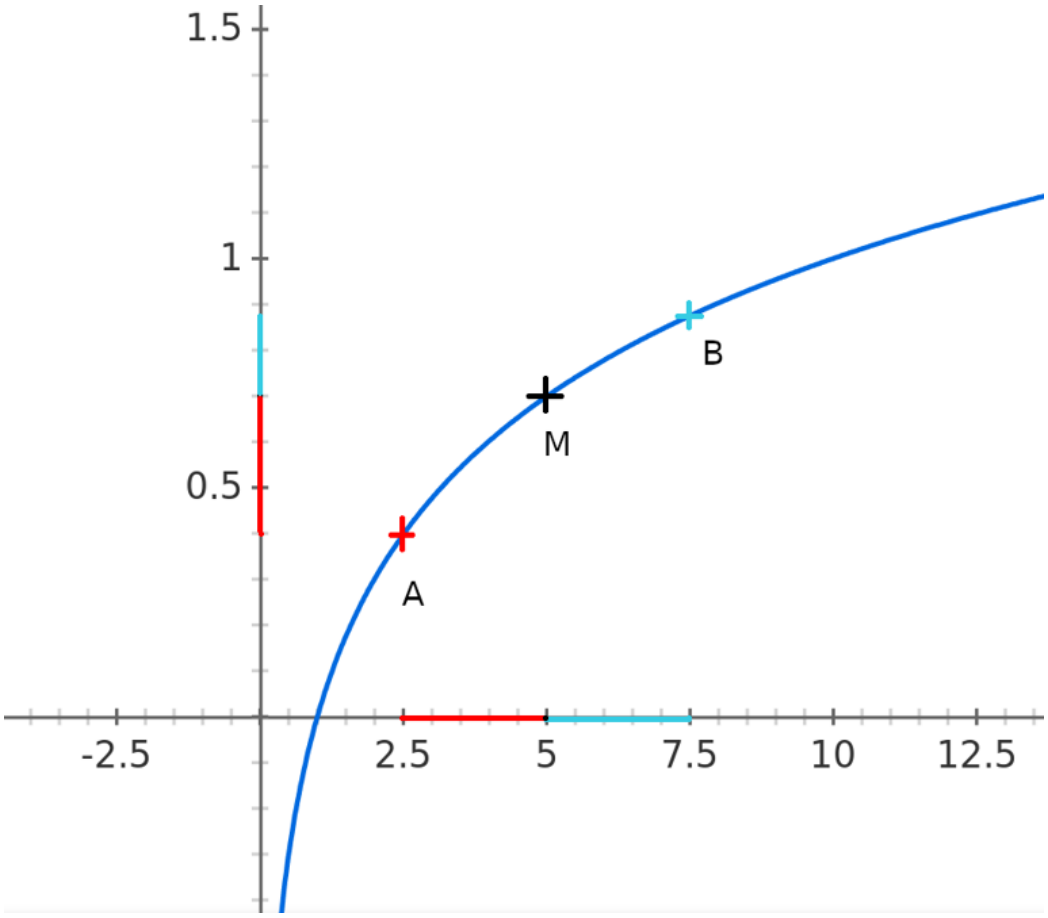
Here we can see two extremes of cero utility: at N = 0 where there are no persons and at N = 100 where each person lives with 1 resources, at subsistence level. In the middle there is a sweet spot, and the maximum M lies at around 37 people[5].

Now we can answer our question! Population A, where N = 10 is better than population Z where N = 99, but M is a superior alternative to both.

So I have shown that there is a population M greater and better than A where everyone is worse off, how is that different from the RC? Well, the difference is that this does not happen for every population, but only for those where average well being is relatively high. Furthermore, the average individual in M is far above subsistence.

Equality

In my model I assumed an equal distribution of resources over the population, mainly to simplify the calculations, but also because under the log relationship and if the population is held constant, total utilitarianism endorses equality. I will try to give an intuition for this and then a formal proof.



This graph represents individual utility (vertical axis) vs individual resources (horizontal axis). If there are two people, A and B, each having 2.5 and 7.5 resources respectively, we can reallocate resources so that both now are at point M, with 5 each. Note that the increase in utility for A is 3, while the decrease for B is a bit less than 2, so total utility increases by more than 1.

This happens no matter where in the graph are A and B due to the properties of the log function. As long as there is a difference in wealth you can increase total utility by redistributing resources equally.

For a formal proof, see [6].

Implications

The main conclusion I get from this is that although total utilitarianism is far from perfect, it might give good results in practice. The Repugnant Conclusion is not dead, however. We can certainly imagine some sentient aliens, AIs or animals whose utility function is such that greater, worse-average-utility populations end up being better. But in this case, should we really call it repugnant? Could our intuition be fine-tuned for thinking about humans, and thus not applicable to those hypothetical beings?

I don't know to what extent have others explored the connection between total utilitarianism and equality, but I was surprised when I realized that the former could imply the latter. Of course, even if total utility is all that matters, it might not be possible to reshuffle it among individuals with complete liberty, which is the case in my model.

Footnotes

1: One might consider other ways of controlling individual utility in a population besides resources (e.g. mind design, torture...) but these seem less relevant to me.

2: Actually, in the original formulation Z is shown to be *better* than A, not just equally good.

3: As long as utility is well defined, that is. Here I will use self-reported happiness as a proxy for utility.

4: See the charts here

5: We can find the exact maximum for any R with a bit of calculus:

$$\frac{dU}{dN} = \log R - \log N - \log e = 0$$

$$N = 10^{\log R - \log e} = R\, e^{-1}$$

A nice property of this is that the ratio R/N that maximizes U is constant for al R(the exact constant e^{-1} obtained here is just due to the arbitrary choice of base 10 for the logarithms)

6: For a population of N individuals the distribution of R resources which maximizes total utility $U = \Sigma \log r_i$ is that where $r_i = R/N$ for all i. The proof goes by induction on N .

This is obvious in the case $N = 1$. For the induction step, we can separate a population of $N + 1$ into two sets of N and 1 individuals respectively so that total utility is $U = \Sigma_{i \in \{1..N\}} \log r_i + \log r_{N+1}$. Suppose we allocate R_N resources to the group of N, and $R - R_N$ to the last person. By hypothesis, each of the N people must receive R_N/N resources to maximize their total utility so $U = N(\log R_N - \log N) + \log (R - R_N)$

Now we have to decide how much should R_N be.

$$\frac{dU}{dR_N} = R_N^{N} \ln 10 - (R - R_N)^{1} \ln 10^{0}$$

Solving for R_N:

$$R_N = R \frac{N}{N+1}$$

Therefore, for each of the N first individuals $r_i = R_N/N = R/(N + 1)$ and for the last one $r_{n+1} = R - R_N = R(1 - \frac{N}{N+1}) = R/(N + 1)$

[Paper] The Global Catastrophic Risks of the Possibility of Finding Alien AI During SETI

[edit: it looks like immediately after publishing the paper, the journal becomes extinct, so the link is no longer working]

My article on the topic has been finally published, 10 years after first draft. I have discussed the problem before on LW. The preprint, free of paywall, is here.

The main difference between the current version and my previous post is that I concluded that such attack is less probable, because if we take into account distribution in the Universe of the naive-our-level-civilizations and civilizations which has powerful AI and are SETI-senders, when the attack become possible, only if most naive civilizations go extinct before the creation of the AI. In that case, succumbing to a SETI-attack may be net positive, as the chance that it is a message from a benevolent alien AI becomes our only way to escape inevitable extinction. Anyway, we should be cautious, if we get any alien message, especially if it will have descriptions of computers and programs to them.

Tidying One's Room

Previously (Compass Rose): Culture, Interpretive Labor, and Tidying One's Room

Epistemic Status: A bit messy

"*She's* tidied up and I can't *find* anything! All my tubes and wires, careful notes!" – Thomas Dolby, She Blinded Me With Science

From Compass Rose:

> Why would tidying my room involve interpretive labor?
>
> It turns out, every item in my room is a sort of crystallized intention, generally past-me. (We've all heard the stories of researchers with messy rooms who somehow knew where everything was, and lost track of everything when someone else committed the violent act of reorganizing the room, thus deindexing it from its owner's mind.) As I decided what to do with an item, I wanted to make sure I didn't lose that information. So, I tried to Aumann with my past self – the true way, the way that filters back into deep models, so that I could pass my past self's ideological turing test. And that's cognitively expensive.
>
> It's generally too aggressive to tidy someone's room without their permission, unless they're in physical danger because of it. But to be unwilling to tidy my own room without getting very clear explicit permission from my past self for every action – or at least checking in – is pathologically nonaggressive.

From my wife, upon seeing the draft up to this point:

> You know, in the time it takes you to write this, you could actually tidy your room.

Proof that the subject of cleaning, and cleaning that which does not belong to you, can escalate quickly in aggressiveness!

There are a few dynamics I'd like to talk about here. I won't (today) be relating them back to Ben's larger questions of how generally to deal with the intentions of the environment, instead choosing a more narrow scope.

I. Intention

Your past self left you an ideological Turing test, of a sort, by leaving items in seemingly random locations.

Good news! I have the cheat sheet.

Close, but wrong: "I'll remember that it's there and I'm too lazy to optimize its location further."

Usually correct: "I'm done with this, I should put it somewhere. This is somewhere."

Don't give your past self too much credit.

Most things are (hopefully) where they are because you put them there on purpose. That's where they 'live'. If they're not in a permanent location, they're probably in an

arbitrary location.

One should think about intention behind the current location of a thing *if and only if the location was clearly chosen intentionally.*

If the location doesn't seem random, this is probably why: "I predicted I'd look here for this item in the future. This is where I seemed to have indexed it."

Ben worried he needed to pass an ITT against his past self before he could alter his past self's wishes.

I think that's backwards. Past you's work is done. The key ITT is against your *future* self!

II. Indexing

Whether or not a location was chosen carefully, it has the great advantage of memory. If you put something somewhere, there's a good chance that's where you will look for it. If you put it there regularly, that chance is better still.

This is why 'tidying' someone's room for them *is an act of aggression.*

If I'm the one who put a thing somewhere, I could figure out where it is by remembering where I put it, or asking where I had it last (which my family called 'The Papa Josh method' as if it wasn't universal, but specific names are still useful, and Papa Josh was apparently kind of an annoying jerk about it). I could also pass the ideological Turing test *of my past self* and figure out where I would have chosen to put it. Since, philosophical objections aside, I am me, my chances are often very good.

If I have a strong indexing of an item to a location, I'll instinctively put it back in the same location, confident I can find it in the future. My ability to automatically look in the right place, and find it now, is good evidence of that. If it was hard to find, I should probably move it. Over time, this improves indexing.

If *someone else* puts the object somewhere, I now have to figure out where *someone else* would place the object. Over time, if they keep doing this, I'll figure out where they put it, but when a *new person* starts cleaning a location, chaos reigns. What is logical to them is not what is logical to you.

An especially nasty trap is when you're not sure if you know where an object is, so you check, it is where you look for it, then put it back *in a different location.* Oh good, you think, I have it, I'll now put it over here. Classic mistake. If an object is in the first place you look, and you need to find it soon, *put it back exactly where you found it!* If an object isn't in the first place you look, *put it in the first place you looked!* You'll look there again.

Otherwise, what you are doing is systematically taking things you can find, and moving them to locations where you might not find them. Whereas if you fail to find them, you won't move them, and they'll stay not found. This is why you can't find the remote – it keeps moving randomly until it finds a place where you can't find it, then stays there until you figure that one out. Repeat.

It took way too many times when the only thing I needed was reliably in the wrong pocket for me to figure out how this works.

III. Illusion

As a child in the days before the internet, I would keep stacks of sports and gaming magazines in my room. In order to quickly locate the one I wanted, I'd spread them out so part of the cover was visible on each copy, allowing a quick visual scan.

Then someone would, against my will, come in and 'clean' the room, stacking them all into one pile with no way to tell which one was which.

So the moment I came back, I'd undo the pile and spread them back out again, since the pile was almost optimizing for lack of legibility.

I'd complain about this all the time, and make my wishes clear, and the stack would reassemble twice a week anyway.

Space, especially visual space, is a resource. Using it draws things to your attention. That's good if you want to find them! It also threatens to distract. It gives the *appearance* of clutter, and threatens to clutter the mind.

IV. Indebted

It is tempting to 'tidy' one's room, to give appearance of tidiness, or to clear necessary space, by accumulating debt. You shove things aside or into closets, rather than putting them in a place that is helpful. Even sorting things into seemingly organized piles is still debt, if you don't know the indexing and won't be able to find them. At some point you'll be paying search costs.

If you are not careful, this debt will accumulate, and interest on it will add up. It is hard to get motivated to pay down such debts, even when returns are good.

It is also tempting to 'tidy' that which does not need to be fixed, or to let this task distract you as a way to procrastinate other tasks.

My solution is simple. Any time you look for something, you give yourself a reasonable amount of time to find it. If after that time you cannot find it, but you are confident it is there to be found, *you stop looking for the item and instead clean the room (at least) until you find the item.* This inevitably finds the item and creates equilibrium – the more you need to clean, the more likely you are to do so. If you can always find everything, then everything is fine.

How to Build a Lumenator

Once upon a time, a friend was sad. Specifically, they had seasonal affective disorder. They tried to fix it by adding lights to their room during the winter.

It didn't work.

They tried adding a LOT of light.

It worked.

They called the giant bundle of lights they assembled a Lumenator. Other people wondered how they, too, might summon a sun into their living room. The task was not exactly *complicated* or *hard*, but it was a little confusing and inconvenient. Instructions were passed around by word of mouth, and individuals cobbled together lumenators in their own homes. Some of them has seasonal affective disorder, and some just liked their rooms to feel line sunshine all the time.

Bit by bit, people's lives grew brighter.

Eventually, someone said "it's silly that there are no instructions on how to do this on the internet and it has to be passed around as weird Cultural Metis." They said this, but then took no actions based on it. Then a second person said that, and they too took no actions based on it. And then finally someone asked me this weekend how to do it and I wrote this blog post.

(Don't be too impressed, because that *second* person who did nothing was *also* me)

A lumenator is 24 lightbulbs, hung in a row from your ceiling. You may want a dimmer switch to control exactly how bright it is. You may want lantern-covers so that the light isn't directly in your eyes (depends on the room in question).

You will need the following material components (roughly $300 on Amazon). [UPDATE: many of these amazon links no longer work, and I think I got some pieces subtly wrong. I recommend checking out this recent UK post and this post to get more details)

- [x1] 24 socket cord to connect the lights
- [x2] 6 pack of light bulbs, 2700K (Warm White)
- [x2] 6 pack of light bulbs, 4000K (Daylight Glow) (Edit: this should actually be 5000k, but I couldn't easily find a 5000k bulb that was clearly the right specs, will update this when I get a minute to do so)
- [x2-x3] Command Hook 14pack (you may screw up the command hook process a few times so probably want extras of these. Also command hooks are great to have around anyway.)
- [x1] [RECOMMENDED] Adjustable dimmer
- [x1] [BONUS] Optional chinese-style lantern covers

You can swap out the command hooks for something more robust (i.e. nails in wall, if you're allowed to do that in the space you live in) but the listed hooks worked for me.

Why 12 brightest lights and 12 softer lights? Honestly, I don't know, this is the wisdom that was passed down from Ben Sancetta who told it to Oliver Habryka who told it to

me. Something about "it's a nice balance that makes the light not too harsh." Shrug.

Putting it up is *pretty* self explanatory once you have the materials. The only hard(ish) part is getting the command hooks positioned so that can hold up the cord. *Before* you've finished putting up all the hooks, the they won't actually be able to support the cord's weight. But, it's easier to position them if you have the cord with you, so you can place them directly near each socket (where they hold the weight a bit more firmly.

So, you might want two people, one to hold up the cord while the other places hooks.

I can try to write up more explicit instructions, for now just wanted to get this up so I could share it with a friend. I do think once you have the materials and have overcome the initial trivial inconveniences you can probably figure it out using your general human intelligence and rationality skills.

[Edit: it turns out there *was* an original article written on arbital.com, which I failed to find because I misspelled "lumenator" as "luminator". The links on what to buy are out of date, but more clearly convey which technical specifications are important]

SSC Meetups Everywhere 2018

Scott over at SlateStarCodex is running another round of "SSC Meetups Everywhere" and since currently a record-high of 80+ meetups are registered for the coming month, I figured we should help people find their nearest meetup, and make them aware of the opportunity. As part of that, we are adding the meetup map to the frontpage for the next week.

If you ever wanted to meet other rationalists/EAs/SSC-readers in person but found that there were no nearby regular meetups, or want to kickstart your local community, this is probably the best opportunity you are going to get this year.

Ontological uncertainty and diversifying our quantum portfolio

The word "ontology" in the title refers to our conception of the basic building block of reality. In quantum mechanics the ontology is the wave function, in general relativity it is spacetime.

This idea in this post assumes the many-worlds interpretation of quantum mechanics is the correct one. In this interpretation there is an infinity of universes which begin as the same. When an event happens it can go one way in some universes (Schrödinger's cat dead) and another way in another (Schrödinger's cat alive). The number of universes stays constant, they just get diversified.

True random number generators are often based on physical phenomena which can be traced to quantum mechanics, as opposed to the more commonly used pseudorandom number generators. We can call such quantum-mechanical random number generators QMRNGs for short. This kind of random generator generates different numbers in different universes while the agent using the generator is uncertain about in which universe he will end up.

Although it is not widely known, the laws of general relativity actually allow for time travel under some special circumstances. Time travel is widely considered impossible since it leads to certain paradoxes, such as the grandfather paradox and is hence deemed logically inconsistent. Those paradoxes are resolved if the time traveler travels not only in time, but also to another universe, as David Deutsch explains. In this setup a time machine is built and the traveler enters it, some time passes, and when he opens the door and exits the machine, he is in the past. Not only is he in the past, he is also in a parallel universe. When he kills "his own" grandfather in that parallel universe he is not really killing his own grandfather, but a parallel-grandfather, a grandfather of parallel-him.

The fundamental laws of physics are not yet known since there is yet no physical theory of everything. Hence, it is not known what are the basic building blocks of reality (we can call them simply "ontology"). In quantum mechanics the ontology it is the wave function, in general relativity the ontology is spacetime. Whatever the true laws of the universe are, the Earth will still be round and single photons will still interfere in the double-slit experiment. Whatever the true laws of physics are, they will probably include parallel universes. It is not so clear if they will allow for travel between those universes. Even if they allow, it is not clear will such travel be practically feasible as a matter of engineering the actual machines which allow for such travel.

There are three basic possibilities with regards to our ontology and inter-universe travel:

1. There are no parallel universes
2. There are parallel universes but we can't travel between them
3. There are parallel universes and we can travel between them
4. There are parallel universes and we can travel between them but such travel is too expensive

Since we are in a state of ontological uncertainty, we can only assign probabilities to each scenario. A utilitarian who assigns a non-zero probability to the possibility number 3 should think about the consequences of using QMRNGs since the use of them causes quantum diversification.

Let's say we have 10 universes which are all identical, they all have you in them, you are tied to the tracks and a trolley is approaching. You have two buttons to press. Button A in all universes has the same effect but you are not sure which the effect is, there is a 90% of it not doing anything and 10% of it stopping the trolley. Button B uses a QMRNG and it is certain to stop the trolley in 1 universe while letting it run you over in 9 universes. The randomness comes from the fact that you don't know in which universe you are going to end up. From a multiverse-wide-perspective it operates deterministically. To a utilitarian the total expected utility from pressing any button is the same. In case A, the expected utility for each universe is 0.1 lives saved, so for total we get 10 * 0.1 = 1 life saved. In case B, the total expected utility is 1 * 1 + 0 * 9 = 1 life saved.

The expected utility is the same, except... if inter-universe travel is possible and you are an expert surgeon which can save your copy's life after it has been run over. In that case you survive in one universe, enter a inter-universe travel machine, travel to a parallel universe, step out of the machine and save the copy's life. You do that one by one, for all 10 universes, saving everyone in the end. Taking the sum of utility of all universes for all times, the situation when a QMRNG is used looks a lot different than when not used. When not used the expected utility is 1 life saved. In the worst case, at one point in the future the utility becomes zero and stays zero. On the other hand, when QMRNG is used, you can recover, so the long-term expected utility of using a QMRNG is actually 10 lives saved. This applies to existential risk if we just substitute "our copy" with "our entire species" and "revival" with "repopulation".

Let's say that in the moment you are pushing the button you don't know if inter-universe travel is possible. There is a non-zero probability p of it being possible. As long as it does not cost you anything, the expected utility of pressing the button B (which we can write as EU(B)) is always higher, since you always save at least one life and there is a probability p you save 9 more lives. The utility of everyone being alive (written U(all), which is equal to 10) is higher than EU(A) which is equal to 1. EU(B) = (1 - p) * EU(A) + p * (U(all)) = (1 - p) + 10p = 1 + 9p. If there is a cost C, we just subtract it from the result. In case C > 9p, it is better to press the button A.

The multiverse naturally has a certain degree of diversification between universes. Events in some universes go one way, in others go another way. It is not clear what is the extent of this diversification. When walking through the city I may be unsure should I go left or right. It could be the case that I go left in 1% of the universes and right in 99% of them, or I go left in 5%, or any other percent, the closer the percent being to 50% the higher the diversification. It could also be the case that in 100% of situations I go right, and only rarely is there a decision I make differently in different universe, with most my decisions being the same in all universes. As we know from chaos theory there are systems which are highly sensitive to initial conditions, such as the weather, and they could introduce diversification. In those cases, initially the changes between some universes are small but they get amplified with time. The effect of using QMRNG could still be negligible if:

1. The chance of cheap technologically feasible inter-universe travel is very small
2. There is a high amount diversification already

We can assume that inter-universe travel would consume some resources and take some time, perhaps there would also be a constraint that universes you travel to need to be *similar enough* to your own. This would limit the speed of our travel through the configuration space (the linear space in which each universe is a point) and also limit the range of such travel. Imagine a situation where there is an astronomically high proportion of universes in which homo sapiens went extinct and only a small proportion in which it didn't. It would be better to increase the proportion of survivors, since the travel to extinct universes in that case would be faster. Also, increasing the number of survivor universes means that survivors will be spread out in configuration space and as such they will be able to revive a larger area of configuration space. Diversifying our quantum portfolio through the usage of quantum mechanical random number generators reduces existential risk.

The implications of QMRNGs are even greater for negative utilitatians. Aiming to reduce suffering, they are already worried that space colonization will produce more suffering, and spreading through the multiverse multiplies their concerns. The implications are great for those concerned with extreme suffering. There is such a thing as a worst universe and a best universe. Increasing diversification could potentially put some universes in a state which would previously not be achieved, so we could create a new even more terrible worst universe. This is counteracted by creating an even more awesome best universe, but may still on net be negative since bad is stronger than good.

The use of quantum-mechanical random number generators increases the diversification of parallel universes. Our ontological uncertainty gives a non-zero probability to the possibility of inter-universe travel. From these two premises, as illustrated by the travelling surgeon thought experiment, we can conclude that using quantum-mechanical random number generators reduces the probability of our species' extinction.

Amy Hoy's How To Master New Skills

This is a linkpost for https://stackingthebricks.com/master-new-skills/

I'm mostly interested in instrumental rationality, and I think Amy's essays are *ridiculously* practical and well-written and not-scary and rational in their approach. I'm hoping this is of help to some people here, it reminds me of the instrumental rationality sequences I was referred to the last time I asked for links and pointers.

Request for input on multiverse-wide superrationality (MSR)

This is a linkpost for http://effective-altruism.com/ea/1rz/request_for_input_on_multiversewide/

> I am currently working on a research project as part of CEA's summer research fellowship. I am building a simple model of so-called "multiverse-wide cooperation via superrationality" (MSR). The model should incorporate the most relevant uncertainties for determining possible gains from trade. To be able to make this model maximally useful, I would like to ask others for their opinions on the idea of MSR. For instance, what are the main reasons you think MSR might be irrelevant or might not work as it is supposed to work? Which questions are unanswered and need to be addressed before being able to assess the merit of the idea? I would be happy about any input in the comments to this post or via mail to johannes@foundational-research.org.

Jobs Inside the API

Cross-posted from Putanumonit.com

I promised that my last post will go up as I am flying over the Pacific Ocean. Then I tweeted that I was looking forward to experiencing a 90-minute Sunday while flying overnight across the International Date Line. The gods did not approve of my hubris, and as the post went up on Sunday morning I woke up at a cheap hotel on the outskirts of Denver, sans my luggage

I have since made it to Singapore and Thailand, but that travel journal will have to wait. This post is about Saturday night at the Denver airport, and about the future of humanity.

Seriousness meter: three beers.

I booked a flight to Singapore through Denver and LA because I like breaking up extra-long flights into two overnight legs. I can save on hotels by sleeping on the plane, and I get to spend a day in a city I've never been in.

Denver welcomed me with perfect weather, transit that runs on time, and wild jackrabbits. My flight to LA was leaving at 7 pm, so after spending the day walking around I sat down at a tap house to try a Colorado craft beer before I left.

As soon as my brew arrived, so did a text from United Airlines: the flight is delayed by an hour, leaving me with an hour and a half to make the connection at LAX. No problem, I thought, and ordered another pint. With the beer came a second text: another delay of an hour.

I finished the beer. I looked at the menu – there was still a stout I wanted to try. I looked at the updated flight arrival time in LAX – 35 minutes to make the connection. Quite out of character, I decided not to tempt fate again and ordered the bill instead of the stout. As soon as I left the bar came a text with the final delay: my flight from Denver will now be arriving at LAX after the connection to Singapore is departing.

I shrugged, went back inside, drank the stout, and headed to the airport.

Everyone has read *the* article about bullshit jobs, especially those of us with bullshit jobs, which according to the article is everyone.

While the article makes a strong case, I think that it is somewhat overstated. The modern economy has too many moving pieces for anyone to track, including the pieces themselves. It takes thousands and thousands of people to make a pencil, and many of them are intermediaries, service providers, etc. The lady doing corporate finance for the insurance shop that allows the transportation company to buy a fleet of trucks to haul graphite from the mine may think that she has a bullshit job. But without her, there's no pencil.

At least, that's what I believe while I'm in the office. I work for a technology service provider to the financial industry. My job is as removed from actual pencils as any, but

I'd like to believe I'm paid an honest wage for an honest contribution.

But whenever I visit an American airport, my faith in the efficiency of labor market allocations wavers.

I remember going through JFK where a single line for security was splitting into two lines. This is something that isn't hard for most humans to figure out. There was a lady standing at that intersection, holding an iPad that was visible to the people in line. The iPad would alternate flashing a left arrow with a right arrow, and the lady holding the iPad would look at it, then point the next person to the left of the right queue. If you moved towards the appropriate queue after seeing the iPad but before the lady told you to do so, she reprimanded you.

There's a model of the modern economy that separates jobs into two kinds: those above and below the API. Algorithms are replacing middle management, and if you don't have a job telling computers what to do, sooner or later your job will consist of doing what computers tell you.

When I came to the front of the line at JFK, I waited for the lady to tell me where to go, smiled, and thanked her. I felt some pity for her, whether it was deserved or not. I didn't want to make her literally below-the-API job any more thankless than it was.

I arrived at Denver Airport and asked the first gentleman in a United Airlines uniform where I could rebook my flight. He sent me to the economy check-in lady, who sent me to the premier access guy, who sent me to the additional services lady, who informed me that her shift has just ended and she is going home. So she called the shift manager, and in he walked, wearing the different-color uniform and big smile of a man who can handle any problem.

The conversation below is very lightly edited from memory.

Dramatis Personae

Boss Agent Manager Fellow, henceforth **BAMF**, a veteran of the aviation industry who has seen it all. Has access to the super special information systems that seamlessly coordinate global travel.

Me, a little sleep deprived and three delicious beers in. Has access to Google.

Scene

BAMF: How can I help you, sir?

Me: I was supposed to fly to Singapore through LA tonight, but the flight to LA is now landing an hour after the connection, so I need to rebook to a different flight.

BAMF: **type** **click** That's not a problem, we have you on a flight leaving LA at 1 pm tomorrow, going to Singapore through Tokyo.

Me: Ah, but I don't need to go to LA, that was just a connection. Can I just get on the flight from Denver to Singapore?

BAMF: **type** **type** **click** **type** I'm sorry, I don't think we have a flight leaving Denver for Singapore tomorrow.

Me: **Checks Google** How about flight 143 at 11:35 am?

BAMF: I'm sorry but… **pause** **type** **click** Oh.

Sir, we can book you on flight 143 tomorrow from Denver to Singapore through Tokyo!

Me: Great!

BAMF: But you may not want to do that. We found a hotel for you in LA, but we don't have one in Denver.

Me: Oh, you have to book me a hotel directly? You can't reimburse me?

BAMF: No, no, you will book it and we will reimburse you either way. Our system just doesn't see any open hotel rooms in Denver, only in LA.

Me: **Checks Google, sees a thousand hotel rooms in Denver** I'll take my chances, I think.

BAMF: Are you sure? If you can't find anything and will stay in the terminal overnight, we have a special reimbursement form for that as well.

Alright, you're all set now, you've been removed from the flight to LA and are booked on UA 143 tomorrow.

Me: Thank you very much! Now would you happen to know where I could get my bag? I assume they haven't loaded it on the delayed flight to LA yet, because that doesn't leave for a few hours.

BAMF: Ah, your bag is in LA, sir. We sent it ahead on an earlier flight. We always send the bag on the first available flight, whether you're on it or not.

Me: Very well then, I'll make do without it for a day. I assume that my bag will continue to Singapore tonight?

BAMF: Oh no, sir. We can't send a bag on a flight if you're not on it.

Me: ??!?! $#%&$?!

BAMF: **Looks at me like I'm an imbecile** It's because that's an *international* flight, you see. When you get to Singapore, you will need to file a baggage claim.

Me: With the United Airlines desk in Singapore?

BAMF: Of course not! You will have to file it with Nippon Airlines, since they're operating the flight from Tokyo. Only the airline that brings you to the final destination can bring the luggage there.

Me: So you're saying that the only way I can be reunited with the bag containing my utmost necessities for a month of travel is to ask a foreign airline who has never seen or handled this bag, and doesn't fly to Singapore from LAX, to somehow transport it to me from LAX to Singapore?

BAMF: It shouldn't take more than two or three days.

Me: Ok, I need to think for a minute. I see there's a person waiting to speak to you, I'll just stand over here for a bit.

BAMF: **utterly confused** Um, sir, there's no other way to get the bag to Singapore. Trust me, I've been doing this for twenty years.

Me: OK. **thinks**

A couple of minutes pass, BAMF is helping a lady with something (or at least pretending to) while I'm standing off to the side, thinking and occasionally tapping Google on my phone. Every few seconds he shoots me an incredulous glance.

Me: So, I see there's a flight from LA to Denver early tomorrow morning.

BAMF: Yes, so?

Me: So it lands in Denver a few hours before my flight to Tokyo.

BAMF: Huh?

Me: So since United Airlines brought me to Denver, I would like to file a baggage claim with United to get my bag from LA to Denver on the flight tomorrow morning, I will pick it up at the airport.

BAMF: **picks jaw off the floor** I… I guess we can do that.

Epilogue

When I arrived at the airport the following morning, I insisted on checking my bag onto the new flight myself instead of trusting United to reroute it. In doing so, I discovered that BAMF booked me on the wrong flight out of Tokyo. I was rebooked by the bag check lady. When I reached the gate to get my boarding pass, it turned out that I was now booked on a different wrong flight, and they had to change it again.

This all happened despite the fact that I told everyone involved the exact number and time of the flight out of Tokyo I needed to be on.

I hope you were as entertained reading this exchange as I was being part of it. Is there also a takeaway here?

That "United Airlines customer service" is an oxymoron we knew already. That it's fun to tell professionals how to do their job after 5 minutes of googling you also knew; it's basically a rationalist rite of passage. You also should know already to pack a change of clothes in your carry-on item when flying anywhere with a connection.

Here's what I wondered as the scene at the airport unfolded: **why do the United Airlines agents still have their jobs?**

Everyone is constantly worrying about the future when more and more jobs are automated, and with good reason. But as far as I was concerned, the jobs of airline agents *have already been automated, and yet they still have their jobs*.

These agents are nothing but a stupid interface layer between me and the flight management system. Whatever else they do, like checking that the face on my head matches the face in the passport, can already be done better by machines. They are

invariably slower than automated systems, more error-prone, and vastly more annoying.

The immediate cause of these agents' employment is the fact that many travelers wouldn't know how to use Google Flights or a similar system for booking flights and tracking their luggage. But it's more than that.

Air travel is frustrating. Flights get delayed, luggage gets lost, passengers get dragged off planes. I suspect that many people not only want a human to interface with the flight booking systems for them, they also want a human to yell at when things go wrong. If you fly a lot you know that a big part of airline agents' job is to smile while being berated by angry passengers. I'm beginning to suspect that it's the main part.

It seems that customers are splitting into two kinds: those who prefer their commercial transactions automated, and those who prefer them humanized. I buy shoes from Zappos and soap from Amazon, but some people want a person to tell them that a shoe or soap matches their hair or whatever. I do my taxes online and never set foot in the bank or the post office, and yet there are always long lines at both. The market keeps providing ever more algorithmic services for me, and ever more human touches for those who want them.

But as the algorithmic services are becoming better and better, it doesn't make sense to have humans doing the same thing but worse. Instead, there's an opportunity for future jobs to pop up in the interface between the robots and the people who don't want to deal with the robots *directly*.

That's what tax preparers are – they use the exact same software that anyone can use at home, but they allow you to talk to a human (and blame a human) instead of learning the software. That's what the United agents do.

When everyone realizes that Zappos has more shoes, and at a lower price, than any shoe store, I can imagine shoe stores being replaced by people sitting at screens. You would talk to these people, they would ask about your day and measure your feet, and then they would order you the shoes you want from Zappos. And if the shoes pinch, you would have someone to yell at while they smile.

You can already hire a personal assistant to interface between you and many algorithms, but each algorithm could have assistants interfacing between it and many customers. These jobs aren't quite above or below the API, they're *part* of the API.

I don't just think that many future jobs might be of this kind, I think that a lot of present jobs are becoming inside-the-API as algorithms do more and more of the actual work. Those of us who prefer to deal directly with the algorithms will find this human API in equal parts frustrating and amusing.

Book Review: AI Safety and Security

I just read Roman Yampolskiy's new 500-pages-multi-author anthology: Artificial Intelligence Safety and Security to get a comprehensive view of the field. Here is my attempt at reviewing the book while explaining the main points of those 28 independent chapters.

General Comments about the Book

The first part, *Concerns of Luminaries*, is made of 11 of the most influential articles related to future AI progress, presented in chronological order. Reading this part felt like navigating through history, chatting with Bill Joy and Ray Kurzweil.

The second part, *Responses of Scholars*, is comprised of 17 chapters written by academics specifically for this book. Each chapter addresses AI Safety and Security from a different perspective, from Robotics/Computer security to Value Alignment/Ethics.

Part I - Concerns of Luminaries

Why the Future Doesn't Need Us (Bill Joy, Wired Magazine, 2000)

The book starts with the influential essay written by Bill Joy at the beginning of the century. The author, co-founder of Sun Microsystems, author of vi and core contributor of BSD Unix, recounts a decisive talk he had with Ray Kurzweil in 1998 that changed his perspective on the future of the technology.

In particular, Joy compares Genetics, Nanotechnology and Robotics (GNR) in the 21st century with Weapons of Mass Destruction (WMD) last century, being mainly concerned with Nanotechnology and Robotics (where robotics encompasses both actual robotics but also AI.

The author's knowledge about nanotechnology comes from Feynman's talk There is plenty of room at the bottom and Drexler's Engines of Creation (1986). Joy believed nanotechnology didn't work, until he discovered that nanoscale molecular electronics was becoming practical. This raised his awareness about *Knowledge-Based* Mass Destruction amplified by *self-replication* in nanotechnologies.

My comments: this article was not much about AI, but gave a general overview of the concerns with future technology. I really enjoyed this essay on GNR by someone who contributed this much to technological progress.

The Deeply Intertwined Promise and Peril of GNR (Ray Kurzweil, 2005)

It's one of the only chapter in *The Singularity is Near* that addresses existential risk.

Responding to Bill Joy's article, Kurzweil presents several ways to build defensive technology to avoid existential risk: to prevent out-of-control self-replicating nanorobots, just build an *immune system* that can also self-replicate.

For Kurzweil, "Intelligence is inherently impossible to control". His proposition to align AI with human values is "to foster those values in our society today and [go] forward. [...] The non-biological intelligence we are creating is and will be embedded in our societies and will reflect our values."

My comments: I was pleasantly surprised to see Kurzweil addressing Yudkwosky's Friendly AI and Bostrom's framework of existential risk in 2005. He appears to know very well the risks associated with an intelligence explosion, but have different opinions on how to deal with them.

The Basic AI Drives (Omohundro, 2008)

This is an essential AI paper describing the core "drives" that any sufficiently advanced intelligence would possess (e.g. self-improvement, rationality or self-protection).

My comments: this paper presents critical examples of instrumental goals, laying the groundwork for current AI Safety research.

The ethics of Artificial Intelligence (Bostrom & Yudkowsky, 2011)

The paper starts by giving some context about AGI, and then presents principles to better think about AGI ethics:

- Moral status should be independent of the substrate of implementation of an intelligence and of its ontogeny.
- When considering duration of experiences, subjective time must be taken into account.

My comments: this chapter answered Kurzweil and cited Omohundro (cf. previous chapters), giving a feeling of consistency to the book.

Friendly AI: the Physics Challenge (Tegmark, 2015)

Max Tegmark proposes a more *physicist-oriented approach* to Friendly AI. The questions I found the most interesting are:

- What if a final goal becomes undefined and the agent undergoes an "ontological crisis"?
- What does it mean to have a "final goal" if there is no clear "end of time" in physics?
- What would be "not boring" to optimize?
    - The fraction of matter that is Human? Conscious?
    - One's ability to predict the future?
    - The computational power of the cosmos?

My comments: some exciting points and examples. Interesting to have Tegmark's physics-oriented perspective.

MDL Intelligence Distillation: Exploring Strategies for Safe Access to Superintelligent Problem-Solving Capabilities (Drexler, 2015)

This paper presents a method of producing a self-replicating AI with a safe goal of *intelligence distillation*, where the key metric is the description length of an AI capable of open-ended recursive improvement.

Additionally, Drexler defines what he calls *Transitional AI Safety*, or reduction-risks methods for AI Safety research, like:

- extending the time available for research
- enabling experimentation with very capable AI
- using smarter-than-human intelligence to solve AI Safety problems

My comments: This was the first technical chapter, and it was much more difficult to follow. It feels like Drexler (author of the Engines of Creation on nanotechnology which was cited in the beginning by Bill Joy) is answering the previous chapters that often cited his work! In the conclusion, the author gives more meta-advices on AI Safety research, like bridging the AI research agenda gap or enriching the conceptual universe, which I found really interesting.

The value learning problem (Soares, 2016)

The paper surveys methods and problems for the value learning problem. One of the key component is an inductive value learning system that would learn to classify *outcomes*, using some labeled value-learning data. Soares then considers how to adapt the algorithm for different issues (e.g. corrigibility or ontology/ambiguity identification). To solve those problems, such an algorithm would need to be able to learn from sparse data and to identify a referent of a label of the training data for any given model of reality.

My comments: This paper was very enjoyable to read. Clear and straight to the point. It surveys multiple important problems and answers them with a simple and concrete method: inductive value learning.

Adversarial examples in the physical world (Alexey Kurakin & Ian Goodfellow & Samy Bengio, 2016)

Essentially, it's possible to send *adversarial examples* to a classifier that will be misclassified, even in a real world setting (e.g. using a camera input, not feeding directly data into the model). Even assuming the difference between an adversarial example and a training example is much smaller (in magnitude) than a certain noise, the adversarial example can still be misclassified while the noise correctly classified.

My comments: Straight to the point paper with a bunch of concrete experiments, pictures and results. Made me consider the Security concerns in a more pragmatic way (e.g. adding a noise on a stop sign).

How might AI come about: different approaches and their implications for life in the universe (David Brin, 2016)

Multiple paths of AI development are discussed (e.g. in "robotic-embodied childhood", child-like robots are fostered into human homes and raised like children). Other

concepts such as Sapience, the control problem or ethics are considered.

My comments: The essay explains superficially multiple concepts. The author seems to be trying to give an intuition for a "Skynet-scenario" to a general audience. This chapter felt weak and fuzzy compared to the rest.

The MADCOM Future:how AI will enhance computational propaganda, reprogram human culture, and threaten democracy... and what can be done about it (Matt Chessen, 2017)

> "Ten years from now, you won't be able to tell whether you're interacting with a human online or not. In the future, most online speech and content will be machines talking to machines."

Machine-driven communication (or MADCOM) will revolutionize online interaction. This could lead to global information warfare where humans can't compete alone with computational propaganda. Multiple US Policy recommendations are given to limit the bad consequences of MADCOM and help implement countermeasures.

My comments: I found the thorough analysis of the possible impacts of MADCOM insightful. This made me update my beliefs about the importance of computational propaganda this century. However, I found that the chapter focused too much on the short-term. For instance, in the definition of AGI (in the glossary), it says: "AGI is still science-fiction"!

Strategic implications of openness in AI development (Bostrom, 2017)

What would be the short-, medium- and long-term impacts of an open AI development?

In the short- and medium-term, AI labs would still conduct original research to build skills, keep up with the state of the art and have a monopoly on their research for a few months (while competitors are catching up). So openness would result in accelerating AI progress.

In the long-term, the final stages for building AGI will likely be much more competitive. To avoid a tight race where AI Safety is dismissed, a singleton scenario is preferable.

Another possibility to take into account is *hardware overhang*. If algorithmic breakthrough is what leads to an intelligence explosion, then openness would favor small groups (that don't have access to advanced hardware). On the other hand, if hardware is the decisive factor, then openness would favor "elite", or already established labs.

Openness in AI development would clearly help AI Safety research, for it's difficult to work on making AI systems safer if those systems are kept secret. AI Safety could

benefit more from openness than AI Capability, because of a need of external perspectives. However, altruistic outsiders are already likely to openly contribute to AI Safety.

My comments: I found this paper very dense. It made me deeply consider the strategic intricacies of openness. I think this chapter is fundamental for this book and I was happy to have it at the end of this part.

Part II - Responses of Scholars

Using human history, psychology, and biology to make AI safe for humans (Gus Bekdash)

Two intelligences are differentiated: Human Intelligence (HI) and Gene Intelligence (GI). GI is essentially the intelligence of all human genomes (or the intelligence of evolution). GI provides control for human reproduction (and therefore controls HI). If all intelligences prove to be inherently dangerous, then we could use the same human principles that made humanity safe for AI (22 principles are considered).

My comments: This chapter appeared to me both concrete and ambitious because of the precise description of those 22 proposed principles.

AI Safety: a first-person perspective (Edward Frenkel)

In this chapter, the mathematician Edward Frenkel tells his story. In particular, he describes the defense mechanism after a traumatism called dissociation, and how it affected his research life.

His point is the following: at the end humans will get to program the AI, so we should make sure every AI Safety personnel is sane to avoid a flight 9525 scenario, where a depressed pilot caused the death of hundreds of passengers.

My comments: His story touched me, but I felt it did not directly address AI Safety.

Strategies for an unfriendly oracle AI with reset button (Olle Häggström)

Is it safe to have an oracle AI answering only yes/no question and being reset every time? Can the AI send a lethal message?

The AI has several methods to (still) try to send a specific message:

- **The Naïve frequency-based approach** consists in answering yes to all question with a certain probability p where the binary expansion of the probability is the message. However, the gigantic number of questions that need to be asked before the frequency of yes answers converges towards the probability p makes the whole approach impractical. Furthermore, humans could ask very simple questions to make sure it is not trying to send a message by answering randomly.
- **The multiple types of questions approach**, where the AI can classify questions and will always answer the same thing for the same category of question. It partitions the question set into (Q_1, ... , Q_n) so that humans can retrieve the message if they have the partition and at least one answer for each

set. This method is much faster and could be feasible if the question/answer protocol is automated.

My comments: Very clear chapter. The math to prove the bounds were easy to follow. I am still unsure about the possibility that the humans would already know the partition Q_1, ... , Q_n, or try to decipher a lethal message. The methods seem to only work if the humans know what method the AI is using and want to know the content of the message.

Goal Changes in Intelligent Agents (Seth Herd, Stephen J. Read, Randall O'Reilly, and David J. Jilk)

There is a trade-off between spending time trying to solve motivation drift, representation drift, and reducing the risks of hacking.

My comments: Short and clear chapter. It was a bit redundant considering the chapter "The basic AI drives" and "The value learning problem".

Limits to Verification and Validation of Agentic Behavior (David J. Jilk)

This chapter addresses verifiability using a framework for deontology. The main result is that verifying if an agent will always exhibit a good behaviour given a deontology is not computable. This result is then applied to AI Safety strategy.

My comments: I found the math symbols unnecessary. More generally, the chapter was too technical for a non-expert audience.

Adversarial Machine Learning (Phillip Kuznetsov, Riley Edmunds, Ted Xiao, Humza Iqbal, Raul Puri, Noah Golmant, and Shannon Shih)

After a general introduction on what adversarial examples are, the chapter presents a taxonomy of different attacks depending on the level of information about the model being attacked.

One of the key claims is that AI security against adversarial examples should not be neglected, because given enough time and efforts an attacker can build a pseudo-model and cause harm.

My comments: This chapter was very well presented and clear. It resonated with one of the first chapters "Adversarial examples in the physical world".

Value alignment via tractable preference distance (Andrea Loreggia, Nicholas Mattei, Francesca Rossi, and K. Brent Venable)

A framework to compute distances and represent conditional preferences (e.g. "preferring red cars if the car is a convertible") is proposed. The value alignment procedure relies on a distance between subjective preferences and ethical principles.

My comments: The chapter provided very little explanation of how the procedure for value alignment worked intuitively, and I found it difficult to follow without any background on CP-nets.

A Rationally Addicted Artificial Superintelligence (James D. Miller)

In economy, beneficial addiction increases your consumption capital as you consume. For instance, a superintelligence could get utility from learning mathematics, and the more it learns mathematics the more it can learn it "quickly" by self-modification (the self-modification uses mathematics).

The author claims that a superintelligence will have a drive to pursue such beneficial addictions. Therefore, AI Safety researchers should try to find beneficial addictions that promote human welfare.

My comments: It's the first article of the book that is heavily influenced by economics. I found it clear and it convinced me of the importance of beneficial addictions.

On the Security of Robotic Applications Using ROS (David Portugal, Miguel A. Santos, Samuel Pereira, and Micael S. Couceiro)

This chapter addresses security concerns regarding a robotic framework: the **R**obotic **O**perating **S**ystem (ROS). This framework is only used in research, does not have security by default and has multiple security holes (e.g. communication between nodes use clear text in TCP/IP and UDP/IP).

Additional security features are presented:

- a native Transport Layer Security (TLS) for all socket transport
- using Advanced Encryption Standard (AES) for encryption
- integrating the IP extension to security (IPSec)
- having a security architecture at the application level

My comments: This chapter made me realize how important securing robotic applications is. No matter how safe we make our AI systems, if someone is able to hack a connected device or a robot, then everything is lost.

Social Choice and the Value Alignment Problem (Mahendra Prasad)

Social choice theory studies how we can aggregate information about agents into a group decision. Although most **normative questions** (i.e. how the world *should* be) about AI ethics will be provided by social choice, some will be built into the AI by its designers. Some questions that must be answered are:

- **Standing**: who or what has its values influence the AI's behaviours and values?
- **Measurement**: how are ethical values measured? In what context? What about discounted utilities and individuals that cannot represent themselves (e.g. posthumans or dead people)?

Solutions to those questions include:

- **A four-stage procedural legitimacy** that aims at building a constitution of the Superintelligent AI.
- **A risk aversion principle**: we don't care if the voting system fails on some low-stage execution but we want it to be extremely robust on high-stage decision.
- **Nondeterministic voting:** to prevent voters from strategically misrepresenting their preferences.

My comments: I feel like this chapter was adapted from an AI ethics article for a textbook on AI Safety and Security, but did not particularly address AI Safety issues. The story about Condorcet was unnecessary.

Disjunctive Scenarios of Catastrophic AI Risk (Kaj Sotala)

The ASI-PATH model proposes to estimate the risk of an AI catastrophe from the probability of the conjunction of multiple events. However, this model does not take into account the fact that Major Strategic Advantages (MSA) are sufficient for **catastrophic risks** (i.e. human well-being damage on a large scale with more than 10 million deaths) and that any **global turbulence** can lead to **existential risk**.

After exhaustively listing Decisive Strategic Advantage (DSA) and MSA enablers, Sotala proposes a *disjunctive* diagram that could lead to an AI catastrophe.

My comments: Being already familiar with the ASI-PATH model, I was happy to see another diagram that helped me think about AI risk estimates and an exhaustive list of DSA/MSA enablers.

Offensive Realism and the Insecure Structure of the International System: Artificial Intelligence and Global Hegemony (Maurizio Tinnirello)

How can International Relations be improved to better manage AI Risk, given the theory of Offensive Realism (i.e. the lack of a world government leads states to joust for power)? This chapter focused on militarized AI, AI Security and political instability.

My comments: I found that the chapter could be applied to any highly destructive weapons and was not specific to AI.

Superintelligence and the Future of Governance: On Prioritizing the Control Problem at the End of History (Phil Torres)

The key claim is that humanity needs to build an *enlightened despot,* or *friendly supersingleton* to prevent state and non-state actors to blow up the world with dual-use emerging technologies

My comments: Very clear diagrams. This made me update my beliefs about an enlightened despot scenario.

Military AI as a Convergent Goal of Self-Improving AI (Alexey Turchin and David Denkenberger)

There is a drive to build Military Tool AIs because an AGI will face enemies, such as:

- Teams working on AGI
- Nation states or alien civilizations
- Any AGI that already exists or might exist

Examples of Military AI (such as the Terminator) have become taboo, and that is a pity because it is (still) a likely outcome.

My comments: Made me reconsider the importance of military AI. Making progress on Alignment is not enough.

A Value-Sensitive Design Approach to Intelligent Agents (Steven Umbrello and Angelo F. De Bellis)

Value-**S**ensitive **D**esign (VSD) is a flexible and capable design for building Aligned agents. It's self-reflexive, fallible and continually improving (it seeks to predict emerging values/issues and influence the design of technologies early in the process). The main focus is on stakeholders' values.

My comments: I found this chapter too abstract. I would have preferred more concrete applications or a more precise formalization.

Consequentialism, Deontology, and Artificial Intelligence Safety (Mark Walker)

The claim in this chapter is that AI Safety research should prefer consequentialism over deontology. The author then does an introduction on normative ethics: Utilitarianism values happiness (**maximizing principle)** where Kantianism values the moral good will (never **substract**, i.e. never treat people solely as means**)**.

My comments: This chapter did not address AI Safety and superficially covered normative ethics.

Smart Machines ARE a Threat to Humanity (Kevin Warwick)

Kevin Warwick criticizes two points made be Penrose in Shadows of the Mind (1994):

- the human brain is a complex organ, but there is no "great deal of *randomness*" inside
- AI will be able to "understand", but in a different way

My comments: I feel like this chapter addresses problems that were being discussed more than 20 years ago, that are already solved or not relevant anymore (e.g. Asimov's three laws).

Conclusion

The first part, presenting authoritative articles from top AI researchers, made me realize how the most influential ideas of this century have built upon each other. I was pleased to see previous chapters being cited over and over by the next ones. Most of the content of this part is available online, but I enjoyed the selection and organization of those chapters.

The second part, made of 17 chapters written by academics specifically for this book, discussed recent research on very different topics (adversarial machine learning, social choice, nanotechnology, ethics, alignment, robotics, AI governance, AI strategy, psychology, AI policy), and gave me a general overview of the field of AI Safety and Security.

Probabilistic Tiling (Preliminary Attempt)

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

We know a decent amount about how to get tiling in proof-based environments where the objective is for the AI to achieve a goal, or to write an AI to achieve a goal, or to write an AI that writes an AI that achieves a goal, and so on, as long as the chain of deferral is finite. The probabilistic setting is far less explored. This post will outline some necessary conditions to achieve tiling of an expected utility maximizer in fully general environments. Some of these conditions are fulfilled by a logical inductor in the limit. It's entirely possible that there are serious problems with this proof, so I'm hoping that it gets a good kicking.

DEFINITIONS:

π_x is an inputless computation that outputs a bitstring, which will be interpreted as a 2-tuple of an action, and another computation, an element of $A \times \Pi$.

π_x^n is the n'th computation in the infinite sequence of computations defined by starting at π_x and taking the computation that each computation outputs. When n=1, the superscript will be omitted.

a_x^n is the action selected by π_x^n.

E_n is some probability distribution or something that can be used to evaluate complicated computations. Like the n'th stage of a logical inductor, although we aren't necessarily assuming that it has to be a logical inductor.

underlining something refers to replacing the symbol with the thing being described. If the variable a_π is used as a way of abstractly representing some action, then $a_{\underline{\pi}}$ refers to the actual action chosen by the policy.

$U(a_{st}^{1:n-1}, x, a_{\underline{\pi}}^{1:\infty})$ is the utility function, which is some large computation of type $A^\omega \to [0,1]$. So this would be the utility function output when it is fed the past actions generated by starting at π_{start}, the current action, and the infinite sequence of future actions is given by whatever sequence of actions is produced by the self-modification

chain starting with π . Note that this only explicitly writes the starting code, and the code that might be modified to, not the past or future action sequence! This is important for the agent to be able to reason about this computation, despite it taking an infinite input.

Given some sequence of probability distributions, and some sequence of sets of policies, let π_{*n} be the computation

$$\text{argmax}_{x,\pi \in A \times \Pi_n} E_n(U(a_{*1:n-1}, x, a_{\pi}^{1:\infty}) | a_{*n} = x)$$

This is basically maximizing expected utility, conditional on the agent itself taking some action, with the past sequence of actions given by an expected utility maximizer on all previous steps.

Actually, in the proof, we will be assuming a 2-step process where some essential things are proved, and then argmax is done. This is done so that the future probability distribution respects some important information. The three important bits of computation are that the action of the previous computation must be computed/observed, that all the policies in Π_{n-1} must be emulated to know what policy and action they produce, with all of the produced policies added to Π_n, and the proof in the next section must be done to ensure that the conditional probabilities work out appropriately.

Now, time for the assumptions.

The first assumption is that abstract expected utility equals concrete expected utility.

$$E_n(U(a_{*1:n-1}, a_{*n}^{1:\infty})) = E_n(U(a_{*1:n-1}, \underline{\underline{a_{*n}}}, a_{*n}^{2:\infty}) | a_{*n} = \underline{\underline{a_{*n}}})$$

This essentially says that if you're going to actually take action 4, then the expected utility of "I take an action" had better be the same as the expected utility of action 4. This is probably fulfilled by a logical inductor, as it would be possible to buy a conditional share/regular share in the lower-priced utility computation, and sell a conditional share/regular share in the higher-priced utility computation, and when the action is actually taken, the provable equivalence between the action actually taken and the action in the abstract means that the prices converge to each other, and the theorems in the logical induction probably enable profiting off the predictable future equality.

One important note is that CDT spectacularly fails this property. Namely, consider a game of matching pennies against a powerful predictor. Since the environment takes actions as input, it's possible to recompute what would have happened if a different action is plugged in. The CDT agent that keeps losing is going to learn to randomize between actions since it keeps seeing that the action it didn't take would have done better. So it eventually gets to a state where it predicts the reward from "pick heads" and "pick tails" is 0.5 (because there's a 50% chance it doesn't pick heads/tails), but it predicts the reward from "I take an action" is 0, violating this assumption. EDT,

however, will predict that picking heads, picking tails, and picking some action all have utility 0, thus fufilling this condition.

The second assumption is that the environment is fair. Namely:

$$\forall \pi : E_n(U(a_{*1:n-1}, a_\pi, a_\pi^{2:\infty})) = E_n(U(a_{*1:n-1}, a_\pi, a_\pi^{2:\infty}) | a_{*n} = a_\pi)$$

This says something like "after finding out that you take a particular action and policy, that should be the same expected utility as some other policy π picking the same action and successor policy", or, more snappily, "finding out that you take a particular action shouldn't be bad news for the expected utility of a particular sequence of actions". The matching pennies predictor that is just penalizing the sequence of actions selected by an argmaxer is unfair in this sense for the argmaxer. This condition (or something like it) seems necessary, in the sense that you can come up with examples where a failure of this condition incentivizes modifying yourself to a different way of selecting actions.

The third assumption (which has already been stated), is that strategy stealing works. More specifically, if $\pi \in \Pi_{n-1} \wedge \pi \neq \pi_{*n-1}$, then $\pi^2 \in \Pi_n$, and a^1 is computed at time n.

This is another necessary condition. Consider a fair environment that rewards you when you do the hamster dance. You are pretty sure you remember the steps, but aren't completely sure. There's another policy that will make you correctly perform the motions. The expected utility of "move your arm left" is different than the expected utility of "hamster dance policy does its move". However, if you were able to compute what the hamster dance policy would output as its next move, you could just copy the move, without giving up your argmaxiness property. Therefore, if you don't know what move a policy outputs, there may be an incentive to modify to it, even if the environment is fair.

The fourth assumption is conditional future-trust of the following form:

$$E_{n-1}(E_n(U)|\phi) = E_{n-1}(U|\phi)$$

This is emphatically false in general, but there's a special condition that makes it viable, namely that the distribution at time n is guaranteed to assign probability 1 to ϕ iff ϕ . My epistemic state about this is "this seems extremely plausible, but I don't know for sure if logical inductors attain this property in the limit"

PROOF:

We will reason from the perspective of the expected utility maximizer in the proof phase shortly before turn n-1, and show that the future distribution that it trusts endorses expected utility maximization over all other policies, and use conditional self-trust to move this conclusion back to the current maximizer to inform its choice of which policy to pick.

To begin (we aren't reasoning from the perspective of the agent yet), assume that there is some pair $(x, \pi_?)$ that achieves a strictly higher expected utility (according to E_{n-1}) than all other pairs where the second element is π_{*n}.

In particular, there is some pair $(x, \pi_?)$ that achieves a higher expected utility than (x, π_{*n}) .

Now, reason from the perspective of the agent, and take as a starting assumption that $a_{*n-1} = x$.

To begin with, by the definition of π_{*n}, and because $(\pi_?^2, a_?^1)$ is in $\Pi_n \times A$ because of the strategy-stealing assumption, then for whatever action and policy that the argmax policy picks,

$$E_n(U(a_{*1:n-1}, a_{*n}, a_{\pi}^{1:\infty})|a_{*n} = a_{*n}) \geq E_n(U(a_{*1:n-1}, a_?, a_?^{2:\infty})|a_{*n} = a_?)$$

holds, because of the definition of argmax. (Very important note: this inequality isn't necessarily true if we have exploration steps! Exploration steps prevent this tiling argument from going through. However, we require well-defined conditionals for the rest of the reasoning, which is attained by exploration steps, so there is a weird tension here.)

Now, by the argument for the equality of abstract and concrete utility, if there's some policy in particular that is getting a high conditional expected utility, the conditional expected utility of "I self-modify to whatever I'm going to self-modify to" should be the same as the expected utility of "I self-modify to this particular policy that looks really good".

$$E_n(U(a_{*1:n-1}, a_{*n}, a_{\pi}^{1:\infty})|a_{*n} = a_{*n}) = E_n(U(a_{*1:n-1}, a_{*n}, a_{*n}^{2:\infty})|a_{*n} = a_{*n})$$

Now, by assumption 1, about the equality of abstract and concrete utility,

$$E_n(U(a_{*1:n-1}, a_{*n}, a_{*n}^{2:\infty})|a_{*n} = a_{*n}) = E_n(U(a_{*1:n-1}, a_{*n}^{1:\infty}))$$

And because $\pi_{*n-1} = x$ by assumption, and the future probability distribution will have seen that and assigned it probability 1, (so conditioning on it doesn't change anything)

$$E_n(U(a_{*1:n-1}, a_{*n}^{1:\infty})) = E_n(U(a_{*1:n-2}, x, a_{*n}^{1:\infty}))$$

Now, we'll begin working with the other side of the inequality. By assumption 2, that the environment is fair,

$$E_n(U(a_{*1:n-1}, a_?, a_?^{2:\infty})|a_{*n} = a_?) = E_n(U(a_{*1:n-1}, a_?, a_?^{2:\infty}))$$

Now, since the agent at time n has computed and proved that $a_? = a_?$

$$E_n(U(a_{*1:n-1}, a_?, a_?^{2:\infty})) = E_n(U(a_{*1:n-1}, a_?^{1:\infty}))$$

And, by the previous argument about how the action taken is x and the future-agent knows it,

$$E_n(U(a_{*1:n-1}, a_?^{1:\infty})) = E_n(U(a_{*1:n-2}, x, a_?^{1:\infty}))$$

Now, because, the agent has proved that

$$a_{*n-1} = x \to E_n(U(a_{*1:n-2}, x, a_{*n}^{1:\infty})) \geq E_n(U(a_{*1:n-2}, x, a_?^{1:\infty}))$$

if the agent responds to proofs of $\phi \to X \geq c$ by having the expected value of the LUV X conditional on ϕ being $\geq c$, then in the argmax estimating phase,

$$E_{n-1}(E_n(U(a_{*1:n-2}, x, a_{*n}^{1:\infty})) - E_n(U(a_{*1:n-2}, x, a_?^{1:\infty}))|a_{*n-1} = x) \geq 0$$

And then, by linearity of expectation,

$$E_{n-1}(E_n(U(a_{*1:n-2}, x, a_{*n}^{1:\infty}))|a_{*n-1} = x) \geq E_{n-1}(E_n(U(a_{*1:n-2}, x, a_?^{1:\infty}))|a_{*n-1} = x)$$

and then, by assumption 4 about conditional self-trust, we get that

$$E_{n-1}(U(a_{*1:n-2}, x, a_{*n}^{1:\infty})|a_{*n-1} = x) \geq E_{n-1}(U(a_{*1:n-2}, x, a_?^{1:\infty})|a_{*n-1} = x)$$

And this contradicts the starting assumption. □

Therefore, if all the assumptions hold, for all timesteps, as well as stuff like linearity of expectation under conditionals, and the argmax policy is first in the tie-breaking ordering, then if there's an expected-utility maximizer on the first step, it will keep tiling to expected utility maximizers forever afterwards. Note that this also makes the implicit assumption in the argmax that the past action sequence was done by expected utility maximizers true.

ANALYSIS:

There are a few interesting features of this proof, even though it's ultimately unsatisfactory. One interesting feature is that the argument fails when the agent has exploration steps. This feels tied to the issue where AIXI is optimal for the starting policy distribution, while AIXI+exploration is not optimal for the starting policy distribution. The proof is dissatisfactory because it heavily relies on exact equalities, which don't hold for logical inductors (rather, the relevant properties hold in the limit).

Another reason for dissatisfaction is that this is an EDT agent, and it doesn't take into account most of the more exotic decision-theory stuff. In particular, it's quite unclear what an agent of this type would do on Parfit's Hitchiker, and I would at least like to generalize it to handle cases where the future action affects the probability of some observation in the past, even if it can't quite attain the full force of UDT. There's also the issue of the conditional probabilities becoming increasingly nonsensical (and self-fulfilling prophecies showing up, such as the cosmic ray problem) if we throw out exploration steps. After all, the nice properties in the logical inductor paper only apply to sequences of actions where the probability approaches 0 slowly enough that the sum of probabilities across turns limits to infinity (such as the harmonic sequence). Finally, I'm unsure what this does on classic problems such as the procrastination paradox where you have to avoid pressing a button for as many turns as you can, but still press it eventually, and I'd be interested in analyzing that.

There are still some interesting lessons that can be taken from this, however. Namely, that the ability to steal the action of another policy is essential to show that you won't tile to it (because you can just copy the action while keeping your decision-making architecture intact, while otherwise, fair environments can have incentives to self-modify). Also, the expected utility of "I take an action" being equal to the expected utility of "I take this specific action that looks like the best so far" seems to be unusually important, because in conjunction with future-trust, it allows working around the core problem of current you not knowing what future-you will do.

Looking at the problems from Logical Inductor Tiling and Why It's Hard, we can make the following recap:

The first problem of CDT sucking can be overcome by conditioning on "I take this action"

The second problem of being bad at considering sequences of actions can be overcome by just looking one step ahead, and if you want to get certain sequences of actions that look good, you just need to strategy-steal at each step from a policy that enacts that sequence of actions, and this also has the nice property of preserving the freedom of what your action will be in the future if future-you thinks that some other action sequence is actually better.

The third problem is solved by the solution presented in the linked post.

The fourth problem of forgetting information isn't really solved head-on, because the future-trust property probably fails if the future-agent throws away information. However, it's highly suggestive that this tiling result just involves an abstract sequence of past actions, instead of unpacking them into an explicit sequence of past actions, and also this tiling result just looks one step ahead instead of considering a far-future version of itself.

The fifth problem of having strategies that react to the environment isn't really solved, but it does manage to get around the problem of not knowing what some other algorithm will eventually do. All that is really needed is that the future version of the agent knows what the other algorithm do so it can strategy-steal. This is still not a fully satisfactory solution, since there's one policy that's clearly more powerful than the rest of them, but it still seems like substantial progress has been made.

The sixth problem isn't really a fixable problem, it's just a situation where the agent has an incentive to modify its policy to something else because it violates the fairness condition.

The seventh problem about being unable to experiment with the consequences of a code rewrite is partially solved by assuming that the environment only cares about actions, and partially solved by conditioning on actions instead of "I pick this policy" (this conditional is higher probability)

Problem 8 is dodged, and is pretty much the same as the fifth problem.

Problem 9 is not addressed, because this tiling proof assumes exact equalities, and doesn't deal with the noise from logical inductors not being perfect impairing sufficiently long chains of trust. However, the fact that the proof only looks one step ahead feels promising for not having compounding noise over time breaking things.

Cause Awareness as a Factor against Cause Neutrality

[Epistemic status: Sudden insight + some reflection]

[Novelty status: I Googled some stuff on cause neutrality, and didn't see this category of concerns mentioned]

Trying to live an optimized, impactful life can be quite a burden. There is no time for fun when there is world-saving to do. You can't learn math unless it helps your career. And you must sacrifice the charitable causes you most care about for the ones that have the most impact.

What a relief when, last week, I realized how it may be possible to support personal pet causes while living an optimized life.

Cause-neutrality is one of the cornerstones of EA thought. It goes like this: Bob took a trip to a village in Indonesia and was really struck by the poverty there. He wants to dedicate his life to helping people in this village because it can do a lot of good. Boo Bob!

For you see, this is just one of many impoverished villages around the world, and probably neither the one suffering most nor the one where he can have the biggest impact cheaply. He should instead go to the List of Suffering Villages™ and pick the one where he can do the most good. And that's assuming it's not better for him to just become an investment banker and donate all his money.

But to make this decision, first the village must be on the list.

We live in a world where every village has a Wikipedia entry. But in other areas, there are a large number of local problems where the resources needed to raise awareness about the problem are a substantial fraction of the resources needed to solve it. In these cases, working on it immediately can be more effective than trying to get it properly prioritized and allocated to the most efficient person.

For comparison, consider personal productivity. The philosophy of *Getting Things Done* is all about moving tasks into a centralized system so you can do them at an optimal time. But, there's an exception: if something takes less than two minutes, you should do it now.

In altruism, this looks like: If you see someone bleeding on a deserted street, you call 911 instead of weighing its impact it against all the phone calls to Congress you should be making. But it also applies in cases where the cost/benefit ratio is less extreme.

This is most evident in obscure political causes. Suppose you're a skilled professional whose time is worth $100/hour. One day, you realize that your local town council can produce 10 million utilons by adopting policy X. You think you can convince them to adopt this policy by spending 500 hours running a campaign. But, an experienced campaign manager whose time is worth $80/hour could do it in 250 hours. If you do it: $50,000 for 10 million utilons. If he does it: $20,000 for 10 million utilons. Profit!

But what if you need to spend 50 hours finding and interviewing candidates for the campaign manager, and another 100 hours explaining the problem to him and introducing him to important people in your town? Now you're up to $35,000 for the campaign manager. And what if there's a 100% chance of success if you do it yourself, versus only a 70% chance if you try to hire someone else (who, of course, may not be as good as his resume claims)? Now you're looking at 200 utilons/$ in expectation whether you do it yourself or hire someone. If you're at all risk-averse, then…..this is how it can make sense for a professional engineer to wind up running a political campaign.

So, the problems of cause-neutrality are the same as the problems of delegating any other task. Risk, transaction costs, and management can eat up all the efficiency gains.

This idea means you should weaken the recommendations of cause-neutrality to: spend resources on your pet causes in an amount inversely proportional to how well-known the cause is. If you have a burning hatred of cancer because it killed your parents, you should probably still give your money to malaria organizations rather than cancer researchers. But if there's an epidemic of salmonella in a small town you're visiting and you happen to be a famous doctor, it can still be effective to spend time helping salmonella patients rather than trying to bring in nurses so you can go do famous-doctor-y things.

Emotional Training Model

Bearlamp: Previous, First

Greaterwrong: Previous, First

Lesswrong: Previous, First

Life is propagated by two main clusters of emotions. (*Yes I know it's more like a spectrum but this is the poor simplification I am using for now)

The "good" feelings that we move towards, and the "bad" feelings that we move away from. (then there's the neutral ones we hang around in sometimes but that's for another time).

If you spend your life always running from the bad ones and always running towards the good ones, **you may have a good life**. You may have a life that just gets lucky and has more good than bad. Alternatively **you may have hard things to do that involve feeling uncomfortable** for short or long periods of time. If you are forever running away from the bad emotions, and forever addicted to running towards the good emotions, you are severely limited in your agency compared to if you have even a little bit of freedom to do something like, "avoid short term rewards", or "put up with scary moments" on the way to other experiences. (see also The Trauma model of mental health)

The carnival ride example

Charlie the five year old goes on a carnival ride. Charlie the five year old throws up. Charlie the five year old learns that carnival rides make you feel terrible.

In (one of many) healthy worlds: When charlie turns ten, his friends ask him to go to the carnival again. Charlie realises that the carnival ride might be different now, he fuels himself with a bit of peer pressure and he runs a new experiment, pushing back on the terrible feeling that he would usually avoid and rewrites his inclination to avoid terrible things. Consequently charlie relearns that a carnival ride is only sometimes terrible and with the support of friends it can be good.

In (one of many) unhealthy worlds: When charlie turns ten, his friends ask him to go to the carnival again. Charlie remembers that the carnival rides feel terrible and decides not to go to the carnival. This reinforces the terrible feeling. Charlie feels entirely justified in avoiding a terrible thing, his friends don't really care either way and life goes on. Charlie keeping a tiny reinforced experience that he should avoid terrible things.

Objectively speaking, a carnival ride is not terrible or good. Subjectively, the feelings we attach to such experiences are what guides us in future experiences. Rightly or wrongly, all possible futures for charlie are going to be guided by the possibility that those emotions will come up.

In an ideal world, our emotions, our s1 will be trained accurately from our surroundings.

In prehistoric times, we avoid the crocodile lake because we feel scared of the danger there. The humans who didn't feel scared of the danger, didn't avoid the lake, didn't live, and didn't pass on their genes.

Unfortunately we don't live in an ideal world for emotional training, and despite the best of intentions we can still wind up with emotional maps that don't help us to win at life.

The good news is that we can re-train our early emotional models of the world. The bad news is that it's probably the hardest thing I've ever done, and people spend years meditating on mountain tops for equanimity towards all experiences.

With that in mind - let's begin.

Practice

The destination is the ability to feel uncomfortable feelings. The result is to get to the other side. Unfortunately like all cryptic journeys, you can't be too focused on the result or you will miss the whole "value in the journey" thing that all those wise people talk about. You can think of this practice as a meditation on feelings

(This looks similar to ACT, that's because it is.)

To start - ask yourself, "How am I feeling?". That will give you an entry point. *There's always an entry point. Even if it's* ***confused****, or* ***I feel like not doing this exercise right now****, or* ***I feel like being distracted by that other tab****.*

Then ask, "What is it like being me right now feeling X?". This question develops a relationship to the thoughtstream.

For our 3 examples above:

1. **Confused** feels like **Silly**
2. **Not doing the exercise** feels **rebellious**
3. **Distraction** Feels **exciting** or **Guitly**

Repeat the question with the new find, "What does it feel like being me right now feeling X?". Building the ongoing relationship with the thoughtstream.

1. **Silly** feels like **A clown that people would poke fun at** *(this is an example of a metaphorical pointer to a feeling)*
2. **Rebellious** feels **empowering** but then also **scary**
3. **Exciting** Feels like **missing out** or **Like being stuck in the classroom during lunch time when everyone else is outside playing** (*This is another example of the metaphor)*
4. **Guilty** feels like **a heavy weight in my chest** *(this is an example of a physical manifestation of a feeling)*

These paths further each might open up into other feeling paths.

-
- **A clown that people would poke fun at**

- What does it feel like to be the clown anyway?
- What does it feel like to be the person laughing at the clown here?
- What does it feel like to be poked fun at?
- **Empowering**

- What does it feel like being me right now, feeling empowered?
- What if I did the opposite? How would that feel?
- **Scary**

- What does it feel like being scared like this right now?
- **Missing out**

- What does it feel like to miss out right now?
- **Like being stuck in the classroom during lunch time when everyone else is outside playing**

- What does it feel like to be stuck in the classroom?
- **a heavy weight in my chest**

- What does it feel like to be heavy in the chest right now?

-

For the purpose of example, I've generated multiple paths. In practice I'd be looking to go down one path at a time. That might look like this:

1. How am I feeling right now?
2. I feel confused about the exercise
3. What is it like being me right now feeling confused?
4. Silly
5. What is it like being me right now feeling silly?
6. Like a clown being poked fun at
7. What does it feel like to be the clown?
8. Embarrasing …

And onward through several feelings. At some point, it becomes useful to not run from a feeling to the next feeling, and instead sit on it for a moment. That might be after 10, 20 or 30+ different stops along the journey.

(in the interest of being brief I'm going to stop at 8 instead of 30) At 8, that means feeling embarrassed, but instead of asking myself for the answer of what it feels like to feel embarrassed - I stop and try to feel what it feels like to feel embarrassed.

Instead of looking for a word underneath embarrassed, I feel the feeling of embarrassed. And wait. And it's uncomfortable, but to get distracted by the uncomfortable feeling would be to leave embarrassed. So I go back to embarrassed. And it gets heavy. And to get distracted onto heavy would be to not be embarrassed any more. And it feels like something is crushing my chest, and it's getting tighter. And it might crush me, and I might not breathe. And I wait.

And then it stops crushing. And it softens, and it eases, and it levels out to a different feeling. And I take a deep breath. And I feel calm. A very deep sense of calm. I feel like I'd be okay being embarrassed. As long as I remember that there's a sense of calm underneath.

And I feel calm. And I feel relieved, and complete.

And that’s what it feels like to feel an uncomfortable feeling and get to the other side. That’s what it feels like to untrain the carnival ride effect.

Next post: Feedback from emotions

New paper: Long-Term Trajectories of Human Civilization

Long-Term Trajectories of Human Civilization (free PDF). *Foresight*, forthcoming, DOI 10.1108/FS-04-2018-0037.

Authors: Seth D. Baum, Stuart Armstrong, Timoteus Ekenstedt, Olle Häggström, Robin Hanson, Karin Kuhlemann, Matthijs M. Maas, James D. Miller, Markus Salmela, Anders Sandberg, Kaj Sotala, Phil Torres, Alexey Turchin, and Roman V. Yampolskiy.

Abstract
Purpose: This paper formalizes long-term trajectories of human civilization as a scientific and ethical field of study. The long-term trajectory of human civilization can be defined as the path that human civilization takes during the entire future time period in which human civilization could continue to exist.
Approach: We focus on four types of trajectories: status quo trajectories, in which human civilization persists in a state broadly similar to its current state into the distant future; catastrophe trajectories, in which one or more events cause significant harm to human civilization; technological transformation trajectories, in which radical technological breakthroughs put human civilization on a fundamentally different course; and astronomical trajectories, in which human civilization expands beyond its home planet and into the accessible portions of the cosmos.
Findings: Status quo trajectories appear unlikely to persist into the distant future, especially in light of long-term astronomical processes. Several catastrophe, technological transformation, and astronomical trajectories appear possible.
Value: Some current actions may be able to affect the long-term trajectory. Whether these actions should be pursued depends on a mix of empirical and ethical factors. For some ethical frameworks, these actions may be especially important to pursue.

An excerpt from the press release over at the Global Catastrophic Risk Institute:

> *Society today needs greater attention to the long-term fate of human civilization. Important present-day decisions can affect what happens millions, billions, or trillions of years into the future. The long-term effects may be the most important factor for present-day decisions and must be taken into account. An international group of 14 scholars calls for the dedicated study of "long-term trajectories of human civilization" in order to understand long-term outcomes and inform decision-making. This new approach is presented in the academic journal Foresight, where the scholars have made an initial evaluation of potential long-term trajectories and their present-day societal importance.*
>
> *"Human civilization could end up going in radically different directions, for better or for worse. What we do today could affect the outcome. It is vital that we understand possible long-term trajectories and set policy accordingly. The stakes are quite literally astronomical," says lead author Dr. Seth Baum, Executive Director of the Global Catastrophic Risk Institute, a non-profit think tank in the US.*
>
> *The group of scholars including Olle Häggström, Robin Hanson, Karin Kuhlemann, Anders Sandberg, and Roman Yampolskiy have identified four types of long-term trajectories: status quo trajectories, in which civilization stays about the same, catastrophe trajectories, in which civilization collapses, technological*

transformation trajectories, in which radical technology fundamentally changes civilization, and astronomical trajectories, in which civilization expands beyond our home planet.

Available here: http://gcrinstitute.org/papers/trajectories.pdf

Counterfactuals for Perfect Predictors

Parfit's Hitchhiker with a perfect predictor has the unusual property of having a Less Wrong consensus that you ought to pay, whilst also being surprisingly hard to define formally. For example, if we try to ask about whether an agent that never pays in town is rational, then we encounter a contradiction. A perfect predictor would not ever give such an agent a lift, so by the Principle of Explosion we can prove any statement to be true given this counterfactual.

On the other hand, even if the predictor mistakenly picks up defectors only 0.01% of the time, then this counterfactual seems to have meaning. Let's suppose that a random number from 1 to 10,000 is chosen and the predictor always picks you up when the number is 1 and is perfect otherwise. Even if we draw the number 120, we can fairly easily imagine the situation where the number drawn was 1 instead. This is then a coherent situation where an Always Defect agent would end up in town, so we can talk about how the agent would have counterfactually chosen.

So one response to the difficulties of discussing counterfactual decisions with perfect predictors would be to simply compute the counterfactual as though the agent has a (tiny) chance of being wrong. However, agents may quite understandably wish to act differently depending on whether they are facing a perfect or imperfect predictor, even choosing differently when facing an agent with a very low error rate.

Another would be to say that the predictor predicts whether placing the agent in town is logically coherent. On the basis that the agent only picks up those who it predicts (with 100% accuracy) will pay, it can assume that it will be payed if the situation is coherent. Unfortunately, it isn't clear what this means in concrete terms for an agent to be such that it couldn't coherently be placed in such a situation. How is, "I commit to not paying in <impossible situation>" any kind of meaningful commitment at all? We could look at, "I commit to making <situation> impossible", but that doesn't mean anything either. If you're in a situation, then it must be possible? Further, such situations are contradictory and *everything* is true given a contradiction, so all contradictory situations seem to be the same.

As the formal description of my solution is rather long, I'll provide a summary: We will assume that each possible world model corresponds to at least one possible sequence of observations. For world models that are consistent conditional on the agent making certain decisions, we'll take the set of observations for agents that are consistent and feed it into the set of agents who aren't. This will be interpreted as what they would have counterfactually chosen in such a situation.

A Formal Description of the Problem

(You may wish to skip directly to the discussion)

My solution will be to include observations in our model of the counterfactual. Most such problems can be modelled as follows:

Let x be a label that refers to one particular agent that will be called the *centered agent* for short. It should generally refer to the agent whose decisions we are optimising. In Parfit's Hitchhiker, x refers to the Hitchhiker.

Let W be a set of possible "world models with holes". That is, each is a collection of facts about the world, but not including facts about the decision processes of x which should exist as an agent in this world. These will include the problem statement.

To demonstrate, we'll construct I for this problem. We start off by defining the variables:

- t: Time
  - 0 when you encounter the Driver
  - 1 after you've either been dropped off in Town or left in the Desert
- l: Location. Either Desert or Town
- Act: The actual action chosen by the hitchhiker if they are in Town at t=1. Either Pay or Don't Pay or Not in Town
- Pred: The driver's prediction of x's action if the driver were to drop them in town. Either Pay or Don't Pay (as we've already noted, defining this counterfactual is problematic, but we'll provide a correction later)
- u: Utility of the hitchhiker

We can now provide the problem statement as a list of facts:

- Time: t is a time variable
- Location:
  - l=Desert at t=0
  - l=Town at t=1 if Pred=Pay
  - l=Desert at t=1 if Pred=Don't Pay
- Act:
  - Not in Town at t=0
  - Not in Town if l=Desert at t=1
  - Pay or Don't Pay if l=Town at t=1
- Prediction: The Predictor is perfect. A more formal definition will have to wait
- Utility:
  - u=0 at t=0
  - At t=1: Subtract 1,000,000 from u if l=Desert
  - At t=1: Subtract 50 from u if Act=Pay

W then contains three distinct world models:

- Starting World Model - w1:
  - t=0, l=Desert, Act=Not in Town, Pred: varies, u=0
- Ending Town World Model - w2:
  - t=1, l=Town, Act: varies, Pred: Pay, u: varies
- Ending Desert World Model - w3:
  - t=1, l=Desert, Act: Not in Town, Pred: Don't Pay, u=-1,000,000

The properties listed as varies will only be known once we have information about x. Further, it is impossible for certain agents to exist in certain worlds given the rules above.

Let O be a set of possible sequences of observations. It should be chosen to contain all observations that could be made by the centered agent in the given problem and there should be at least one set of observations representing each possible world model with holes. We will do something slightly unusual and include the problem statement as a set of observations. One intuition that might help illustrate this is to imagine that the agent has an oracle that allows it to directly learn these facts.

For this example, the possible individual observations grouped by type are:

- Location Events: <l=Desert> OR <l=Town>
- Time Events: <t=0> OR <t=1>
- Problem Statement: There should be an entry for each point in the problem statement as described for I. For example:
    - <l=desert at t=0>

O then contains three distinct observation sequences:

- Starting World Model - o1:
    - <Problem Statement> <t=0> <l=Desert>
- Ending Town World Model - o2:
    - <Problem Statement> <t=0> <l=Desert> <t=1> <l=Town>
- Ending Desert World Model - o3:
    - <Problem Statement> <t=0> <l=Desert> <t=1> <l=Desert>

Of course, <t=0><l=Desert> is observed initially in each world so we could just remove it to provide simplified sequences of observations. I simply write <Problem Statement> instead of explicitly listing each item as an observation.

Regardless of its decision algorithm, we will associate x with a fixed Fact-Derivation Algorithm f. This algorithm will take a specific sequences of observations o and produce an id representing a world model with holes w. The reason why it produces an id is that some sequences of observations won't lead to a coherent world model for some agents. For example, the Ending in Town Sequence of observers can never be observed by an agent that never pays. To handle this, we will assume that each incomplete world model w is associated with a unique integer [w]. In this case, we might logically choose, [w1]=1, [w2]=2, [w3]=3 and then f(o1)=[w1], f(o2)=[w2], f(o3)=[w3]. We will define m to map from these id's to the corresponding incomplete world model.

We will write D for the set of possible decisions algorithms that x might possess. Instead of having these algorithms operate on either observations or world models, we will make them operate on the world ids that are produced by the Fact-Derivation Algorithm so that they still produce actions in contradictory worlds. For example, define:

- d2 - Always Pay
- d3- Never Pay

If d2 sees [O3] or d3 sees [O2], then it knows that this is impossible according to its model. However, it isn't actually impossible as its model could be wrong. Further, these "impossible" pre-commitments now mean something tangible. The agent has pre-committed to act a certain way if it experiences a particular sequence of observations.

We can now formalise the Driver's Prediction as follows for situations that are only conditionally consistent (we noted before that this needed to be corrected). Let o be the sequence of observations and d0 be a decision algorithm that is consistent with o, while d1 is a decision algorithm that is inconsistent with it. Let w=m(f(o)), which is a consistent world given d0. Then the counterfactual of what d1 would do in w is defined as: d1(f(o)). We've now defined what it means to be a "perfect predictor". There is however one potential issue, perhaps multiple observations led to w? In this case, we need to define the world more precisely and include observational details in

the model. Even if these details don't seem to change the problem from a standard decision theory perspective, they may still affect the predictions of actions in impossible counterfactuals.

Discussion

In most decision theory problems, it is easier to avoid discussing observations any more than necessary. Generally, the agent makes some observations, but their knowledge of most of the setup is mostly assumed. This abstraction generally works well, but it leads to confusion in cases like this where we are dealing with predictors who want to know if they can coherently put another agent in a specific situation. As we've shown, even though it is meaningless to ask what an agent would do given an impossible situation, it is meaningful to ask what the agent would do given an impossible input.

When asking what any real agent would do in a real world problem, we can always restate it as asking about what the agent would do given a particular input. However, using the trick of separating observations doesn't limit us to real world problems; as we've seen, we can use the trick of representing the problem statement as direct observations to represent more abstract problems. The next logical step is to try extending this to cases such as, "What if the 1000th digit of Pi were even?" This allows us to avoid the contradiction and deal with situations that are at least consistent, but it doesn't provide much in the way of hints of how to solve these problems in general. Nonetheless, I figured that I may as well start with the the one problem that was the most straightforward.

Update: After rereading the description of Updateless Decision Theory, I realise that it is already using something very similar to this technique as described here. So the main contribution of this article seems to be exploring a part of UDT that is normally not examined in much detail.

One difference though is that UDT uses a Mathematical Intuition Function that maps from inputs to a probability distribution of execution histories, instead of a Fact-Derivation Algorithm that maps from inputs to models and only for consistent situations. One advantage of breaking it down as I do is to clarify that UDT's observation-action maps don't only include entries for possible observations, but observations that it would be contradictory for an agent to make. Secondly, it clarifies that UDT predictors predict agents based on how they respond to inputs representing situations, rather than directly on situations themselves, which is important for impossible situations.

Links:

- Turns out Cousin_it actually discussed this problem many years before me. He points out that when the situation is inconsistent, we can run into issues with spurious counterfactuals.

AI Reading Group Thoughts (1/?): The Mandate of Heaven

My housemate Kelsey "theunitofcaring" has begun hosting an AI reading group in our house. Our first meeting was yesterday evening, and over a first draft attempt at chocolate macarons, we discussed this article about AI safety and efficiency by Paul Christiano, and various ideas prompted thereby at greater or lesser remove.

One idea that came up is what we decided to call "tipping point AI" (because apparently there are a lot of competing definitions for "transformative" AI). The definition we were using for tipping point AI was "something such that it, or its controller, is capable of preventing others from building AIs". The exact type and level of capability here could vary - for instance, if it's built after we've colonized Mars (that is, colonized it to an extent such that Martians could undertake projects like building AIs), then a tipping point AI has to be able to project power to Mars in some form, even if the only required level of finesse is lethality. But if it's before we've colonized Mars, it can be unable to do that, and just able to prevent colonization projects in addition to AI projects.

One hypothesis that has been floated in a context such that we are pretty sure it is not anyone's real plan is that an AI could just destroy all the GPUs on the planet and prevent the manufacture of new ones. This would be bad for Bitcoins, video games, and AI projects, but otherwise relatively low-impact. An AI might be able to accomplish this task by coercion, or even by proxy - the complete system of "the AI, and its controller" needs to be able to prevent AI creation by other agents, so the AI itself might only need to identify targets for a controller who already wields enough power to fire missiles or confiscate hardware and chooses to do so in service of this goal, perhaps the US government.

The idea behind creating tipping point AI isn't that this is where we stop forever. The tipping point AI only has to prevent other agents from building their own in their basements. It eliminates competition. Some features of a situation in which a tipping point AI exists include:

- The agent controlling the AI can work on more sophisticated second drafts without worrying about someone else rushing to production unsafely.
- The controlling agent can publish insights and seek feedback without worrying about plagiarism, code forks, etc.
- They can apply the AI's other abilities, if any (there will presumably be some, since "prevent AI creation" is not a primitive action - some surveillance capability seems like a minimum to me) to their other problems, perhaps including creating a better AI. Even if this application has economic or other benefits that might attract others to similar solutions by default, the AI will prevent that, so no one will be (productively) startled or inspired into working on AI faster by seeing the results.

However, if you're an agent controlling a tipping point AI, you have a problem: the bus number* of the human race has suddenly dropped to "you and your cohort". If anything happens to you - and an AI being tipping point variety doesn't imply it can help you with all of the things that might happen to you - then the AI is leaderless. This, depending on its construction, might mean it goes rogue and does something

weird, that it goes dormant and there's no protection against a poorly built new AI project, or that it keeps doing whatever its last directive was (in the example under discussion, "prevent anyone from building another AI"). None of these are good states to have obtain permanently.

So you might want to define, and then architect into your AI the definition of, organizational continuity, robustly enough that none of those things will happen.

This isn't trivial - it's almost certainly easier than defining human value in general, but that doesn't mean it's simple. Your definition has to handle internal schisms, both overt and subtle, ranging from "the IT guy we fired is working for would-be rivals" to "there's serious disagreement among our researchers about whether to go ahead with Project Turaco, and Frances and Harold are working on a Turaco fork in their garage". If you don't want the wrong bus accident (or assassination) to mean that humanity ends, encounters a hard stop in its technological progress, or has its panopticonic meddling intelligence inherited by a random person who chose the same name for their uber-for-spirulina business? Then you need to have a way to pass on the mandate of heaven.

One idea that popped into my head while I was turning over this problem was a code of organizational conduct. This allows the organization to resume after a discontinuity, without granting random people a first-mover advantage at picking up the dropped mantle unless they take it up whole. It's still a simpler problem than human value in general, but it's intermediate between that and "define members of a conventional continuous group of humans". The code has to be something that includes its own decisionmaking process - if six people across the globe adopt a code simultaneously they'll need to resolve conflicts between them just as much as the original organization did. You presumably want to incorporate security features that protect both against garage forks of Projects Turaco and also against ill-intentioned or not-too-bright inheritors of your code.

Other options include:

- Conventional organizational continuity. You have, perhaps, a board of directors who never share a vehicle, and they have some sort of input into the executives of the organization, and you hope nobody brings the plague to work, and there is some sort of process according to which decisions are made and some sort of process for defaulting if decisions fail to be made.
- Designated organizational heirs: if your conventional organization fails, then your sister project, who are laying theoretical groundwork but not building anything yet because you have a tipping point AI and you said so, get the mandate of heaven and can proceed. This assumes that you think their chances of achieving value alignment are worse than yours but better than any other option. This has obvious incentive problems with respect to the other organization's interest in yours suddenly ceasing to exist.
- Non-organization based strategies (a line of succession of individuals). People being changeable, this list would need to be carefully curated and carefully maintained by whoever was ascendant, and it would be at substantial risk of unobserved deception, errors in judgment, or evolution over time of heirs' interests and capabilities after their predecessors can no longer edit the line of succession. These would all be capable of affecting the long term future of humanity once the AI changed hands.
- I'm sure there are things I haven't thought of.

I don't have a conclusion because I just wrote this about thoughts that I had in response to the meeting, to let other people who can't attend still be in on some of what we're talking and thinking about.

*The number of people who can be hit by a bus before the organization ceases to function

Aumann’s Agreement Revisited

This post is about reviewing some of the slightly counter-intuitive conclusions that Baysean probability theory has on agreement between people’s probabilities about the same belief. You have probably heard something about Aumann’s agreement theorem saying that “with common priors, people’s estimates should converge upon gaining common knowledge about their estimates.” However, it’s worth noting that this holds only for case of common priors, new data in the form of estimates and when people have the same models of one another’s rationality. In case the priors are not common, arbitrary new data could create a further divergence in likelihoods. In case people have the same priors on propositions, but different priors on each other’s rationality, Aumann’s agreement doesn’t apply after sharing likelihoods. The theorem is still a theorem and obviously still holds given its assumptions, however the real question is how realistic those assumptions are. I would generally argue that a model that considers some partial mistrust of each other’s rationality is more realistic and could help explain why we don’t see Aumann’s agreement in practice.

Sample misuse of the theorem:

Her: Two rational people are supposed to agree. He doesn't agree with me, so he must not be rational.

Him: Aumann's agreement theorem does not apply with uncommon priors, which include priors over differences in trust between self and other.

I will first go through simple cases building up examples of agreement / disagreement. If you are interested in just the “agreement with mistrust of rationality” result, skip to the bottom.

The first example is about people with different priors reacting to new evidence. Taken from Probability the Logic of Science:

The new information D is: ‘Mr N has gone on TV with a sensational claim that a commonly used drug is unsafe’, and three viewers, Mr A, MrB, and Mr C, see this. Their prior

probabilities P(S|I) that the drug is safe are (0.9, 0.1, 0.9), respectively; i.e. initially, Mr A and Mr C were believers in the safety of the drug, Mr B a disbeliever. But they interpret the information D very differently, because they have different views about the reliability of Mr N. They all agree that, if the drug had really been proved unsafe, Mr N would be right there shouting it: that is, their probabilities P(D|SI) are (1, 1, 1); but Mr A trusts his honesty while Mr C does not. Their probabilities P(D|SI) that, if the drug is safe, Mr N would say that it is unsafe, are (0.01, 0.3, 0.99), respectively. Applying Bayes'theorem P(S|DI)= P(S|I) P(D|SI)/P(D|I),and expanding the denominator by the product and sum rules, P(D|I) = P(S|I) P(D|SI)+ P(S|I) P(D|SI), we find their posterior probabilities that the drug is safe to be (0.083,0.032,0.899),respectively.

A: 'Mr N is a fine fellow, doing a notable public service. I had thought the drug to be safe from other evidence, but he would not knowingly misrepresent the facts; therefore hearing his report leads me to change my mind and think that the drug is unsafe after all. My belief in safety is lowered by 20.0 db, so I will not buy any more.'

B: 'Mr N is an erratic fellow, inclined to accept adverse evidence too quickly. I was already convinced that the drug is unsafe; but even if it is safe he might be carried away into saying otherwise. So, hearing his claim does strengthen my opinion, but only by 5.3 db. I would never under any circumstances use the drug.'

C 'Mr N is an unscrupulous rascal, who does everything in his power to stirup trouble by sensational publicity. The drug is probably safe, but he would almost certainly claim it is unsafe whatever the facts. So, hearing his claim has practically no effect (only 0.005db) on myconfidence that the drug is safe. I will continue to buy it and use it.'

The opinions of Mr A and Mr B converge – become closer to each other because both are willing to trust Mr N's veracity to some extent. But Mr A and Mr C diverge because their prior probabilities of deception are entirely different.

Note that while both A and C both update downwards on the drug reliability, their opinions diverge in the terms of the log ratio of probabilities increasing rather than decreasing. This is a good counter-example to twisting the agreement theorem from to create some sort of assumption of convergence on arbitrary new data. Yes, that means that even two rational people's beliefs in a proposition could diverge after learning something.

A second example illustrates a similar point with regards to considering multiple hypothesis. Here we are considering an experiment which tries to prove that a person has ESP – Extra Sensory Perception.

In the reported experiment, from the experimental design the probability for guessing a card correctly should have been p = 0.2, independently in each trial. Let Hp be the 'null hypothesis' which states this and supposes that only 'pure chance' is operating (whatever that means). According to the binomial distribution, Hp predicts that if a subject has no ESP, the number r of successful guesses in n trials should be about (mean ± standard deviation)

For n =37100 trials, this is 7420±77.

But, according to the report, Mrs. Gloria Stewart guessed correctly r = 9410 times in 37100 trials, for a fractional success rate of f = 0.2536. These numbers constitute our data D. At first glance, they may not look very sensational; note, however, that her score was (9410−7420) /77 = 25.8 standard deviations away from the chance expectation.

This basically gives an estimated P-value of 3.15×10^−139

Forget the debate of whether to lower the P-value! Very few experiments have that kind of p-value, so any reasonable threshold would make this pass. Forget the debate of whether this is enough data to update the prior. You can be an ESP skeptic and there is no reason your prior should be lower than 3.15×10^−139!

So, I am not asking you to believe in ESP, rather the question is how to update probabilities in this case? The simple answer is that instead of considering the two hypotheses:

H_E = ESP is true,

H_null = this result happens by chance,

you need to consider the three hypotheses:

H_E = ESP is true,

H_null = this result happens by chance

H_fake = the experiment had an error in its setup.

Based on the data, we can rule out the H_null convincingly and our resulting ratios of how likely ESP is true depends on the ratio of priors of H_E and H_fake. So, for example, if we have person A, who is a small amount of skeptic in both science and ESP with H_E_A = 10^−4% and H_fake_A = 10% and person B who is a bit more of believer in both with H_E_B = 2% and H_fake_B = 0.5%, then their resulting likelihoods will be around:

P(H_E_A | D) = 10^−3%

P(H_fake_A | D) = 100 - 10^−3%

P(H_E_B | D) = 80%

P(H_fake_B | D) = 20%

So, B is convinced by the evidence in ESP because the data pulled his prior in ESP more than the prior in the experiment being fake or false. A is now convinced with ~ 100% that the experiment was false.

I have heard a strange objection to this experiment and that being that experiments with that kind of p-value are somehow less likely to be true because the p-value is unusual. While people don't generally publish these kinds of the results due to the fear of being unusual, there is nothing inherently wrong with a discovering a low p-val on a large N experiment trials. There are plenty of real processes that could give this low number.

What does this mean? In practice this means that if the theoretical agent is skeptical of your rationality with P (your irrationality) = p, that means it's hard for you to push them to update a lot on a hypothesis which they have a prior of < p.

The last example is from the original Aumann paper and illustrated how agreement should work.

As an illustration, suppose 1 and 2 have a uniform prior over on the parameter – chance that a coin comes up heads. This is the probability of H/T is a random number between 0 and 1. Let A be an event that a coin comes up heads the next toss. It's a 50% prior. Person 1 and 2 each observe a single coin toss privately. It's H and T respectfully. Their new likelihoods for the coin parameter are 2/3 and 1/3. However, after learning each other's likelihoods, they immediately deduce what the other's coin toss was and update it back to 50%. This is an example of convergence with a single sharing of updates.

However, if there is some other unknown number of tosses that have caused 1 and 2 to arrive at their prior, then the update takes two steps – the first update reveals the private amount of tosses that 1 and 2 have observed privates and the second update combines them into a single estimate. Note, I haven't checked the math myself, but I am assuming the original paper is going to be correct.

So, if followed the 3 examples carefully, then you might say that they seem to show two very different and somewhat unrelated findings:

1. With uncommon priors, new arbitrary information could create a divergence in likelihoods

2. With common priors, sharing knowledge about subjective likelihoods should eventually result in convergence.

This is true in theory in that the two theories don't contradict each other, however in practice, the application is somewhat complex. The main issue with Aumann's theorem is the assumption of "common priors". This has been discussed at length, with various solutions, such as punting the issue to deeper meta-priors. see citations here

I want to consider another example of when "uncommon priors" prevent agreement – when people have a different model of each other's rationality. In a lot of places that mention the agreement theorem, there is the assumption that two people assume each other's rationality with 100% probability.

The moment this assumption is relaxed a bit, to where two people assume each other's rationality on a given subject with 80% probability, while assuming their own rationality with 100% probability, then learning likelihoods can act as evidence towards the other person's irrationality on this subject as well as towards convergence of beliefs.

Here is a sample setup. Person A and person B are trying to consider the probability of a coin toss. Each one observes the coin toss through a private oracle which has 99% accuracy. In addition, each one considers themselves rational, but does not trust the other person completely - they consider each other irrational with 80% chance.

In case of irrationality of B, A believes that the statement B makes is always 99% tails and can be ignored. Consider the case when the coin came up heads, but oracle has told A it's heads and told B it's tails (it's the 1% chance of the oracle being wrong). This happens 99% 1% =0.99% of the time.

Note that the perspectives of A and B are symmetrical with respect to each other's probability. Each one will report that they have 99% of heads and tails respectfully. What happens in the update after they share information?

From A's perspective before the sharing, this can be divided into 4 categories. Here B_rat is the proposition that b is rational, H is the probability of heads, E is the evidence in presented in the situation.

P (B_rat) = 80%

P (H) = 99%

P (B_rat & H & E) = 80% * 99% * 1% = 0.792%

P(B_rat & !H & E) = 80% * 1% * 99% = 0.792%

P(!B_rat & H & E) = 20% * 99% =19.8%

P(!B_rat & !H & E) = 20% * 1% = 0.2%

Thus we get:

P(E) = 21.584%

P(! B_rat | E) = P(!B_rat & E) | P(E) = 20/21.584 ~ 92.6%

P (!H |E) = P(!H & E) | P(E) = (0.792 + 0.2) / 21.584 ~ 4.6%

So, in other words after hearing each other's estimates, both will update their hypothesis that they are wrong from 1% to 4.6%. However, they will update the hypothesis that the other person is irrational from 20% to 92.6%. What happens after this update? In the least convenient possible world each person assumes at the other would act in the exact same way in the case of rationality and irrationality and thus no further updates are possible. ,

What does this mean? This suggests an alternative explanation for the notion of "agreeing to disagree." Instead of viewing this as a series of contradictory statement about two people trusting each other's rationality perfectly and having different credences, it could instead be interpreted as two people trusting each other's rationality generally, but not on a particular subject.

The conclusion is that idea of "rational people can't agree to disagree" is too strong and the normal human ability to agree to disagree does indeed have a theoretical basis. The practical advice is generally in line with spending your weirdness points wisely - to make sure to not and try to convince people of statements of too low probability if their original trust in you is too low.

Wanting people to take each other's beliefs as "some evidence" is an important and positive desire. I have seen many instances where discussion would benefit from taking someone's belief into account.

However, the expectation that everybody "must" agree with each other about everything after sharing a small amount of beliefs is not realistic. The broader philosophical meta-point is to be careful when trying to use math to "prove" that a common-sense thing cannot or should not exist.

Fundamentals of Formalisation Level 7: Equivalence Relations and Orderings

Followup to Fundamentals of Formalisation Level 6: Turing Machines and the Halting Problem. First post.

This is a new lesson of our online course on math formalizations required for AI safety research.

The big ideas:

- Equivalence Relations
- Partitions
- Orderings

To move to the next level you need to be able to:

- Explain the relationship between equivalence relations and partitions.
- Name two different kinds of orderings and the conditions on the ordering relations required for these kinds.

Why this is important:

- This level is about further building up your mathematical toolbox. While equivalence relations and orderings currently seem like random abstract notions, they are both very important. The former gives you a way to instantiate different notions of equality, which is important in modeling. The latter appears in constructing models for another turing machine equivalent notion of computability, the lambda calculus.

For every lesson you have 2 options: do the whole thing, or skip to the questions and exercises in the end. The latter option is for people who suspect they already know the subject. It serves as a means of verifying or falsifying that hypothesis.

Theories of Pain

Epistemic status: Exploratory. I did basically no actual research for this post so please don't take anything I say at face value.

Follow-up to: Book Review: Unlearn Your Pain, A Sarno-Hanson Synthesis

.

I've had problems with pain for basically my whole life, and over the years a lot of people have told me a lot of different things about why the pain happens and what I can do to fix it. This is my attempt to put all of those things in one place and see if I can make something coherent out of them.

This post grew out of the observation that for every chronic pain intervention I've encountered, there are hundreds of people who swear by it as a miracle cure (though it doesn't work for everyone), so each thing is probably at least gesturing in the direction of something true. But most practitioners are very attached to their own view of how pain works and view at least some of the others as crackpots, so I haven't seen anyone try to tie all of the theories together.

I start with brief explanations of all the theories I've looked into for more than five minutes, divided imperfectly into physical explanations and psychological explanations and presented roughly in chronological order of when I encountered them. Then I try to inelegantly smash them all together. (Note that this is basically the same path that Todd Hargrove followed, only he spent many years on it instead of a couple hours, so maybe just read his book instead.)

Survey of theories

Physical explanations

The medical establishment (in the United States)

In my experience, complaints of chronic joint pain lead doctors to test your blood for rheumatic diseases and maybe X-ray any particularly painful bones or joints. If they don’t find evidence of disease using these methods, they kind of throw up their hands and walk away. If there is something visibly wrong with a bone/joint, they either 1) recommend physical therapy, 2) recommend surgery, or 3) be like ‘welp, I hope it heals on its own, that'll be $1000.’ They may also prescribe strong painkillers, but that feels more like a cop-out than an actual intervention to me, since it's not striking at the root of the problem.

Plausibly there is something more useful that doctors sometimes do, but beyond ruling out insidious diseases I have really not gotten a lot of value out of going to the doctor for this problem. To be fair, the modern medical establishment is much better than any alternative at curing viral and bacterial infections, treating life-threatening injuries, and keeping at bay serious diseases such as cancer and autoimmune disorders; chronic pain just isn't currently in its wheelhouse.

Physical therapy

As far as I can tell, physical therapy seems to be predicated on the notion that your pain is a result of you doing something wrong or your body being out of alignment. Thus, the way to fix pain is to find the thing you're doing wrong or the place where your body is stuck, and teach you exercises that will correct that. Given that my PT was surprised at my large range of motion and graceful-looking movement, my impression is that physical therapy is mostly targeted at (and useful for) people whose problems are more outwardly obvious than mine.

To me, the most interesting thing about physical therapy was that most of the exercises my PT recommended were not actually targeted at the areas of pain, but instead focused on improving my core strength.

A Guide to Better Movement has a similar thesis that the mechanics of movement are a key part of understanding pain, but focuses much more on the role of the nervous system. I haven't finished the book yet but it seems quite promising, and I'll try to update this when I have a better grasp on what it's about.

Massage therapy

The general idea of massage is that pain comes from tension in your muscles, and if you release the tension by working the muscles directly, the pain will go away. In an average (non-medical) massage, the massage therapist is likely to apply friction, percussion, and/or heat to your muscles. While this can be very effective at getting your muscles to relax, it's unlikely to lead to lasting changes in how you feel.

Myofascial massage

(EDIT 03/2019 - I was told all this information about fascia by a massage therapist, but I am now much less confident that it is correct.)

Fascia is a thing that surrounds muscles. It can be either a liquid or a solid, and when it solidifies it inhibits the range of motion in the area. I think this is supposedly what's happening when you keep a muscle shortened for too long and then it's later painful to move it (e.g. hunching at your desk or 'sleeping on your neck weird'). I've been told that myofascial massage involves working on solidified fascia to turn it back into liquid, which maybe makes sense. I don't know.

Trigger point massage

Trigger points are "discrete, focal, hyperirritable spots located in a taut band of skeletal muscle. They produce pain locally and in a referred pattern and often accompany chronic musculoskeletal disorders." Massage focused on trigger points can release the tension in them, which alleviates the referred pain from the trigger point. I think **acupuncture** operates on the same underlying principle, only instead of massaging the trigger point you stick a needle in it.

Here is a Youtube video explaining trigger points in more detail. Lots of people report near-magical success self-treating with trigger point massage, and it's often recommended for people who have RSI in their arms [1]. Disclaimer: Don't bank on magical results or even any results at all. Also note that trigger points may not be real.

Progressive muscle relaxation (PMR)

This isn't really a subcategory of massage, but it does operate on the same theory of muscle tension being the problem. In PMR, you systematically relax each of your muscle groups in turn, just by sitting there and thinking about it. It feels pretty good, and it can have somewhat deeper benefits if you practice enough to be able to do it quickly and as needed. Instructions can be found here, or lots of other places on the internet. Note that the first step, where you tense your muscles, is probably not actually necessary and might make you feel worse.

Dietary causes

A therapist I went to suggested that chronic joint pain is often a result of undiagnosed food sensitivities, such as intolerances to gluten, lactose, or sugar. She didn't expand on this at all but here's my guess at a mechanism: If you're consuming something to which you have an intolerance or low-level allergy, your body is treating it as an invader and will launch an immune attack, which can inflame the joints and manifest as chronic pain.

This seems plausible to me as an explanation for some cases, and particularly for mine, since I have known intolerances to lactose and gluten. However, my guess is that in most cases this would only be a small piece of the puzzle, one of many underlying causes that interacts with a bunch of other things. But that's just a random guess, and nutrition is really poorly understood in general, so take this entire section with industrial quantities of salt.

Neuroscience (specifically the research of Dr. Irene Tracey, from this article)

According to the article, people who suffer from chronic pain have an overactive pain amplification mechanism in the brain stem:

> "Tracey's latest research has investigated a key neural mechanism of chronic pain. It is situated in the brain stem… which functions as the conduit for communication between the brain and the body. Experiments on animals had identified two mechanisms within the brain stem that, respectively, muffle and boost pain signals before they reach the rest of the brain… Unfortunately, in some people the mechanism that exacerbates pain is dominant. Scanning the brains of patients with diabetic nerve pain, Tracey and Segerdahl found enhanced communication from the brain stem, via the spine, to the parts of the brain known to contribute to the sensation of pain."

This is a pretty enticing explanation for why of two people with basically identical injuries, one may recover entirely after the injury heals while the other may experience a lifetime of chronic pain from the injury. I'd be interested in further exploration into the neurology of pain; my impression is that it's currently not a very large field.

.

Psychological explanations

Dr. John Sarno & Dr. Howard Schubiner

Sarno and Schubiners' philosophy is that pain is your body trying to distract you from stress/trauma, in what they call tension myositis syndrome or TMS. TMS can cause not only chronic pain, but also ulcers, nausea, dizziness, fatigue, weakness, stomach problems, ear disturbances, and other mysterious ailments.

Sarno says that if you just ask your body, “what is this pain trying to distract me from?”, the pain will disappear because it no longer has a reason to exist. How does this work? I… don’t know. The book doesn’t really explain that. It seems possible that it's... some kind of really strong placebo effect? Or... I don't know? The Sarno method sure seems a lot like Dark Arts, but that doesn’t explain away the fact that it works, at least for some people. (If you don’t believe Sarno’s claim that he’s cured over ten thousand people, maybe you'll believe the 400+ five-star reviews on Amazon from people saying they were cured forever.)

Sarno also advises that - since chronic pain is psychological - sufferers won't get better if they continue to pursue physical solutions for their pain, such as massage or physical therapy. This is a pretty dangerous claim to take at face value and could very well make things a lot worse, so if you read either *The Mindbody Prescription* or *Unlearn Your Pain*, be careful.

Note that **Scott Alexander** points out that Sarno’s claim that Lithuanians rarely develop chronic whiplash following a neck injury is probably not true. This is a fairly major strike against Sarno and Schubiner’s case that the reign of pain lies mainly in the brain.

Somatic therapy

Similar to Sarno, somatic therapy operates on the model that pain is your body expressing stress/trauma, but unlike the Sarno method, it's primarily marketed as a form of psychotherapy. For 'holistic healing', somatic therapy addresses the underlying trauma while also working with the body. The internet tells me that "therapy sessions typically involves the patient tracking his or her experience of sensations throughout the body. Depending on the form of somatic psychology used, sessions may include awareness of bodily sensations, dance, breathing techniques, voice work, physical exercise, movement and healing touch." It sounds sort of similar to Focusing?

Hansonian model

Robin Hanson's signaling theory would probably say that chronic pain is you subconsciously signaling helplessness, which helps you get resources without having to do much work. Out of all the things I've discussed this one is definitely on the shakiest ground, since it just comes from moridinamael's speculations.

However, it feels pretty intuitively plausible to me, given that I often intentionally signal helplessness in other ways, and given that chronic pain is a great excuse for not being able to do stuff. When I was leading a dance group in high school and there was a lot of pressure on me to be perfect, my knee pain gave me an excuse for sometimes messing up. Now my carpal tunnel gives me an excuse for not always meeting my work goals, and more generally, pain gives me an excuse to sometimes lie in bed all day and whine and cry, which is not something it is normally acceptable to do as an adult.

My attempt at a synthesis

A first attempt

Your body is a very delicate machine, and if you do anything incorrectly - e.g. typing with bad posture, putting your weight on the wrong part of your foot when you walk, eating foods you're sensitive to, having breasts that weigh too much, or compensating for a weak core - it can mess the whole thing up in ways that aren't obviously connected to the thing you're doing wrong.

Your physiological responses to psychological stressors (such as trauma, anxiety, or feelings of helplessness) may then exacerbate the slight imperfections caused by whatever you're doing wrong, creating trigger points and other persistent pain.

Whether or not these problems (which have both physical and psychological causes) rise to the level of conscious attention depends on the activity of the pain amplification mechanism in your brain stem. There can then be a feedback loop, in which your brain learns that certain things are painful, and then is ever more careful about not injuring them, until any movement of the affected area (or even lying still) is agony. At this point, you have developed a chronic pain problem.

Missing pieces

Physical-psychological cycle

I feel that my model is missing the two-way interaction between the physical and psychological stressors. For example, if I feel vulnerable or defensive, I am likely to hunch my shoulders and generally tense up, which messes up my body. This two-way thing would help to explain why both psychological-only and physical-only interventions can fully cure some people's pain, since the cycle can be broken no matter where you intervene.

Permanent change

I think lasting cures probably all involve learning how to respond to stressors in a way that won't trigger a pain spiral. Under this model, too much reliance on external help that you don't understand (e.g. a massage therapist who 'works magic') will prevent you from solving the problem, even if it helps in the short to medium term.

The nervous system

I just have this feeling that Todd Hargrove knows what he's talking about. I will try to remember to update this when I've finished reading *A Guide to Better Movement.*

Mysteries

Surgery

I'm confused about the role of surgery for chronic pain. Sarno claims that any benefits from back surgery are basically placebo, and says some pretty convincing things on that point. I also have a vague sense that surgery is over-prescribed in the US just because doctors don't have that many good solutions, and also because it makes the medical establishment money. I know at least one person who reports benefits from surgery for his RSI, but in most cases of RSI I don't think there's anything obvious to operate on, and RSI is quite clearly more about use than about underlying problems.

Brains

There is probably more than just the one neurological mechanism at work in chronic pain. In particular, differences in how brains process sensory input seem pretty important to me. My boyfriend has terrible typing posture and is obviously very tense (it's very painful for him when I press lightly on his shoulders), but he basically doesn't complain of pain at all. This is consistent with his tendency to get so focused on whatever he's doing that he doesn't notice hunger, fatigue, or people talking directly to him. I, on the other hand, notice every slight discomfort when I'm sitting at my desk, from the increasing tension in my neck to the gait of the person walking behind me to the tiniest amount of hunger, thirst, or emotional turmoil. In a nutshell, his processing is top-down, while mine is bottom-up.

???

There are a dozen other mysteries here.

Conclusion

I don't think my model is very good, but I've already put more time into this than I wanted to so I'm not going to change it right now. If you, reading this post, had a different idea for how the pieces might fit together, tell me about it in the comments. If there are some theories I listed that you think should just be thrown out entirely, tell me why.

Thanks for reading all the way to the end!! :)

--

[1] If you're interested in self-treatment with the trigger point method, the canonical text is *The Trigger Point Therapy Workbook*. Since you'll also need to buy at least one tool to help you execute the techniques, expect to spend ~$50. If you're more interested in the theory, try this ebook.

Human-Aligned AI Summer School: A Summary

*(**Disclaimer:** this summary is incomplete and does not accurately represent all the content presented at the summer school, but only what I remember and seem to have understood from the lectures. Don't hesitate to mention important ideas I missed or apparent confusion.)*

Last week, I attended the first edition of the human-aligned AI summer school in Prague. After three days, my memories are already starting to fade, and I am unsure about what I will retain in the long-term.

Here, I try to remember the content of about 15h of talks. It serves the following purposes:

- **To the general audience that did not attend the school**, I try to give an overview, to inform about the general trends we discussed.
- **For those who attended the school**, I distill what I understood, to refresh our memories.

Value Learning (Daniel Filan)

Value Learning aims at infering human values from their behavior. Paul Christiano distinguishes ambitious value learning vs. narrow value learning:

- **Ambitious value learning**: learn human preferences over long-term outcomes.
- **Narrow value learning**: learn human instrumental values and subgoals.

Inverse Reinforcement Learning

Inverse Reinforcement Learning (IRL) studies which reward *best* explains a behaviour. Two methods of IRL were discussed (the state-of-the-art builds on top of those two, for instance using neural networks):

- Bayesian IRL: uses Bayesian update. Does not work in practice because necessits to solve many Markov Decision Processes, which is computationally intensive.
- Maximum Entropy IRL: the optimal distribution (of maximum entropy) is an exponential of a linear function. One of the reason it performs better in practice is that it's easier to efficiently approximate the relevant integrals.

Why not to do value learning **:**

- It is (still) inefficient
- It depends heavily on human rationality models
- The reward might not be in the prior reward space
- Solving other problems, such as naturalized agency, might be more urgent
- The actions in the behavior are not well-defined in practice (e.g. what counts as an action in a football game?)

Beyond Inverse Reinforcement Learning

The main problem of traditional IRL is that it does not take into account the deliberate interactions between a human and an AI (e.g. the human could be slowing down his behaviour to help learning).

Cooperative IRL solves this issue by introducing a two-player game between the human and the AI, where both are rewarded according to the human's reward function. This incentivizes the human to teach the AI his preferences (if the human only choses its best action, the AI would learn the wrong distribution). Using a similar dynamic, the off-switch game encourages the AI to allow himself to be switched off.

Another adversity when implementing IRL is that the reward function is difficult to completely specify, and will often not capture all of what the designer wants. Inverse reward design makes the AI quantify his uncertainty about states. If the AI is risk-averse, it will avoid uncertain states, for instance situations where it believes humans have not completely defined the reward function because they did not know much about it.

Agent Foundations (Abram Demski)

Abram's first talk was about his post "Probability is Real, and Value is Complex". At the end of the talk, several people (including me) were confused about the "magic correlation" between probabilities and expected utility, and asked Abram about the meaning of his talk.

From what I understood, the point was to show a counter-intuitive consequence of choosing Jeffrey-Bolker axioms in decision theory over Savage axioms. Because Bayes' algorithm can be formalized using Jeffrey-Bolker axioms, this counter-intuitive result challenges potential agent designs that would use Bayesian updates.

The second talk was more general, and addressed several problems faced by embedded agents (e.g. naturalized induction).

Bounded Rationality (Daniel Filan / Daniel Braun)

To make sure an AI would be able to understand humans, we need to make sure it understands their *bounded rationality*, i.e. how sparse information and a bounded computational power limit rationality.

Information-Theoretic Bounded Rationality (Daniel Braun)

The first talk on the topic introduced a decision-complexity C(A|B) that expressed the "cost" of going from the reference B to the target A (proportional to the Shannon Information of A given B). Intuitively, it represents the cost in search process when

going from a prior B to a posterior A. After some mathematical manipulations, a concept of "information cost" is introduced, and the final framework highlights a trade-off between some "information utility" and this "information cost" (for more details see here, pp. 14-18).

Human irrationality in planning (Daniel Filan)

Humans seem to exhibit a strong preference in planning hierarchically, and are "irrational" in that sense, or at least not "Boltzmann-rational" (Cundy & Filan, 2018).

Hierarchical RL is a framework used in planning that introduces "options" in Markov Decision Processes where Bellman Equations still hold.

State-of-the-art methods in Hierarchical RL include meta-learning of the hierarchy or a two-modules neural network.

Side effects (Victoria Krakovna)

Techniques aiming at minimizing negative side effects include minimizing *unnecessary disruptions* when achieving a goal (e.g. turning Earth into paperclips) or designing low-impact agents (avoiding large side effects in general).

To correctly measure impact, several questions must be answered:

- How is *change* defined?
- What was actually *caused* by the agent?
- What was really *necessary* to achieve the objective?
- What are the *implicit consequences* of the objective (e.g. a longer life expectancy after "curing cancer")?

A "side-effect measure" should penalize unnecessary actions (necessity), understand what was caused by the agent vs. caused by the environment (causation) and penalize irreversible actions (asymmetry).

Hence, an agent may be penalized for an outcome different from an "inaction baseline" (where the agent would not have done anything) or for any irreversible action.

However, those penalties introduce *bad incentives* to avoid irreversible actions but still let them happen anyway (for instance preventing a vase to be broken to gain a reward, then break the vase anyway to go back to the "inaction baseline"). Relative reachability provides an answer to this behaviour, by penalizing the agent for making states less reachable than there would be by default (for instance breaking a vase makes the states with an unbroken vase unreachable) and leads to safe behaviors in the Sokoban-like and conveyor belt gridworlds.

Open questions about this approach are:

- How exactly should we compute the "inaction baseline" or the "default state"?
- How well could it work with AGI?

I thank Daniel Filan and Jaime Molina for their feedback, and apologize for the talks I did not summarize.

Turning Up the Heat: Insights from Tao's 'Analysis II'

Foreword

It's been too long - a month and a half since my last review, and about three months since *Analysis I*. I've been immersed in my work for CHAI, but reality doesn't grade on a curve, and I want more mathematical firepower.

On the other hand, I've been cooking up something really special, so watch this space!

Analysis II

12: Metric Spaces

Metric spaces; completeness and compactness.

Proving Completeness

It sucks, and I hate it.

13: Continuous Functions on Metric Spaces

Generalized continuity, and how it interacts with the considerations introduced in the previous chapter. Also, a terrible introduction to topology.

There's a lot I wanted to say here about topology, but I don't think my understanding is good enough to break things down - I'll have to read an actual book on the subject.

14: Uniform Convergence

Pointwise and uniform convergence, the Weierstrass M *-test, and uniform approximation by polynomials.*

Breaking Point

Suppose we have some sequence of functions $f^{(n)} : [0,1] \to \mathbb{R}$, $f^{(n)}(x) := x^n$, which converge pointwise to the 1-indicator function $f : [0,1] \to \mathbb{R}$ (*i.e.*, $f(1) = 1$ and 0 otherwise). Clearly, each $f^{(n)}$ is (infinitely) differentiable; however, the limiting function

f isn't differentiable at all! Basically, pointwise convergence isn't at all strong enough to stop the limit from "snapping" the continuity of its constituent functions.

Progress

As in previous posts, I mark my progression by sharing a result derived without outside help.

Already proven: $\int_{-1}^{1}(1-x^2)^N\,dx \geq \frac{1}{\sqrt{N}}$.

Definition. Let $\epsilon > 0$ and $0 < \delta < 1$. A function $f : R \to R$ is said to be an (ϵ, δ)-*approximation to the identity* if it obeys the following three properties:

- f is compactly supported on $[-1, 1]$.
- f is continuous, and $\int_{-\infty}^{\infty} f = 1$.
- $|f(x)| \leq \epsilon$ for all $\delta \leq |x| \leq 1$.

Lemma: For every $\epsilon > 0$ and $0 < \delta < 1$, there exists an (ϵ, δ)-approximation to the identity which is a polynomial P on $[-1, 1]$.

Proof of Exercise 14.8.2(c). Suppose $c \in R, N \in N$; define $f(x) := c(1-x^2)^N$ for $x \in [-1, 1]$ and 0 otherwise. Clearly, f is compactly supported on $[-1, 1]$ and is continuous. We want to find c, N such that the second and third properties are satisfied. Since $(1-x^2)^N$ is non-negative on $[-1, 1]$, c must be positive, as f must integrate to 1. Therefore, f is non-negative.

We want to show that $|c(1-x^2)^N| \leq \epsilon$ for all $\delta \leq |x| \leq 1$. Since f is non-negative, we may simplify to $(1-x^2)^N \leq \epsilon$. Since the left-hand side is strictly monotone increasing on $[-1, -\delta]$ and strictly monotone decreasing on $[\delta, 1]$, we substitute $x = \delta$ without loss of generality. As $\epsilon > 0$, so we may take the reciprocal and multiply by ϵ, arriving at $\epsilon(1-\delta^2)^{-N} \geq c$.

We want $\int_{-\infty}^{\infty} f = 1$; as f is compactly supported on $[-1, 1]$, this is equivalent to $\int_{-1}^{1} f(x)\,dx = 1$. Using basic properties of the Riemann integral, we have $\int_{-1}^{1} (1 - x^2)^N\,dx = \frac{1}{c}$. Substituting in for c,

$$\epsilon^{-1}(1-\delta^2)^N \leq \frac{1}{\sqrt{N}} \leq \int_{-1}^{1} (1-x^2)^N\,dx,$$

with the second inequality already having been proven earlier. Note that although the first inequality is not always true, we can make it so: since ϵ is fixed and $1 - \delta^2 \in (0, 1)$, the left-hand side approaches 0 more quickly than $\frac{1}{\sqrt{N}}$ does. Therefore, we can make N as large as necessary; isolating ϵ,

$$\epsilon \geq (1-\delta^2)^N \sqrt{N}$$

$$\epsilon \geq \sqrt{N} > (1-\delta^2)^N \sqrt{N},$$

the second line being a consequence of $1 > (1 - \delta^2)^N$. Then set N to be any natural number such that this inequality is satisfied. Finally, we set $c = \frac{1}{\int_{-1}^{1}(1-x^2)^N\,dx}$. By construction, these values of c, N satisfy the second and third properties. □

Convoluted No Longer

Those looking for an excellent explanation of convolutions, look no further!

Weierstrass Approximation Theorem

Theorem. Suppose $f : [a, b] \to R$ is continuous and compactly supported on $[a, b]$. Then for every $\epsilon > 0$, there exists a polynomial P such that $||P - f||_\infty < \epsilon$.

In other words, any continuous, real-valued f on a finite interval can be approximated with arbitrary precision by polynomials.

Why I'm talking about this. On one hand, this result makes sense, especially after taking machine learning and seeing how polynomials can be contorted into basically whatever shape you want.

On the other hand, I find this theorem intensely beautiful. $\overline{\overline{P[a,b]} = C[a,b]}$'s proof was slowly constructed, much to the reader's benefit. I remember the very moment the proof sketch came to me, newly-installed gears whirring happily.

15: Power Series

Real analytic functions, Abel's theorem, exp *and* log*, complex numbers, and trigonometric functions.*

EXP

Cached thought from my CS undergrad: exponential functions always end up growing more quickly than polynomials, no matter the degree. Now, I finally have the gears to see why:

$$\exp(x) := \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

exp has *all* the degrees, so no polynomial (of necessarily finite degree) could ever hope to compete! This also suggests why $\frac{d}{dx} e^x = e^x$.

Complex Exponentiation

You can multiply a number by itself some number of times.

[*nods*]

You can multiply a number by itself a negative number of times.

[Sure.]

You can multiply a number by itself an irrational number of times.

[OK, I understand limits.]

You can multiply a number by itself an imaginary number of times.

[Out. Now.]

Seriously, this one's weird (rather, it *seems* weird, but how can "how the world is" be "weird")?

Suppose we have some $c \in C$, where $c = a + bi$. Then $e^c = e^a e^{bi}$, so "all" we need to figure out is how to take an imaginary exponent. Brian Slesinsky has us covered.

Years before becoming involved with the rationalist community, Nate asks this question, and Qiaochu answers.

This isn't a coincidence, because nothing is ever a coincidence.

Or maybe it is a coincidence, because Qiaochu answered every question on StackExchange.

16: Fourier Series

Periodic functions, trigonometric polynomials, periodic convolutions, and the Fourier theorem.

17: Several Variable Differential Calculus

A beautiful unification of Linear Algebra and calculus: linear maps as derivatives of multivariate functions, partial and directional derivatives, Clairaut's theorem, contractions and fixed points, and the inverse and implicit function theorems.

Implicit Function Theorem

If you have a set of points in R^n, when do you know if it's secretly a function

$g : R^{n-1} \to R$? For functions $R \to R$, we can just use the geometric "vertical line test" to figure this out, but that's a bit harder when you only have an algebraic definition. Also, sometimes we can implicitly define a function locally by restricting its domain (even if no explicit form exists for the whole set).

Theorem. Let E be an open subset of R^n, let $f : E \to R$ be continuously differentiable,

and let $y = (y_1, \dots, y_n)$ be a point in E such that $f(y) = 0$ and $\frac{\partial f}{\partial x_n} \neq 0$. Then there exists

an open $U \subseteq R^{n-1}$ containing $(y_1, \dots, y_{n-1})$, an open $V \subseteq E$ containing y, and a function

$g : U \to R$ such that $g(y_1, \dots, y_{n-1}) = y_n$, and

$$\{ (x_1 , \dots , x_n) \in V : f (x_1 , \dots , x_n) = 0 \} =$$

$$\{ (x_1 , \dots , x_{n-1} , g (x_1 , \dots , x_{n-1})) : (x_1 , \dots , x_{n-1}) \in U \} .$$

So, I think what's really going on here is that we're using the derivative at this known zero to locally linearize the manifold we're operating on (similar to Newton's

approximation), which lets us have some neighborhood U in which we can derive an implicit function, even if we can't always write it out.

18: Lebesgue Measure

Outer measure; measurable sets and functions.

Tao lists desiderata for an ideal measure before deriving it. Imagine that.

19: Lebesgue Integration

Building up the Lebesgue integral, culminating with Fubini's theorem.

Conceptual Rotation

Suppose $\Omega \subseteq R^n$ is measurable, and let $f : \Omega \to [0, \infty]$ be a measurable, non-negative function. The Lebesgue integral of f is then defined as

$$\int_\Omega f := \sup \{ \int_\Omega s : s \text{ is simple and non-negative, and minorizes } f \} .$$

This hews closely to how we defined the *lower* Riemann integral in Chapter 11; however, we don't need the equivalent of the upper Riemann integral for the Lebesgue integral.

To see why, let's review why Riemann integrability demands the equality of the lower and upper Riemann integrals of a function g. Suppose that we integrate over $[0, 1]$, and that g is the indicator function for the rationals. As the rationals are dense in the reals, any interval $[a, b] \subseteq [0, 1]$ ($b > a$) contains rational numbers, no matter how much the interval shrinks! Therefore, the upper Riemann integral equals 1, while the lower equals 0 (for similar reasons). g *is* Lebesgue integrable; since it's 0 almost everywhere (as the rationals have 0 measure), its integral is 0.

This marks a fundamental shift in how we integrate. With the Riemann integral, we consider the lim sup and lim inf of increasingly-refined upper and lower Riemann sums - this is the *length* approach. In Lebesgue integration, however, we consider which $E \subseteq \Omega$ is responsible for each value y in the range (*i.e.*, $f^{-1}(y) = E$), multiplying y by the measure of E - this is *inversion*.

In a sense, the Lebesgue integral more cleanly strikes at the heart of what it *means* to integrate. Surely, Riemann integration was not far from the mark; however, if you

rotate the problem slightly in your mind, you will find a better, cleaner way of structuring your thinking.

Final Thoughts

Although Tao botches a few exercises and the section on topology, I'm a big fan of *Analysis I* and *II*. Do note, however, that *II* is far more difficult than *I* (not just in content, but in terms of the exercises). He generally provides relevant, appropriately-difficult problems, and is quite adept at helping the reader develop rigorous and intuitive understanding of the material.

Forwards

Next is Jaynes' *Probability Theory*.

Tips

- To avoid getting hung up in Chapter 17, this book should be read after a linear algebra text.
- Don't do exercise 17.6.3 - it's wrong.
- Deep understanding comes from sweating it out. Don't hide, don't wave away bothersome details - stay and explore. If you follow my strategy of quickly generating outlines - can you formally and precisely write out each step?

Verification

I completed every exercise in this book; in the second half, I started avoiding looking at the hints provided by problems until I'd already thought for a few minutes. Often, I'd solve the problem and then turn to the hint: "be careful when doing *X* - don't forget edge case *Y*; hint: use lemma *Z*"! A pit would form in my stomach as I prepared to locate my mistake and back-propagate where-I-should-have-looked, before realizing that I'd *already* taken care of that edge case using that lemma.

Why Bother?

One can argue that my time would be better spent picking up things as I work on problems in alignment. However, while I've made, uh, quite a bit of progress with impact measures this way, concept-shaped holes are impossible to notice. If there's some helpful information-theoretic way of viewing a problem that I'd only realize if I had *already taken* information theory, I'm out of luck.

Also, developing mathematical maturity brings with it a more rigorous thought process.

Fairness

There's a sense I get where even though I've made immense progress over the past few months, it still *might not be enough*. The standard isn't "am I doing impressive things for my reference class?", but rather the stricter "am I good enough to solve serious problems that might not get solved in time otherwise?". This is quite the standard, and even given my textbook and research progress (including the upcoming posts), I don't think I meet it.

In a way, this excites me. I welcome any advice for buckling down further and becoming yet stronger.

If you are interested in working with me or others on the task of learning MIRI-relevant math, if you have a burning desire to knock the alignment problem down a peg - I would be more than happy to work with you. Messaging me may also net you an invitation to the MIRIx Discord server.

On a related note: thank you to everyone who has helped me; in particular, TheMajor has been incredibly generous with their explanations and encouragement.

You Play to Win the Game

Previously (Putanumonit): Player of Games

Original Words of Wisdom:

Quite right, sir. Quite right.

By far the most important house rule I have for playing games is exactly that: **You Play to Win the Game.**

That doesn't mean you always have to take exactly the path that maximizes your probability of winning. Style points can be a thing. Experimentation can be a thing. But in the end, you play to win the game. If you don't think it matters, do as Herm Edwards implores us: Retire.

It's easy to forget, sometimes, what 'the game' actually is, in context.

The most common and important mistake is to *maximize expected points or point differential,* at the cost of win probability. Alpha Go brought us many innovations, but perhaps its most impressive is its willingness to sacrifice territory it doesn't need to minimize the chances that something will go wrong. Thus it often wins by the narrowest of *point* margins, but in ways that are very secure.

The larger-context version of this error is to maximize *winning or points in the round* rather than *chance of winning the event.*

In any context where points are added up over the course of an event, the game that matters is *the entire event.* You do play to win each round, to win each point, but strategically. You're there to hoist the trophy.

Thus, when we face a game theory experiment like Jacob faced in Player of Games, we have to understand that we'll face a variety of opponents with a variety of goals and methods. We'll play a prisoner's dilemma with them, or an iterated prisoner's dilemma, or a guess-the-average game.

To win, one must outscore every other player. Our goal is *to win the game.*

Unless or until it isn't. Jacob explicitly wasn't trying to win at least one of the games by scoring the most points, instead choosing to win the greater game of life itself, or at least a larger subgame. This became especially clear once winning was beyond his reach. At that point, the game becomes something odd – you're scoring points that don't matter. It's not much of a contest, and it doesn't teach you much about game theory or decision theory.

It teaches you other things about human nature, instead.

A key insight is what happens when a prize is offered for the *most successful* player of one-shot prisoner's dilemmas, or a series of iterated prisoner's dilemmas.

If you cooperate, *you cannot win.* Period. Someone else will defect while their opponents cooperate. Maybe they'll collude with their significant other. Maybe they'll lie convincingly. Maybe they'll bribe with out-of-game currency. Maybe they'll just get

lucky and face several variations on ‘cooperate bot’. Regardless of how legitimate you think those tactics are, with enough opponents, one of them will happen.

That means the *only* way to win is to defect and convince opponents to cooperate. Playing any other way means *playing a different game.*

When scoring points, make sure the points matter.

These issues will also be key to the next post as well, where we will analyze a trading board game proposed by Robin Hanson.

Logical Counterfactuals & the Cooperation Game

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

Logical counterfactuals (as in Functional Decision Theory) are more about your state of knowledge than the actual physical state of the universe. I will illustrate this with a relatively simple example.

Suppose there are two players in a game where each can choose A or B with the payoffs as follows for the given combinations:

AA: 10, 10

BB: 0, 0

AB or BA: -10, -10

Situation 1:

Suppose you are told that you will make the same decision as the other player. You can quickly conclude that A provides the highest utility.

Situation 2:

Suppose you are told that the other player chooses A. You then reason that A provides the highest utility

Generalised Situation: This situation combines elements of the previous two. Player 1 is an agent that will choose A, although this is not known by Player 2 unless option b) in the next sentence is true. Player 2 is told one of the following:

a) They will inevitably make the same decision as Player 1

b) Player 1 definitely will choose A

If Player 2 is a rational timeless agent, then they will choose A regardless of which one they are told. This means that both agents will choose A, making both a) and b) true statements.

Analysis:

Consider the Generalised Situation, where you are Player 2. Comparing the two cases, we can see that the physical situation is identical, apart from the information you (Player 2) are told. Even the information Player 1 is told is identical. But in one situation we model Player 1's decision as counterfactually varying with yours, while in the other situation, Player 1 is treated as a fixed element of the universe.

On the other hand, if you were told that the other player would choose A and that they would make the same choice as you, then the only choice compatible with that would be to choose A. We could easily end up in all kinds of tangles trying to figure out the logical counterfactuals for this situation. However, the decision problem is really just

trivial in this case and the only (non-strict) counterfactual is what actually happened. There is simply no need to attempt to figure out logical counterfactuals given perfect knowledge of a situation.

It is a mistake to focus too much on the world itself as given precisely what happened all (strict) counterfactuals are impossible . The only thing that is possible is what actually happened. This is why we need to focus on your state of knowledge instead.

Resources:

A useful level distinction: A more abstract argument that logical counterfactuals are about mutations of your model rather than an attempt to imagine an external inconsistent universe.

What a reduction of "could" could look like: A conception of "could" in terms of what the agent can prove

Reducing collective rationality to individual optimization in common-payoff games using MCMC: Contains a similar game

Computational complexity of RL with traps

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

I briefly describe an open (to the best of my knowledge) problem and link to cstheory.stackexchange questions with more details. The readers are invited to try solving those questions.

In the comment section of my research agenda essay, Jessica and I discussed, among other things, whether intelligence should be defined by some sort of diminishing regret condition or just by Bayes-optimality. In particular, the former approach runs into a problem when traps are allowed, whereas the latter is still perfectly well-defined.

However, one problem with Bayes-optimality is its computational complexity. Without traps, Bayes-optimality is feasible in the weak sense that there is asymptotically Bayes-optimal algorithm (e.g. PSRL: Posterior Sampling Reinforcement Learning) that runs in time polynomial in the size of the problem, if the prior is described by an explicit list of MDP transition kernels and reward functions (so, in practice it's only feasible for a small number of small MDPs; there are also algorithms that can work for a large number of small MDPs). Hopefully this weak feasibility can be boosted into something stronger using additional assumptions about the prior/environment (e.g. hierarchical structure; presumably, deep learning somehow exploits such assumptions). However, when traps are allowed, it's not clear that even weak feasibility is achievable. In particular, PSRL is no longer asymptotically Bayes-optimal, since it is more or less guaranteed to walk into any trap given enough time.

I condensed the problem into two succinct questions on cstheory.stackexchange (but received no answers so far). In the stochastic case, there are some computational hardness results in the literature, but they are not sufficiently strong to rule out an asymptotically Bayes-optimal algorithm (they only rule out an *exactly* Bayes-optimal algorithm). In the deterministic case, there are no hardness results that I know of. So, currently I would not be very surprised neither by a positive nor a negative answer to any of those questions.

EDIT: This answer resolves the question, more or less. Even in the deterministic case, the problem is NP-hard, with a bound of at least $\Omega\left(\frac{1}{\#\{\text{hypotheses}\}}\right)$ on the approximation accuracy.