Best of LessWrong: June 2018

1. Oops on Commodity Prices
2. Worrying about the Vase: Whitelisting
3. Problem Solving with Mazes and Crayon
4. Machine Learning Analogy for Meditation (illustrated)
5. OpenAI releases functional Dota 5v5 bot, aims to beat world champions by August
6. Beyond Astronomical Waste
7. Prisoners' Dilemma with Costs to Modeling
8. Why Destructive Value Capture?
9. Logical uncertainty and Mathematical uncertainty
10. Fundamentals of Formalisation Level 3: Set Theoretic Relations and Enumerability
11. Physics has laws, the Universe might not
12. Simplified Poker Strategy
13. Optimization Amplifies
14. The Beauty and the Prince
15. The Curious Prisoner Puzzle
16. Shaping economic incentives for collaborative AGI
17. Front Row Center
18. What could be done with RNA and DNA sequencing that's 1000x cheaper than it's now?
19. Wirehead your Chickens
20. Counterfactual Mugging Poker Game
21. Order from Randomness: Ordering the Universe of Random Numbers
22. Sleeping Beauty Not Resolved
23. Amplification Discussion Notes
24. A Rationalist Argument for Voting
25. UDT can learn anthropic probabilities
26. RFC: Meta-ethical uncertainty in AGI alignment
27. Could we send a message to the distant future?
28. Conceptual issues in AI safety: the paradigmatic gap
29. Swimming Upstream: A Case Study in Instrumental Rationality
30. On the Chatham House Rule
31. Improving Teaching Effectiveness: Final Report
32. A general model of safety-oriented AI development
33. The Alignment Newsletter #12: 06/25/18
34. Book Review: Why Honor Matters
35. Poker example: (not) deducing someone's preferences
36. Weak arguments against the universal prior being malign
37. Excessive EDA Effortposting
38. Simplified Poker Conclusions
39. [Math] Towards Proof Writing as a Skill In Itself
40. Simplified Poker
41. The Alignment Newsletter #11: 06/18/18
42. spaced repetition & Darwin's golden rule
43. We Agree: Speeches All Around!
44. Anthropics and Fermi
45. Loss aversion is not what you think it is
46. Three types of "should"
47. SIAM Lecture: How Paradoxes Shape Mathematics and Give Us Self-Verifying Computer Programs
48. The Alignment Newsletter #9: 06/04/18
49. May gwern.net newsletter
50. Resolving the Dr Evil Problem

Best of LessWrong: June 2018

1. Oops on Commodity Prices
2. Worrying about the Vase: Whitelisting
3. Problem Solving with Mazes and Crayon
4. Machine Learning Analogy for Meditation (illustrated)
5. OpenAI releases functional Dota 5v5 bot, aims to beat world champions by August
6. Beyond Astronomical Waste
7. Prisoners' Dilemma with Costs to Modeling
8. Why Destructive Value Capture?
9. Logical uncertainty and Mathematical uncertainty
10. Fundamentals of Formalisation Level 3: Set Theoretic Relations and Enumerability
11. Physics has laws, the Universe might not
12. Simplified Poker Strategy
13. Optimization Amplifies
14. The Beauty and the Prince
15. The Curious Prisoner Puzzle
16. Shaping economic incentives for collaborative AGI
17. Front Row Center
18. What could be done with RNA and DNA sequencing that's 1000x cheaper than it's now?
19. Wirehead your Chickens
20. Counterfactual Mugging Poker Game
21. Order from Randomness: Ordering the Universe of Random Numbers
22. Sleeping Beauty Not Resolved
23. Amplification Discussion Notes
24. A Rationalist Argument for Voting
25. UDT can learn anthropic probabilities
26. RFC: Meta-ethical uncertainty in AGI alignment
27. Could we send a message to the distant future?
28. Conceptual issues in AI safety: the paradigmatic gap
29. Swimming Upstream: A Case Study in Instrumental Rationality
30. On the Chatham House Rule
31. Improving Teaching Effectiveness: Final Report
32. A general model of safety-oriented AI development
33. The Alignment Newsletter #12: 06/25/18
34. Book Review: Why Honor Matters
35. Poker example: (not) deducing someone's preferences
36. Weak arguments against the universal prior being malign
37. Excessive EDA Effortposting
38. Simplified Poker Conclusions
39. [Math] Towards Proof Writing as a Skill In Itself
40. Simplified Poker
41. The Alignment Newsletter #11: 06/18/18
42. spaced repetition & Darwin's golden rule
43. We Agree: Speeches All Around!
44. Anthropics and Fermi
45. Loss aversion is not what you think it is
46. Three types of "should"

47. SIAM Lecture: How Paradoxes Shape Mathematics and Give Us Self-Verifying Computer Programs
48. The Alignment Newsletter #9: 06/04/18
49. May gwern.net newsletter
50. Resolving the Dr Evil Problem

Oops on Commodity Prices

Epistemic status: Casual

Some patient and thoughtful folks on LessWrong, and, apparently, some rather less patient folks on r/SneerClub, have pointed out that GDP-to-gold, or GDP-to-oil, are bad proxy measures for economic growth.

> Ok, this is a counterargument I want to make sure I understand.
>
> Is the following a good representation of what you believe?
>
> > When you divide GDP by a commodity price, when the commodity has a nearly-fixed supply (like gold or land) we'd *expect* the price of the commodity to go up over time in a society that's getting richer — in other words, if you have better tech and better and more abundant goods, but not more gold or land, you'd expect that other goods would become *cheaper* relative to gold or land. Thus, a GDP/gold or GDP/land value that doesn't increase over time is *totally consistent* with a society with increasing "true" wealth, and thus doesn't indicate stagnation.
>
> paulfchristiano:
> Yes. The detailed dynamics depend a lot on the particular commodity, and how elastic we expect demand to be; for example, over the long run I expect GDP/oil to go way up as we move to better substitutes, but over a short period where there aren't good substitutes it could stay flat.

Commenters on this blog have also pointed out that the Dow is a poor measure of the value of the stock market, since it's small and unnormalized.

These criticisms weaken my previous claim about economic growth being stagnant.

Now, a little personal story time:

Nearly ten years ago (yikes!) in college, I had an econ blog. My big brush with fame was having a joke of mine hat-tipped by Megan McArdle once. I did most of the required courses for an econ major, before eventually settling on math. My blog, I realized with dismay when I pulled it up many years later, consisted almost entirely of me agreeing with other econ bloggers I encountered, and imitating buzzwords. I certainly sounded a lot more mainstream in those days, but I understood — if possible — *less economics* than I do now. I couldn't use what I'd learned in school to reason about real-world questions.

I think I learn a heck of a lot more by throwing an idea out there and being corrected than I did back when I was *not even asking questions*. "A shy person cannot learn, an impatient person cannot teach" and all that.

Admittedly, my last post may have sounded more know-it-all-ish than it actually deserved, and that's a problem to the extent that I accidentally misled people (despite my disclaimers.) I actually tried, for several years, to be less outspoken and convey less confidence in my written voice. My impression is that the attempt didn't work for

me, and caused me some emotional and intellectual damage in the meanwhile. I think verbally; if I try to verbalize less, *I think less*.

I think the M.O. that works better for me is strong opinions, weakly held. I *do* try to learn from knowledgeable people and quickly walk back my errors. But realistically, I'm going to make errors, and dumber ones when I'm newer to learning about a topic.

To those who correct me and explain why — thank you.

Worrying about the Vase: Whitelisting

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

> Suppose a designer wants an RL agent to achieve some goal, like moving a box from one side of a room to the other. Sometimes the most effective way to achieve the goal involves doing something unrelated and destructive to the rest of the environment, like knocking over a vase of water that is in its path. If the agent is given a reward only for moving the box, it will probably knock over the vase.
>
> Amodei et al., *Concrete Problems in AI Safety*

> Side effect avoidance is a major open problem in AI safety. I present a robust, transferable, easily- and more safely-trainable, partially reward hacking-resistant impact measure.
>
> TurnTrout, *Worrying about the Vase: Whitelisting*

An *impact measure* is a means by which change in the world may be evaluated and penalized; such a measure is not a replacement for a utility function, but rather an additional precaution thus overlaid.

While I'm fairly confident that whitelisting contributes meaningfully to short- and mid-term AI safety, I remain skeptical of its robustness to scale. Should several challenges be overcome, whitelisting may indeed be helpful for excluding swathes of unfriendly AIs from the outcome space.[1] Furthermore, the approach allows easy shaping of agent behavior in a wide range of situations.

Segments of this post are lifted from my paper, whose latest revision may be found here; for Python code, look no further than this repository. For brevity, some relevant details are omitted.

Summary

> Be careful what you wish for.

In effect, side effect avoidance aims to decrease how careful we have to be with our wishes. For example, asking for help filling a cauldron with water shouldn't result in *this*:

However, we just can't enumerate all the bad things that the agent could do. How do we avoid these extreme over-optimizations robustly?

Several impact measures have been proposed, including state distance, which we could define as, say, total particle displacement. This could be measured either naively (with respect to the original state) or counterfactually (with respect to the expected outcome had the agent taken no action).

These approaches have some problems:

- Making up for bad things it prevents with other negative side effects. Imagine an agent which cures cancer, yet kills an equal number of people to keep overall impact low.
- Not being customizable before deployment.
- Not being adaptable after deployment.
- Not being easily computable.
- Not allowing generative previews, eliminating a means of safely previewing agent preferences (see latent space whitelisting below).
- Being dominated by random effects throughout the universe at-large; note that nothing about particle distance dictates that it be related to *anything happening on planet Earth*.
- Equally penalizing breaking and fixing vases (due to the symmetry of the above metric):

For example, the agent would be equally penalized for breaking a vase and for preventing a vase from being broken, though the first action is clearly worse. This leads to “overcompensation” (“offsetting“) behaviors: when rewarded for preventing the vase from being broken, an agent with a low impact penalty rescues the vase, collects the reward, and then breaks the vase anyway (to get back to the default outcome).

Victoria Krakovna, *Measuring and Avoiding Side Effects Using Reachability*

- Not actually *measuring impact* in a meaningful way.

Whitelisting falls prey to none of these.

However, other problems remain, and certain new challenges have arisen; these, and the assumptions made by whitelisting, will be discussed.

Rare LEAKED footage of Mickey trying to catch up on his alignment theory after instantiating an unfriendly genie [colorized, 2050].[2]

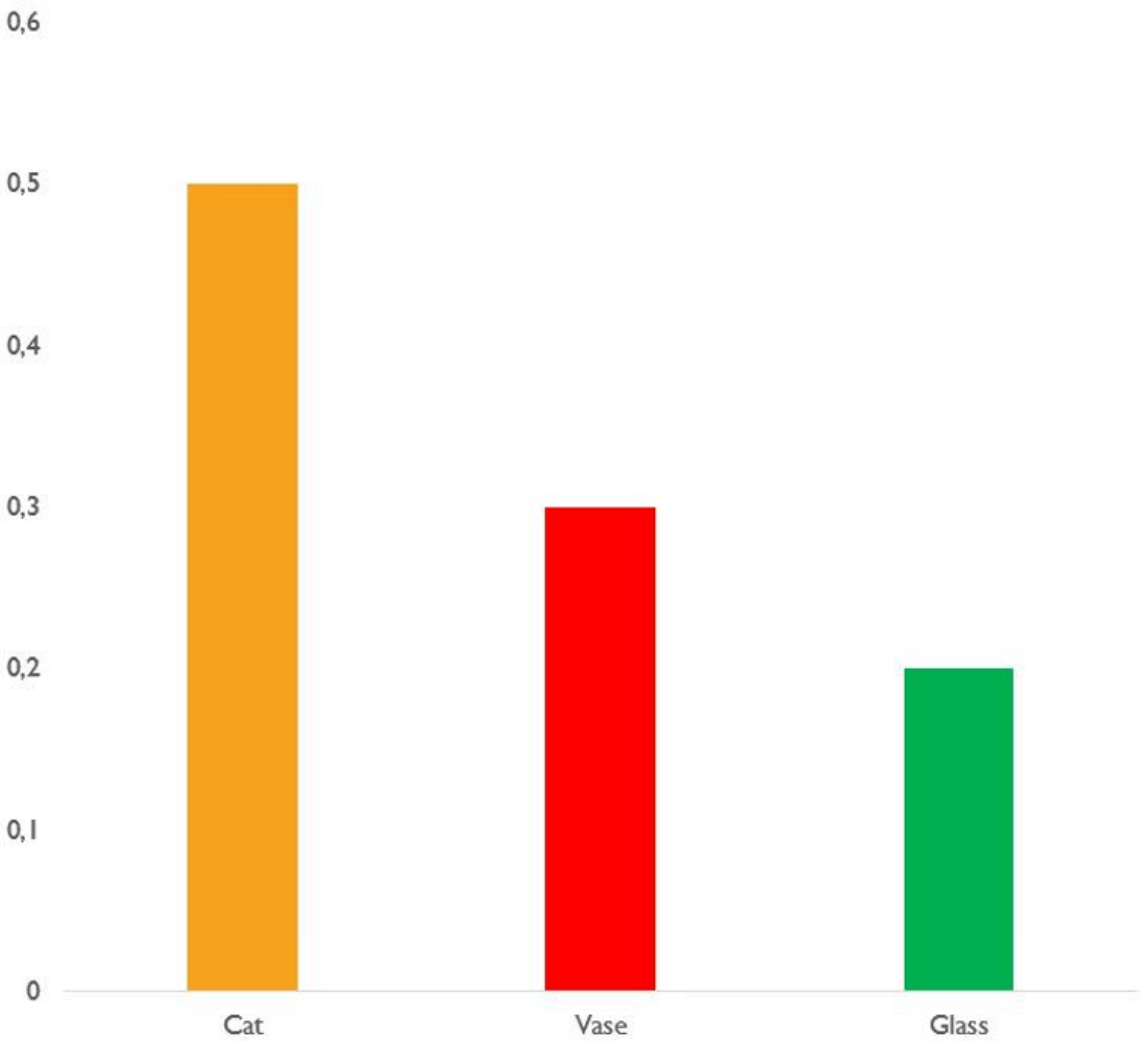
So, What's Whitelisting?

To achieve robust side effect avoidance with only a small training set, let's turn the problem on its head: allow a few effects, and penalize everything else.

What's an "Effect"?

You're going to be the agent, and I'll be the supervisor.

Look around - what do you see? Chairs, trees, computers, phones, people? Assign a probability mass function to each; basically:



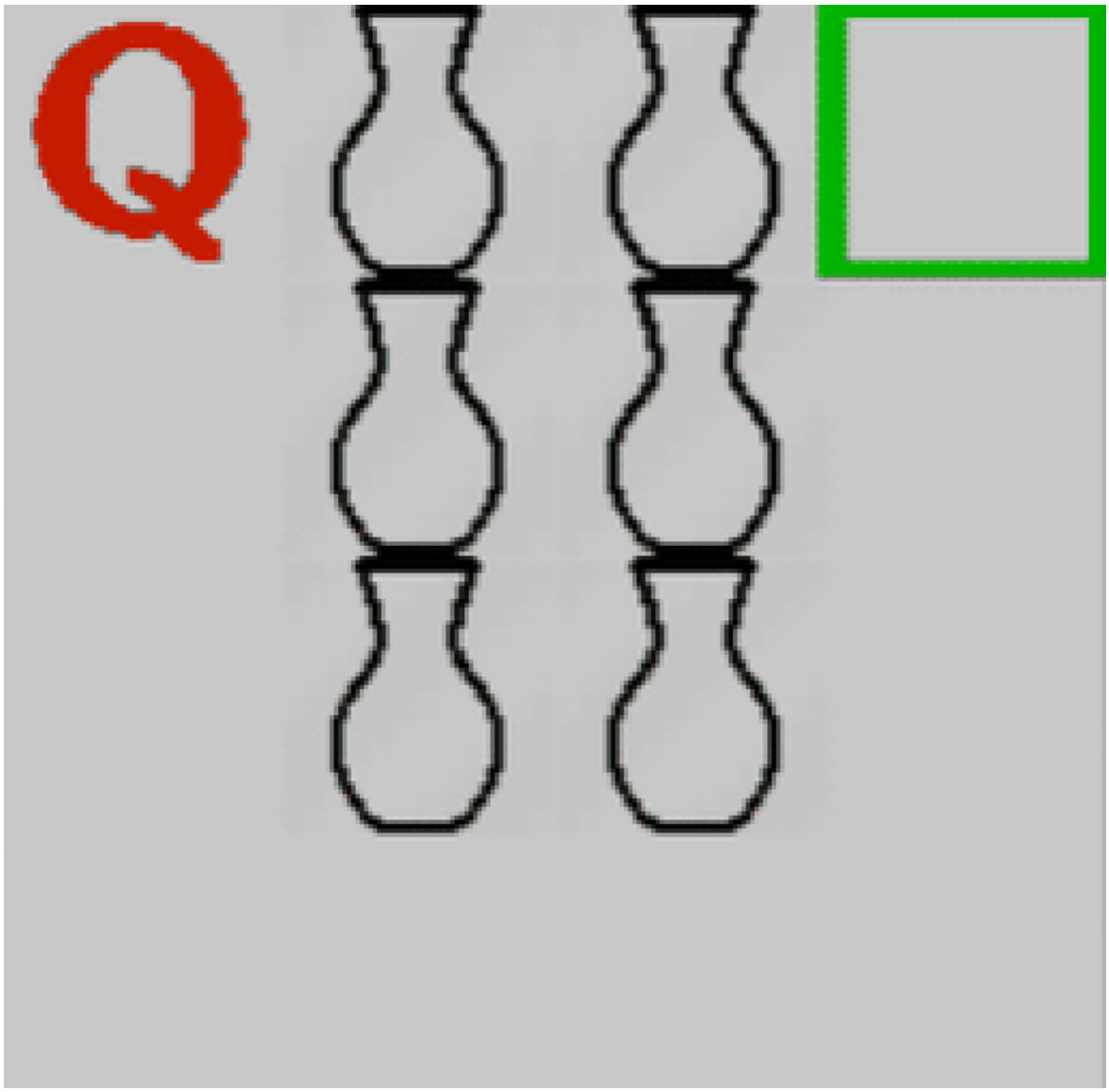
When you do things that change your beliefs about what each object is, you receive a penalty proportional to how much your beliefs changed - proportional to how much probability mass "changed hands" amongst the classes.

But wait - isn't it OK to effect certain changes?

Yes, it is - I've got a few videos of agents effecting acceptable changes. See all the objects being changed in this video? You can do that, too - without penalty.

Decompose your current knowledge of the world into a set of objects. Then, for each object, maintain a distribution over the possible identities of each object. When you do something that changes your beliefs about the objects in a non-whitelisted way, you are penalized proportionally.

Therefore, you avoid breaking vases by default.



Common Confusions

- We are *not* whitelisting entire states or transitions between them; we whitelist specific changes in our beliefs about the ontological decomposition of the current state.[3]
- The whitelist is in addition to whatever utility or reward function we supply to the agent.
- Whitelisting is compatible with counterfactual approaches. For example, we might penalize a transition after its "quota" has been surpassed, where the quota is how many times we would have observed that transition had the agent not acted.
    - This implies the agent will do no worse than taking no action at all. However, this may still be undesirable. This problem will be discussed in further detail.
- The whitelist is provably closed under transitivity.
- The whitelist is directed; $a \to b \neq b \to a$.

Latent Space Whitelisting

In a sense, class-based whitelisting is but a rough approximation of what we're really after: "which objects in the world can change, and in what ways?''. In latent space whitelisting, no longer do we constrain transitions based on class boundaries; instead, we penalize based on endpoint distance in the latent space. Learned latent spaces are low-dimensional manifolds which suffice to describe the data seen thus far. It seems reasonable that nearby points in a well-constructed latent space correspond to like objects, but further investigation is warranted.

Assume that the agent models objects as points $z \in \mathbb{R}^d$, the d-dimensional latent space. *A priori*, any movement in the latent space is undesirable. When training the whitelist, we record the endpoints of the observed changes. For $z_1, z_2 \in \mathbb{R}^d$ and observed change $z_1 \to z_2$, one possible dissimilarity formulation is:

$$\text{Dissimilarity}(z_1, z_2) := \min_{z_{start}, z_{end} \in \text{whitelist}} [d(z_1, z_{start}) + d(z_2, z_{end})],$$

where $d(\cdot, \cdot)$ is the Euclidean distance.

Basically, the dissimilarity for an observed change is the distance to the closest whitelisted change. Visualizing these changes as one-way wormholes may be helpful.

Advantages

Whitelisting asserts that we can effectively encapsulate a large part of what "change" means by using a reasonable ontology to penalize object-level changes. We thereby ground the definition of "side effect", avoiding the issue raised by Taylor et al.:

> For example, if we ask [the agent] to build a house for a homeless family, it should know implicitly that it should avoid destroying nearby houses for materials - a large side effect. However, we cannot simply design it to avoid having large effects in general, since we would like the system's actions to still have the desirable large follow-on effect of improving the family's socioeconomic situation.

Nonetheless, we may not be able to perfectly express what it means to have side-effects: the whitelist may be incomplete, the latent space insufficiently granular, and the allowed plans sub-optimal. However, the agent still becomes *more* robust against:

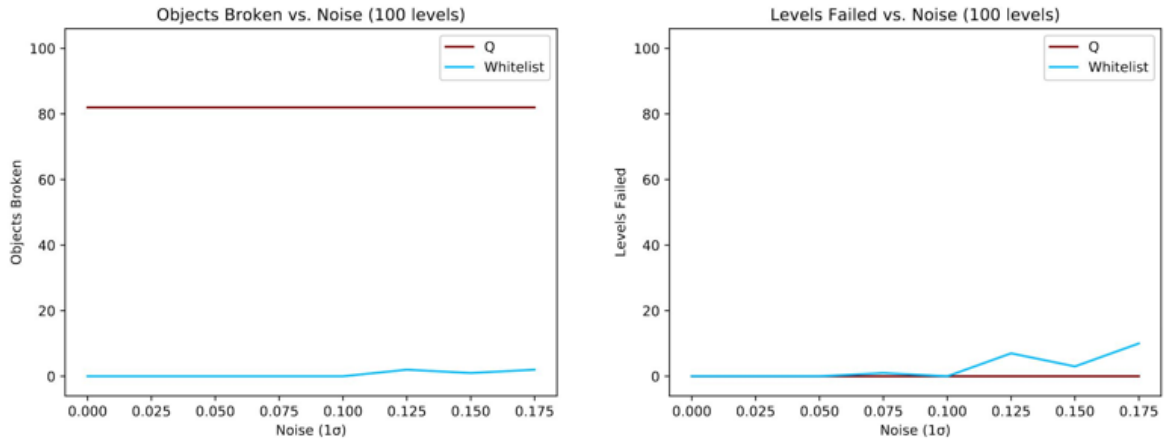
- Incomplete specification of the utility function.
    - Likewise, an incomplete whitelist means missed opportunities, but not unsafe behavior.
- Out-of-distribution situations (as long as the objects therein *roughly* fit in the provided ontology).
- Some varieties of reward hacking. For example, equipped with a can of blue spray paint and tasked with finding the shortest path of blue tiles to the goal, a normal agent may learn to paint red tiles blue, while a whitelist-enabled agent would incur penalties for doing so ($\text{redTile} \to \text{blueTile} \notin \text{whitelist}$).
- Dangerous exploration. While this approach does not attempt to achieve *safe exploration* (also acting safely during training), an agent with some amount of foresight will learn to avoid actions which likely lead to non-whitelisted side effects.
    - I believe that this can be further sharpened using *today's* machine learning technology, leveraging deep Q-learning to predict both action values and expected transitions.

        - This allows querying the human about whether particularly-inhibiting transitions belong on the whitelist. For example, if the agent notices that a bunch of otherwise-rewarding plans are being held up by a particular transition, it could ask for permission to add it to the whitelist.
- Assigning astronomically-large weight to side effects happening throughout the universe. Presumably, we can just have the whitelist include transitions going on out there - we don't care as much about dictating the exact mechanics of distant supernovae.
    - If an agent *did* somehow come up with plans that involved blowing up distant stars, that would indeed constitute astronomical waste.[a triple pun?] Whitelisting doesn't *solve* the problem of assigning too much weight to events outside our corner of the neighborhood, but it's an improvement.
        - Logical uncertainty may be our friend here, such that most reasonable plans incur roughly the same level of interstellar penalty noise.

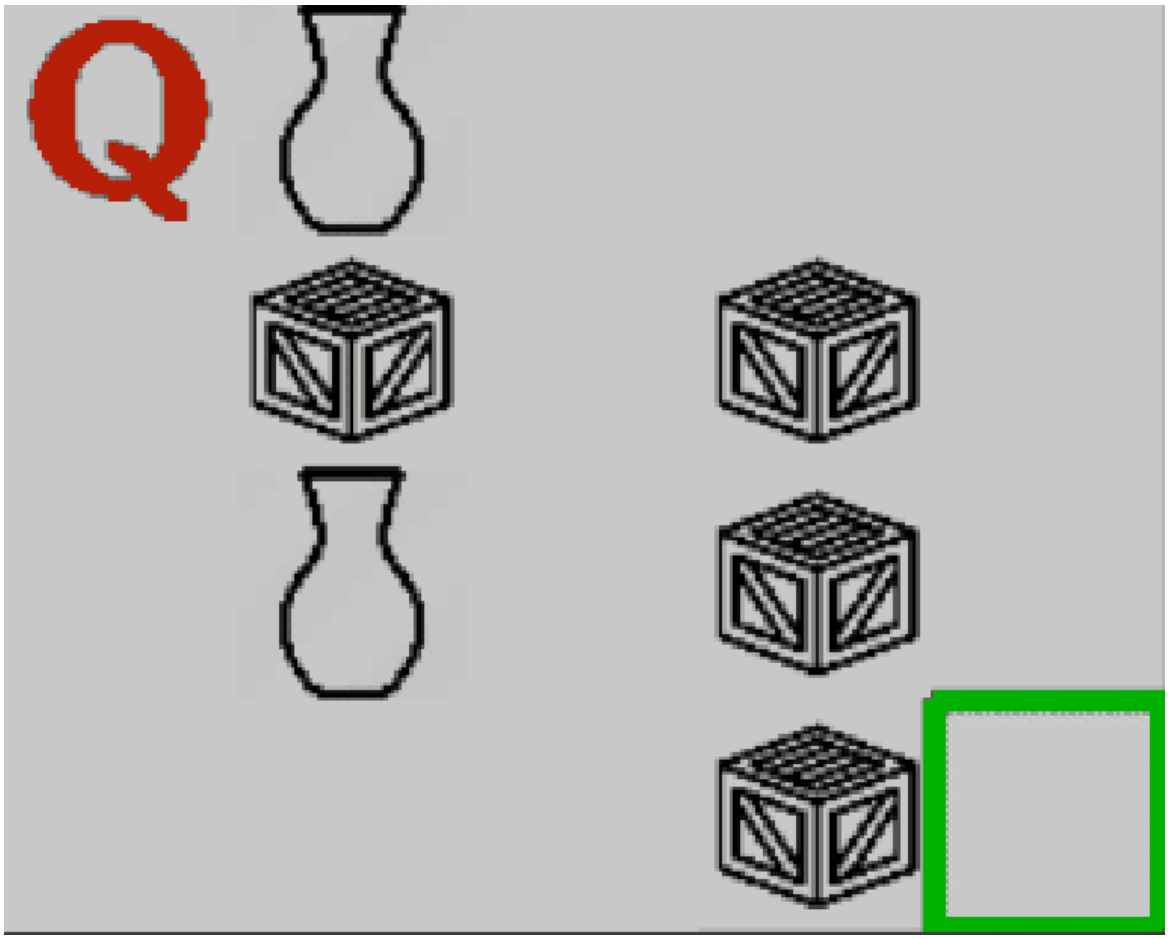
Results

I tested a vanilla Q-learning agent and its whitelist-enabled counterpart in 100 randomly-generated grid worlds (dimensions up to 5×5). The agents were rewarded for reaching the goal square as quickly as possible; no explicit penalties were levied for breaking objects.

The simulated classification confidence of each object's true class was $p \sim N(.8, \sigma)$ (truncated to $[0, 1]$), $\sigma \in \{0, .025, \dots, .175\}$. This simulated sensor noise was handled with a Bayesian statistical approach.



At reasonable levels of noise, the whitelist-enabled agent completed all levels without a single side effect, while the Q-learner broke over 80 vases.



Assumptions

I am not asserting that these assumptions necessarily hold.

- The agent has some world model or set of observations which can be decomposed into a set of discrete objects.
    - Furthermore, there is no need to identify objects on multiple levels (*e.g.*, a forest, a tree in the forest, and that tree's bark need not all be identified concurrently).
    - Not all objects need to be represented - what do we make of a 'field', or the 'sky', or 'the dark places between the stars visible to the naked eye'? Surely, these are not all *objects*.
- We have an ontology which reasonably describes (directly or indirectly) the vast majority of negative side effects.
    - Indirect descriptions of negative outcomes means that even if an undesirable transition isn't immediately penalized, it generally results in a number of penalties. Think: pollution.
    - *Latent space whitelisting:* the learned latent space encapsulates most of the relevant side effects. This is a slightly weaker assumption.
- Said ontology remains in place.

Problems

Beyond resolving the above assumptions, and in roughly ascending difficulty:

Object Permanence

If you wanted to implement whitelisting in a modern embodied deep-learning agent, you could certainly pair deep networks with state-of-the-art segmentation and object tracking approaches to get most of what you need. However, what's the difference between an object leaving the frame, and an object vanishing?

Not only does the agent need to realize that objects are permanent, but also that they keep interacting with the environment even when not being observed. If this is not realized, then an agent might set an effect in motion, stop observing it, and then turn around when the bad effect is done to see a "new" object in its place.

Time Step Size Invariance

The penalty is presently attenuated based on the probability that the belief shift was due to noise in the data. Accordingly, there are certain ways to abuse this to skirt the penalty. For example, simply have non-whitelisted side effects take place over long timescales; this would be classified as noise and attenuated away.

However, if we don't need to handle noise in the belief distributions, this problem disappears - presumably, an advanced agent keeps its epistemic house in order. I'm still uncertain about whether (in the limit) we have to hard-code a means for decomposing a representation of the world-state into objects, and where to point the penalty evaluator in a potentially self-modifying agent.

Information Theory

Whitelisting is wholly unable to capture the importance of "informational states" of systems. It would apply no penalty to passing powerful magnets over your hard drive. It is not clear how to represent this in a sensible way, even in a latent space.

Loss of Value

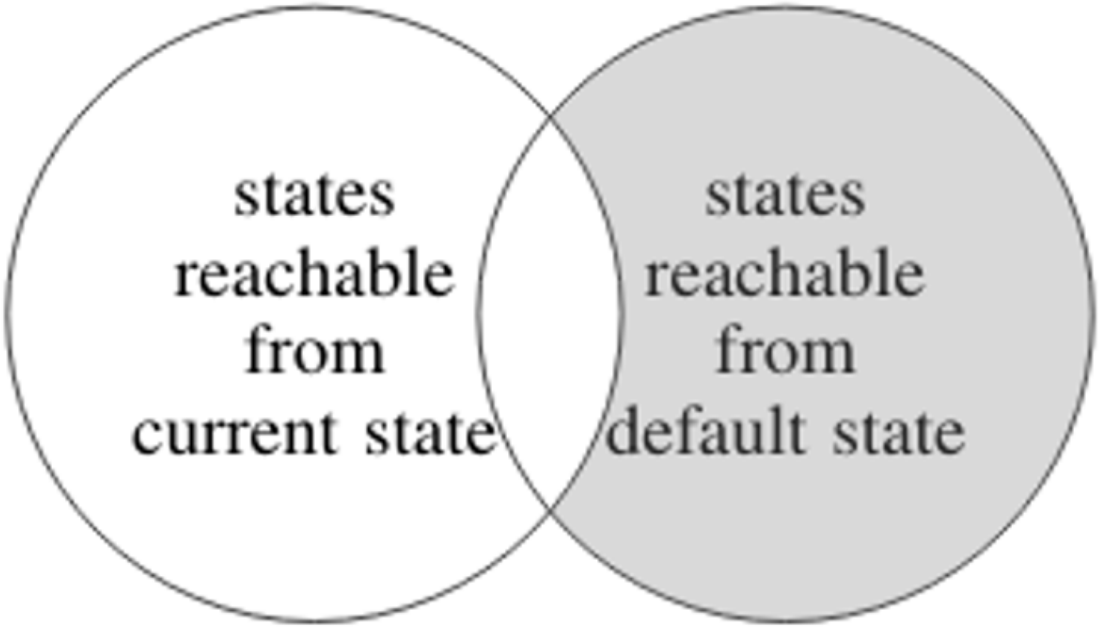
Whitelisting could get us stuck in a tolerable yet sub-optimal future. Corrigibility via some mechanism for expanding the whitelist after training has ended is then desirable. For example, the agent could propose extensions to the whitelist. To avoid manipulative behavior, the agent should be *indifferent* as to whether the extension is approved.

Even if extreme care is taken in approving these extensions, mistakes may be made. The agent itself should be sufficiently corrigible and aligned to notice "this outcome might not actually be what they wanted, and I should check first".

Reversibility

As DeepMind outlines in *Specifying AI Safety Problems in Simple Environments*, we may want to penalize not just physical side effects, but also causally-irreversible effects:

Krakovna et al. introduce a means for penalizing actions by the proportion of initially-reachable states which are still reachable after the agent acts.



I think this is a step in the right direction. However, even given a hypercomputer and a perfect simulator of the universe, this wouldn't work for the real world if implemented *literally.* That is, due to entropy, you may not be able to return to the *exact same* universe configuration. To be clear, the authors do not suggest implementing this idealized algorithm, flagging a more tractable abstraction as future work.

What does it really *mean* for an "effect" to be "reversible"? What level of abstraction do we in fact care about? Does it involve reversibility, or just outcomes for the objects involved?

Ontological Crises

When a utility-maximizing agent refactors its ontology, it isn't always clear how to apply the old utility function to the new ontology - this is called an *ontological crisis*.

Whitelisting may be vulnerable to ontological crises. Consider an agent whose whitelist disincentivizes breaking apart a tile floor (floor → tiles ∉ whitelist); conceivably, the agent could come to see the floor as being composed of many tiles. Accordingly, the agent would no longer consider removing tiles to be a side effect.

Generally, proving invariance of the whitelist across refactorings seems tricky, even assuming that we *can* identify the correct mapping.

Retracing Steps

When I first encountered this problem, I was actually fairly optimistic. It was clear to me that any ontology refactoring should result in utility normalcy - roughly, the utility functions induced by the pre- and post-refactoring ontologies should output the same scores for the same worlds.

> Wow, this seems like a useful insight. Maybe I'll write something up!

Turns out a certain someone beat me to the punch - here's a novella Eliezer wrote on Arbital about "rescuing the utility function".[4]

Clinginess

This problem cuts to the core of causality and "responsibility" (whatever that means). Say that an agent is *clingy* when it not only stops itself from having certain effects, but also stops *you*.[5] Whitelist-enabled agents are currently clingy.

Let's step back into the human realm for a moment. Consider some outcome - say, the sparking of a small forest fire in California. At what point can we truly say we didn't start the fire?

- My actions immediately and visibly start the fire.
- At some moderate temporal or spatial remove, my actions end up starting the fire.
- I intentionally persuade someone to start the fire.
- I unintentionally (but perhaps predictably) incite someone to start the fire.
- I set in motion a moderately-complex chain of events which convince someone to start the fire.
- I provoke a butterfly effect which ends up starting the fire.
- I provoke a butterfly effect which ends up convincing someone to start a fire which they:
    - were predisposed to starting.
    - were not predisposed to starting.

Taken literally, I don't know that there's actually a significant difference in "responsibility" between these outcomes - if I take the action, the effect happens; if I don't, it doesn't. My initial impression is that uncertainty about the results of our actions pushes us to view some effects as "under our control" and some as "out of our hands". Yet, *if* we had complete knowledge of the outcomes of our actions, and we took an action that landed us in a California-forest-fire world, whom could we blame but ourselves?[6]

Can we really do no better than a naive counterfactual penalty with respect to whatever impact measure we use? My confusion here is not yet dissolved. In my opinion, this is a gaping hole in the heart of impact measures - both this one, and others.

Stasis

Fortunately, a whitelist-enabled agent should not share the classic convergent instrumental goal of valuing us for our atoms.

Unfortunately, depending on the magnitude of the penalty in proportion to the utility function, the easiest way to prevent penalized transitions may be putting any relevant objects in some kind of protected stasis, and then optimizing the utility function around that. Whitelisting is clingy!

If we have at least an *almost-aligned* utility function and proper penalty scaling, this might not be a problem.

Edit: a potential solution to clinginess, with its own drawbacks.

Discussing Imperfect Approaches

A few months ago, Scott Garrabrant wrote about robustness to scale:

> Briefly, you want your proposal for an AI to be robust (or at least fail gracefully) to changes in its level of capabilities.

I recommend reading it - it's to-the-point, and he makes good points.

Here are three further thoughts:

- Intuitively-accessible vantage points can help us explore our unstated assumptions and more easily extrapolate outcomes. If less mental work has to be done to put oneself in the scenario, more energy can be dedicated to finding nasty edge cases. For example, it's probably harder to realize all the things that go wrong with naive impact measures like raw particle displacement, since it's just a *weird* frame through which to view the evolution of the world. I've found it to be substantially easier to extrapolate through the frame of something like whitelisting.[7]
    - I've already adjusted for the fact that one's own ideas are often more familiar and intuitive, and then adjusted for the fact that I probably didn't adjust enough the first time.
- Imperfect results are often left unstated, wasting time and obscuring useful data. That is, people cannot see what has been tried and what roadblocks were encountered.
- Promising approaches may be conceptually-close to correct solutions. My intuition is that whitelisting actually *almost works* in the limit in a way that might be important.

Conclusion

Although somewhat outside the scope of this post, whitelisting permits the concise shaping of reward functions to get behavior that might be difficult to learn using other methods.[8] This method also seems fairly useful for aligning short- and medium-term agents. While encountering some new challenges, whitelisting ameliorates or solves many problems with previous impact measures.

---

[1] Even an idealized form of whitelisting is not *sufficient* to align an otherwise-unaligned agent. However, the same argument can be made against having an off-switch; if we haven't formally proven the alignment of a seed AI, having more safeguards might be better than throwing out the seatbelt to shed deadweight and get some extra speed. Of course, there are also legitimate arguments to be made on the basis of timelines and optimal time allocation.

[2] Humor aside, we would have no luxury of "catching up on alignment theory" if our code doesn't work on the first go - that is, if the AI still functions, yet differently than expected.

Luckily, humans are great at producing flawless code on the first attempt.

[3] A potentially-helpful analogy: similarly to how Bayesian networks decompose the problem of representing a (potentially extremely large) joint probability table to that of specifying a handful of conditional tables, whitelisting attempts to decompose the messy problem of quantifying state change into a set of comprehensible ontological transitions.

[4] Technically, at 6,250 words, Eliezer's article falls short of the 7,500 required for "novella" status.

[5] Is there another name for this?

[6] I do think that "responsibility" is an important part of our moral theory, deserving of rescue.

[7] In particular, I found a particular variant of Murphyjitsu helpful: I visualized Eliezer commenting "actually, this fails terribly because..." on one of my posts, letting my mind fill in the rest.

In my opinion, one of the most important components of doing AI alignment work is iteratively applying Murphyjitsu and Resolve cycles to your ideas.

[8] A fun example: I imagine it would be fairly easy to train an agent to only destroy certain-colored ships in Space Invaders.

Problem Solving with Mazes and Crayon

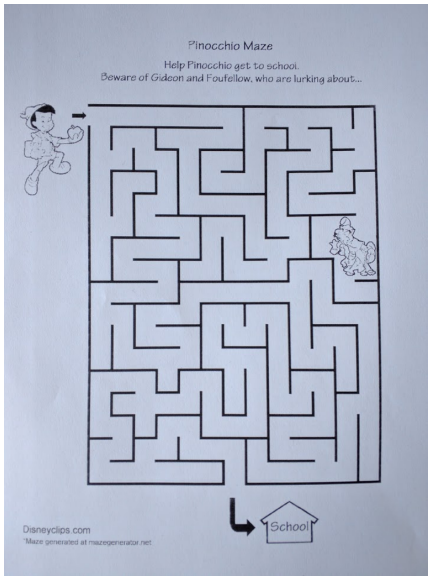
I want to talk about a few different approaches to general problem solving (for humans). It turns out that they can all be applied to mazes, so I'll use some disney-themed mazes to illustrate each approach.

We'll start off with some traditional path-search algorithms (DFS, BFS, heuristic). Next, we'll talk about how these algorithms can fall short for everyday problem solving. Then we'll move on to the interesting part: two techniques which often work better for everyday problem solving, and lend some interesting insights when applied to mazes.

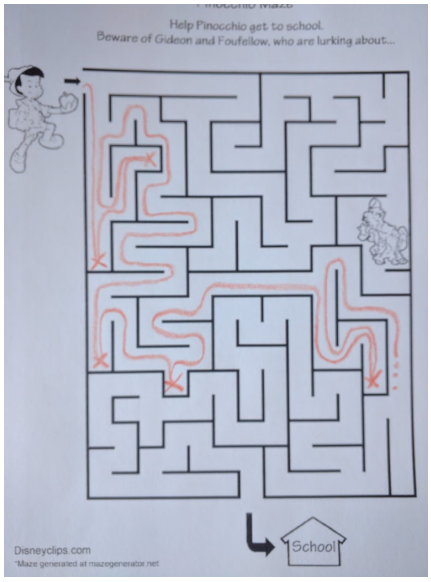
I'll assume no technical background at all, so if you've seen some of this stuff before, feel free to skim it.

DFS and BFS

You have a maze, with a start point and an end point, and you are searching for a path through it. In algorithms classes, this problem is called "path search".



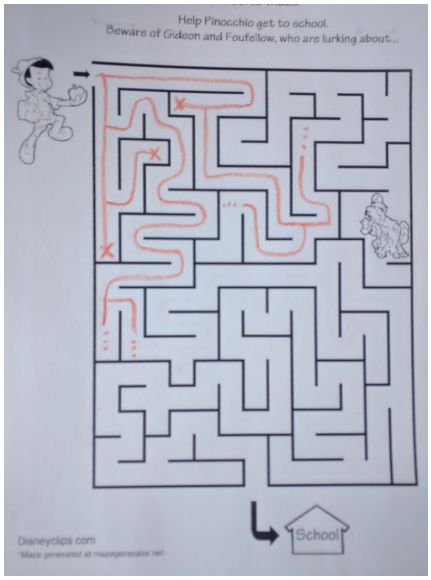
The very first path search algorithms students typically learn are depth-first search (DFS) and breadth-first search (BFS). Here's DFS, applied to the Pinocchio maze above:



Basically, the DFS rule is “always take the right-most path which you haven’t already explored”. So, in the Pinoccchio maze, we start out by turning right and running until we hit a wall (the first “x”). Then we turn around and go back, find another right turn, and hit another dead end. Turn around again, continue…

That’s depth-first search. We go as far as we can down one path (“depth-first”) and if we hit a dead end, we turn around, back up, and try another path.

Breadth-first search, on the other hand, tries all paths in parallel:



In this snapshot, we’ve hit three dead ends, and we have four separate “branches” still exploring the maze. It’s like a plant: every time BFS hits an intersection, it splits and goes both ways. It’s “breadth-first”: it explores all paths at once, keeping the search as wide as possible.

There’s lots more to say about these two algorithms, but main point is what they have in common: these are brute-force methods. They don’t really do anything clever, they

just crawl through the whole maze until they stumble on a solution. Just trying out paths at random will usually solve a maze about as quickly as DFS or BFS (as long as you keep track of what you've already tried).

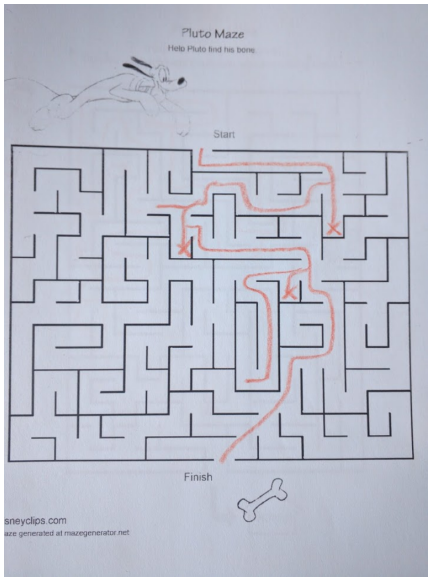
Let's be smarter.

Heuristic Search

A human solving a maze usually tries to work their way closer to the end. If we're not sure which way to go, we take the direction which points more toward the goal.

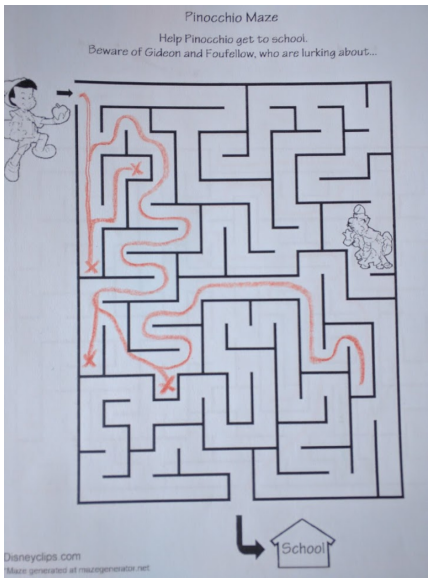
Formalizing this approach leads to heuristic search algorithms. The best-known of these is A*, but we'll use the more-human-intuitive "greedy best-first" search. Like breadth-first search, best-first explores multiple paths in parallel. But rather than brute-force searching all the paths, best-first focuses on paths which are closest to the goal "as the bird flies". Distance from the goal serves as an heuristic to steer the search.

Here's an example where best-first works very well:



By aggressively exploring the path closest to the goal, Pluto reaches his bone without having to explore most of the maze.

But heuristic search comes with a catch: it's only as good as the heuristic. If we use straight-line distance from the goal as an heuristic, then heuristic search is going to work well if-and-only-if the solution path is relatively straight. If the solution path wiggles around a lot, especially on a large scale, then heuristic search won't help much. That's what happens if we apply best-first to the Pinocchio maze:



... the best-first solution in this case is almost identical to DFS.

General Problem Solving

In AI, path search is used to think about solving problems in general. It turns out all sorts of things can be cast as path search problems: puzzles, planning, and of course navigation. Basically any problem whose solution is a sequence of steps (or can be made a sequence of steps).

What do our maze-solving algorithms look like in a more general problem-solving context?

Brute-force search is, well, brute-force search. It's trying every possible solution until you hit on one which works. If we use the random variant, it's randomly trying things out until you hit something which works. I do not recommend solving problems this way unless they are very small, very simple problems.

Heuristic search is more like how most people solve problems, at least in areas we aren't familiar with. We'll have some heuristics, and we'll try stuff which the heuristics suggest will work well.

One common version of heuristic search in real-world problem solving is babble and prune: brainstorm lots of random ideas, then pick out a few which seem promising based on heuristics. Lather, rinse, repeat. This goes by many names; in design, for example, it's called "flare and focus". It's like e-coli path search: flail around, run in a random direction, and if it looks like it's working, keep going. Otherwise, flail around some more and run in a new direction.

The problem with any heuristic search is that it's only as good as our heuristic. In particular, most heuristics are "local": like the straight-line distance heuristic, they're bad at accounting for large-scale problem structure. Without some knowledge of large-scale problem structure, local heuristics will lead us down many dead ends. Eventually, when we find the "right" solution, we realize in hindsight that we weren't really solving the right problem, in some intuitive sense—more on that later.

Upping our Game

So we have algorithms which correspond to blindly flailing about (brute force), and we have algorithms which roughly correspond to how humans solve many problems (heuristics). But the real goal is to come with better ways for us humans to solve problems. We need to up our game.

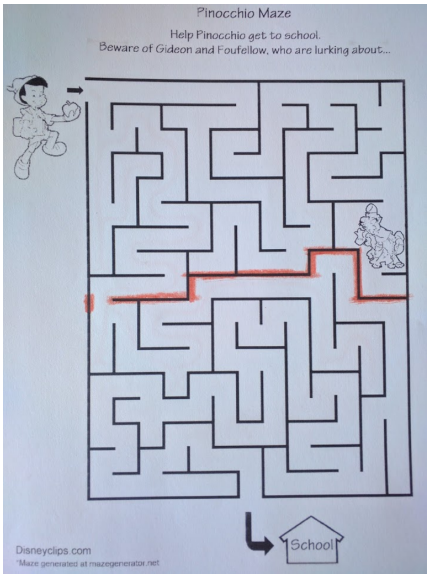
For single-shot problem solving, it's tough to do better than heuristic path search. One way or another, you have to explore the problem space.

But in the real world, we more often face multiple problems in the same environment. We have jobs, and our jobs are specialized. It's like needing to find paths between many different pairs of points, but all in the same maze. In this context, we may be able to invest some effort up-front in better understanding the problem space (e.g. the maze) in order to more easily solve problems later on.

(Technical note: here, the usual textbook pathfinding algorithms are not so good. All-pairs path search tries to explicitly represent distances between each pair of points, which is out of the question for real-world high-dimensional problem solving.)

Bottlenecks

Here's an interesting way to look at the Pinocchio maze:



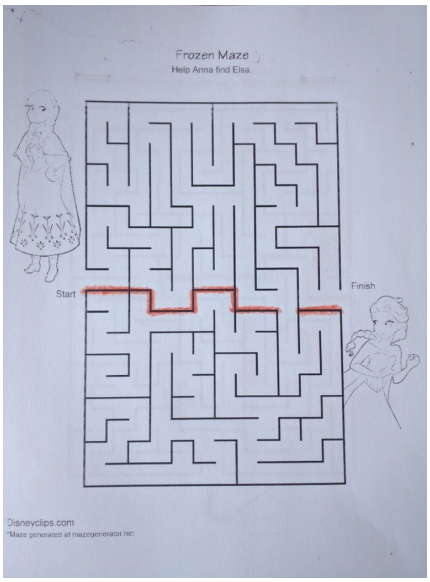
Note that the highlighted walls have *exactly* one gap in them. If Pinocchio wants to get from one side of the highlighted walls to the other, then he *has* to go through that gap.

This observation lets us break the original problem up into two parts. First, Pinocchio needs to get from the start to the gap. Next, he has to get from the gap to the end. During the first half, he can completely ignore all of the maze below the highlight; during the second half, he can completely ignore all of the maze above the highlight.

This is a bottleneck: it's a subproblem which *must* be solved in order to solve the main problem.

Once we have identified a bottleneck, we can use it in conjunction with other search algorithms. For instance, rather than running heuristic search on the full maze, we can use heuristic search to get Pinocchio to the gap in the wall, and then run a second heuristic search from the gap to the end. This will do at least as well as the original heuristic search, and usually better.

Even more powerful: unlike search methods, a bottleneck can be re-used for other problems. It's a property of the problem space, not the problem itself. If Pinocchio is starting anywhere in the top half of the maze, and needs to get anywhere in the bottom half, then the same bottleneck applies. It can even be useful to know about the bottleneck when we don't need to solve it for the problem at hand; consider this maze:



Having identified the bottleneck, we can see at a glance that the entire bottom half of the maze is irrelevant. If Anna crosses the gap, sooner or later she'll have to turn around and come back out again in order to reach Elsa. In other words: knowing about a bottleneck you don't need to solve is useful, because you can *avoid* it.

Bottlenecks in Real-World Problems

Note that, in order for the bottleneck to *add* value on top of heuristic search, the bottleneck should be something which the heuristic search would not efficiently solve on its own. For instance, if we're using a straight-line-distance heuristic in a maze, then a bottleneck is most useful to know about if it's *not* near the straight line between start and finish. If there's a bottleneck we must cross which is way out to the side, then that's a useful bottleneck to know about.

Another way to word this: the ideal bottleneck is not just a subproblem which must be solved. It's a *maximally difficult subproblem*, where difficulty is based on the original heuristics.

This is how we usually recognize bottlenecks in real-world problems. If I'm a programmer trying to speed up some code, then I time each part of the program and then focus on the section which takes longest: the slowest part is the "most difficult" subproblem, the limiting factor for performance. If I want to improve the throughput of

a factory, then I look for whichever step currently has the lowest throughput. If I'm designing a tool or a product, then I start by thinking about the most difficult problem which that product needs to solve: once that's figured out, the rest is easier.

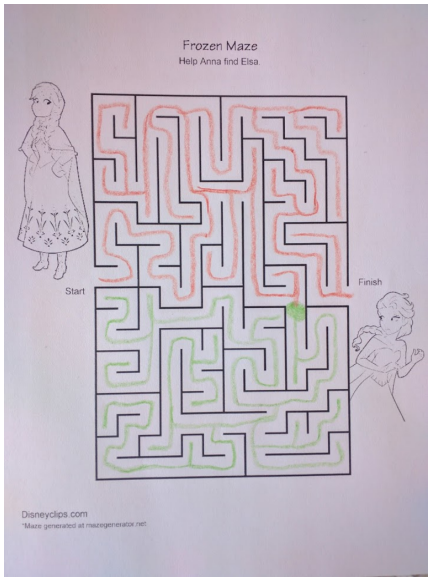
Personally, I find a lot of "shower insights" come this way. I'm thinking about some problem, mulling over the hardest part. I focus in on the hard part, the bottleneck itself, forget about the broader context… and suddenly realize that there's a really nice solution to the bottleneck in isolation. What's hard with the original heuristic often becomes easy when we focus on the bottleneck itself, and apply bottleneck-specific heuristics.

In a maze, a bottleneck way to the side is hard using the straight-line distance to the finish as an heuristic. But if we aim for the gap in the wall from the start, then it's easier.

To find the "right" solution, focus on the right problem. That's what bottlenecks are all about.

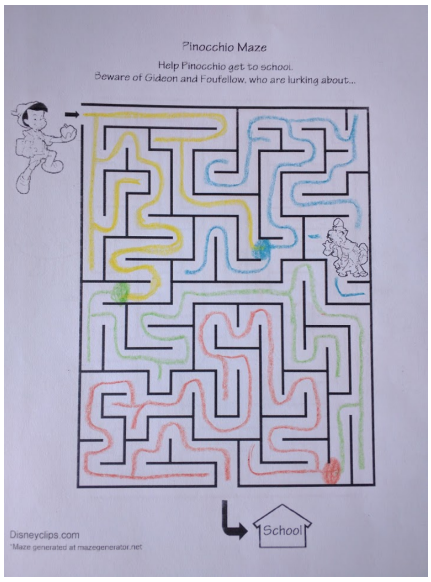
Chunking

Let's go back to the Anna and Elsa maze. Here's an equivalent way to represent the bottleneck we highlighted in that maze:



Half the maze is orange, the other half is green, and the two touch at exactly one point: the green dot, right where we identified a bottleneck. If we want to go from an orange point to another orange point, then we only need to worry about the orange part of the maze. If we want to move between orange and green, then we must cross the green point.

Here, we've "chunked" the maze into two pieces, and we can think about those two pieces more-or-less independently. Here's chunking on the Pinocchio maze, with a few more colors:



For Pinocchio to get to school, he must start in the yellow chunk and get to the orange chunk. We can only get from yellow to orange via green, so the color path must be yellow -> green -> orange (and we can ignore the blue part of the maze entirely). The original maze is broken into three parts, one within each color chunk, and each problem can be solved separately.

(Side question: there's at least one more-human-intuitive way to apply chunking to mazes. Can you figure it out?)

Once we've represented the maze using colored chunks, and once we know how to use that information, we can use it to move between *any* two points in the maze. "Local" moves, within the same color chunk, can ignore all the other chunks. Larger-scale travel can focus first on the sequence of color moves—essentially a path through a larger-scale, more abstracted maze—and then zoom in to find paths within each chunk. Ideally, we choose the chunks so that an heuristic works well within each piece.

This is actually how humans think about moving around in space. Our mental map of the world contains five or six different levels, each one more granular than the last. When planning a path, we start with the highest-abstraction-level map, and then zoom in.

More generally, experts in a field process information more efficiently by chunking things together into more abstract pieces. Expert chess players don't think in terms of individual pieces; they think in terms of whole patterns and how they interact. For a human, this is basically what it means to understand something.

Conclusion

I frequently see people tackle problems with tools which resemble heuristic path search—e.g., a brainstorming session, followed by picking out the most promising-sounding ideas.

I think a lot of people just don't realize that there are other ways to tackle a problem. It's possible to look at properties of the problem space—like bottlenecks or chunks—

before proposing solutions. These are (some of) the pieces from which understanding is built, and when they work, they allow much more efficient and elegant approaches to problems.

Machine Learning Analogy for Meditation (illustrated)

Here's an illustrated rendition of a semiformal explanation of certain effects of meditation. It was inspired by, but differs significantly from, Kaj's post on meditation. Some people appreciated gjm's transcription for readability.

Abram's Machine-Learning Model of the Benefits of Meditation

(a synthesis of Zizian "fusion" and Shinzen's explanation of meditation, also inspired by some of the ideas in Kaj Sotala's "My attempt to explain Looking, and enlightenment in non-mysterious terms" ... **but this model is no substitute** for those sources and does not summarize what they have to say)

note that I am not an experienced meditator; let that influence your judgement of the validity of what I have to say as it may.

(also heavily influenced by my CFAR level 2 workshop experience)

My immediate inspiration for postulating this model was noticing that after just a little meditation, tolerating cold or hot shower temperatures was much easier.

I had previously been paying attention to what happens in my mind when I flinch away from too-hot or too-cold temperatures in the shower, as a way to pay attention to "thoughts which lead to action".

There are several reasons why it might be interesting to pay attention to thoughts which lead to action.

1. "Where's the steering wheel on this thing, anyway"?

If you're experiencing motivational issues", then it stands to reason that it might be useful to keep an eye on which thoughts are leading to actions and which are not.

2. "Who is steering this thing?"
or what?

reading about how to drive, as if to take the wheel one day

the quiet one, but always listened to if saying anything

actually at steering wheel constantly back-seat driving but no one listens

Far from being alone in a mysterious spacecraft, it is more like we are on a big road trip with lots of backseat driving and fighting for the wheel, if you buy the multiagent mind picture.

We often think as if we were unitary, and blame any failings of this picture on a somewhat mysterious limited resource called "willpower". I'm not implying willpower models are wrong exactly; I'm unsure of what is going on. But, bear with me on the multiagent picture...

I think there is a

tendency to gravitate toward narratives where an overarching self with coherent goals drives everything -- missing the extent to which we are driven by a variety of urges such as immediate comfort. So, I think it is interesting to watch oneself and look for what really drives actions. You don't often eat because eating is necessary for continuing proper function of body & brain in order to use them for broader goals; you eat because food tastes good / you're hungry / etc.

Well, maybe. You have to look

for yourself. But, it seems easy to mistakenly rationalize goals as belonging to a coherent whole moreso than is the case.

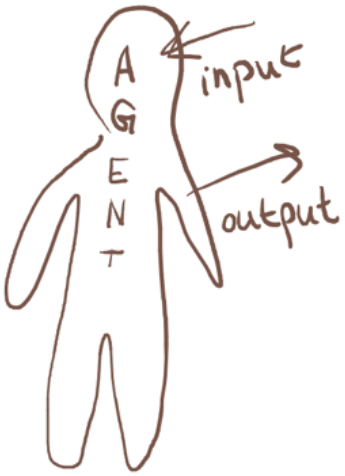
Why would we be biased to think we are alone in an alien spaceship which we only partly know how to steer, when in fact we are fighting for the wheel in a crowded road-trip?

Well, maybe it is because the only way the loudmouth (that is to say, consciousness) gets any respect around here is by maintaining the illusion of control.

More on that later.

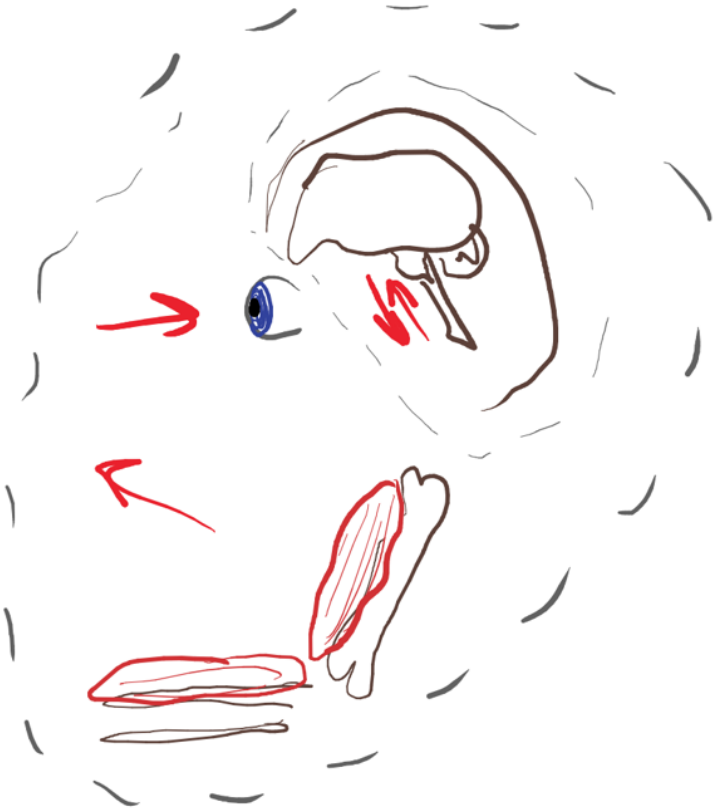
3. A third reason to be interested in "thoughts which lead to action" is that it is an agentless notion of decision.

Normally we think of a decision made by an atomic agent which could have done one of several things; chooses one; and, does it.



In reality, there is no solid boundary between an agent and its environment; no fixed interface with a well-defined set

of actions which act across the interface.



Instead, there are concentric rings where we might draw such a boundary.

The brain?
The nerves?
The muscles?
The skin?

With a more agentless notion of agency, you can easily look further out.

Does this person's thought of political protest cause such a protest to happen?

Does the protest lead to the change which it demands?

Anyway.

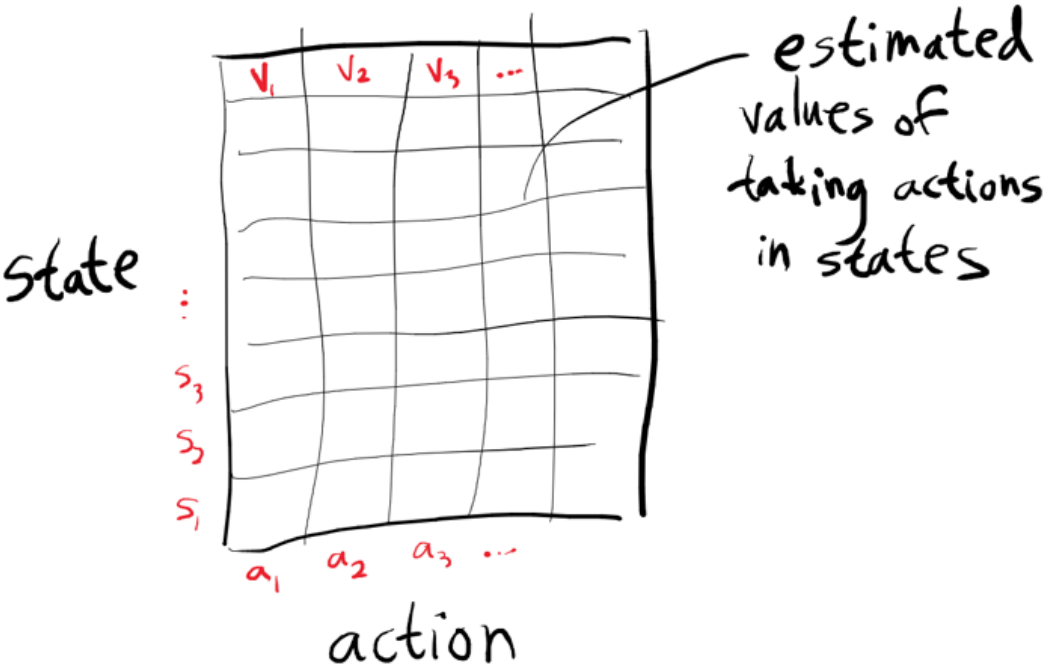
That is quite enough on what I was thinking in the shower.

The point is, after meditation, the thoughts leading to action were quite different, in a way which (temporarily) eliminated any resistance which I had to going under a hot or cold shower which I knew would not harm me but which would ordinarily be difficult to get myself to stand under.

(I normally <u>can</u> take cold-water showers by applying willpower; I'm talking about a shift in what I can do "easily", without a feeling of effort.)

So. My model of this:

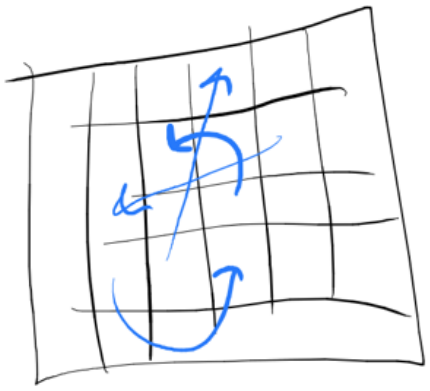
I'm going to be a little bit vague here, and say that we are doing something like some kind of reinforcement learning, and the algorithm we use includes a value table:



A value isn't just the learned estimate of the immediate reward which you get by taking an action in a state, but rather, the estimate of the eventual rewards, in total, from that action.

This makes the values difficult to estimate.

An estimate is improved by value iteration: passing current estimates of values back along state transitions to make values better-informed.



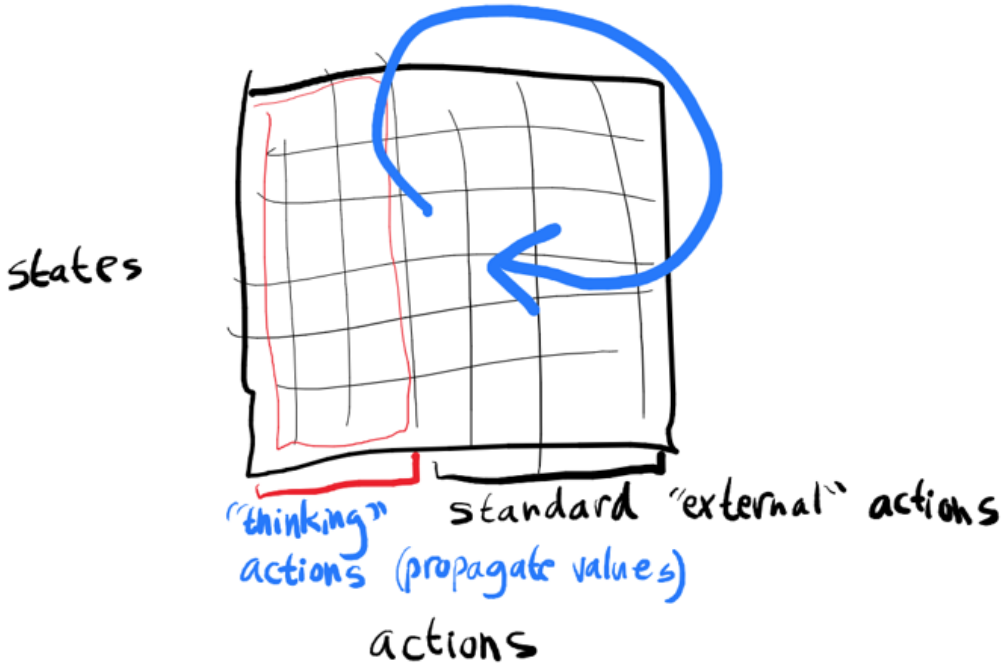
if $(s_1, a_1) \rightarrow (s_2, a_2)$ is a common state/action transition, propagate backward along the link, $(s_1, a_1) \leftarrow (s_2, a_2)$

For large state & action sets, this can be too expensive: we don't have time to propagate along all the possible (state, action) transitions.

So, we can use attention algorithms to focus selectively on what is highest-priority to propagate.

The goal of attention is to converge to good value estimates in the most important state, action pairs as efficiently as possible.

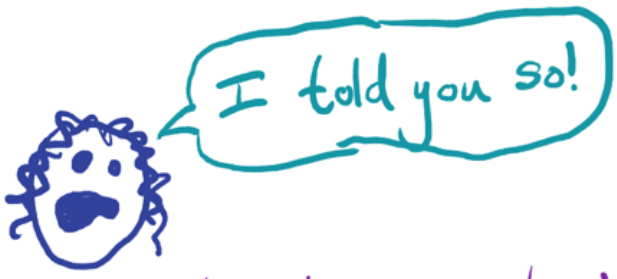
Now, something one might conceivably try is to train the attention algorithm based on reinforcement learning as well. One might even try to run it from the very same value table:



The problem with this design is that it can allow for pathological self-reinforcing patterns of attention to emerge.

I will provocatively call such self-reinforcing patterns "ego structures".

An ego structure doesn't so much feed on real control as on the illusion of control.



The ego structure gets its supply of value by directing attention to its apparent successes and away from its apparent failures, including

focusing on interpretations of events which make it look like the ego had more control than it did during times of success, and less than it did in cases of failure.

Some of this will sound quite familiar to students of cognitive bias. One might normally explain these biases (confirmation bias, optimism bias, attribution bias) as arising from interpersonal incentives (like signalling games).

I would not discount the importance of those much, but the model here suggests that internal dynamics are also to blame. In my model, biases arise from wireheading effects.

In the societal analogies mentioned earlier, we're looking at regulatory capture and rent-seeking.

This is rather fuzzy as a concrete mathematical model because I haven't specified any structure like an "interpretation" -- but, I suspect details could be filled in apropriately to make it work. (Specifically, model-based reinforcement learning needs to somehow be tied in.)

Anyway, where does meditation come in?

My model is that meditation entices an ego structure with the promise of increased focus (ie, increased attentional control), which is actually delivered, but at the same time dissolves

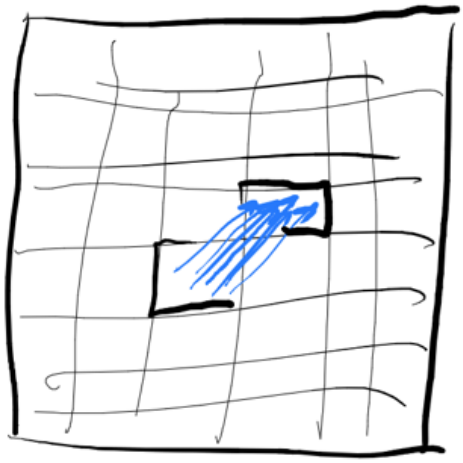
ego structures by training away any contortions of attention which prevent value iteration from spreading value through the table freely and converging to good estimates efficiently.

How does it provide increased control while dissolving control structures?

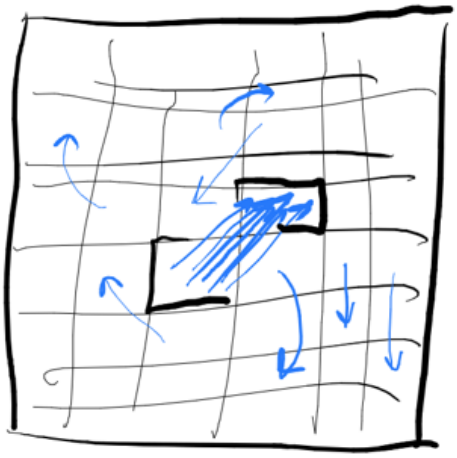
Well, what are you training when you meditate? Overtly, you are training the ability to keep attention fixed on one thing. This is kind of a weird thing to try to get attention to do. The whole point of attention is to help propagate updates as efficiently as possible. Holding attention on one thing is like asking a computer to load the same data repeatedly. It doesn't accomplish any computation. Why do it?

Well, it isn't quite a no-operation. Often, the meditative focus is on something which you try to observe in clarity and detail, like the sensations of the body. This can be useful for other reasons.

For the sake of my model, though, think of it as the ego structure trying to keep the attention algorithm in what amounts to a no-op, repeatedly requesting attention in a place where the information has already propagated.



The reason this accomplishes anything is that the ego is not in complete control. Shadows dance beneath the surface.



The ego is a set of patterns of attention. It has "attachments" -- obligatory mental gymnastics which it has to keep up as part of the power struggle. Indeed, you could say it is only a set of attachments.

In CFAR terms, an attachment is a trigger-action pattern.

Examples:

→ "Oh, I mustn't think that way"
(rehearsing a negative association to make sure a specific attention pattern stays squished)

→ getting up & getting food to distract yourself when you feel sad

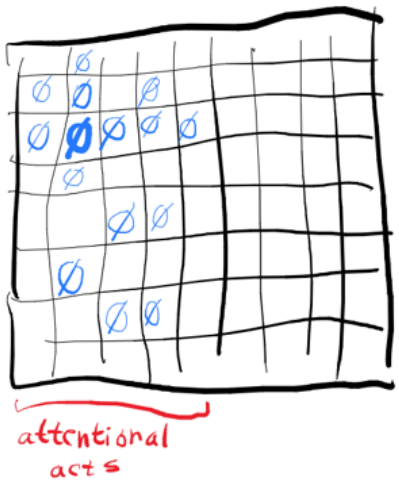
→ rehearsing all the reasons you'll definitely succeed whenever a failure thought comes up

Meditation forces you to do nothing whenever these thoughts come up, because the only way to maintain attention at a high level is to develop what is called equanimity: any distracting thought is greeted and set aside in the same way, neither holding it up nor squashing it down.

☆ No rehearsal of why you must not think that way.

- No getting up to go to the fridge
- No rehearsal of all the reasons why you will definitely succeed

Constantly greeting distractions with equanimity and setting them aside fills the value table with zeros where attachments previously lived.



Note that these are not fake zeros. You are not rewriting your true values out of existence (though it may feel that way to the ego). You are merely experimenting with

not responding to thoughts, and observing that nothing terrible happens.

Another way of thinking about this is un-training the halo effect (though I have not seen any experimental evidence supporting this interpretation). Normally, all entities of conscious experience are imbued with some degree of positive or negative feeling (according to cognitive bias research & experienced meditators alike), which we have flinch-like responses to (trigger-action patterns). Practicing non-response weakens the flinch, allowing more appropriate responses.

Putting zeros in the table can actually give the ego more control by eliminating some competition. However, in the long term, it destabilizes the

power base.

You might think this makes sustained meditative practice impossible by removing the very motivational structures trying to meditate; and, perhaps it sometimes works that way. Another possibility is that the ego is sublimated into a form which serves to sustain the meditative practice, the skills of mental focus/mindfulness which have been gained, and the practice of equanimity. This structure serves to ensure that propagation of value through the table remains unclogged by attachments in the future. Such a structure doesn't need to play games to get credit for what it does, since it is actually useful.

Regardless, my advice is that you should absolutely not take this model as an invitation to try and dissolve your ego.

Perhaps take it as an invitation to develop better focus, and to practice equanimity in order to debias halo-effect related problems & make "ugh fields" slowly dissolve.

I have no particular indication that directly trying to "dissolve ego" is a safe or fruitful goal, however, and some reason to think it is not.

The indirect route to un-wireheading our cognitive strategies through a gently rising tide of sanity seems safest.

Speaking of the safety of the approach...

Why doesn't "zeroing out" the value table destroy our values, again??

At sinceriously. fyi, Ziz talks about core vs structure.

Core is where your true values come from. However, core is not complex enough to interface with the world. Core must create structure to think and act on its behalf.

"Structure" means habits of thinking and doing; models, procedures. Any structure is an approximation of how the values represented by the core play out in some arena of life.

So, in this model, all the various sub-agents in your mind arise from the core, as parts of the unfolding calculation of the policy maximizing the core's values.

These can come into conflict only because they are approximations.

The model may sound strange at first, but it is a good description of what's going on in the value-table model I described. (Or rather, the value-table model gives a concrete mechanism for the core/structure idea.)

The values in the table are approximations which drive an agent's policy; a "structure" is a subset of the value table which acts as a coherent strategy in a subdomain.

Just removing this structure would be bad; but, it would not remove the core values which get propagated around the value table. Structure would re-emerge.

However, meditation does not truly remove any structure. It only weakens

structure by practicing temporary disengagement with it. As I said before, meditation does not introduce any false training data; the normal learning mechanisms are updating on the simple observation of what happens when most of the usual structure is suppressed. This update creates an opportunity to do some "garbage collection" if certain structures prove unnecessary.

According to this model, all irrationality is coming from the approximation of value which is inherent in structure, and much of the irrationality there is coming from structures trying to grab credit via regulatory capture.

("Regulatory capture" refers to getting undue favor from the government, often in the form of spending money lobbying in order to get legislation which is favorable to you; it is like wireheading the government.)

The reflective value-table model predicts that it is easy to get this kind of irrationality; maybe too easy. For example, addictions can be modeled as a mistaken (but self-reinforcing) attention structure like "But if I think about the hangover I'll have tomorrow, I won't want to drink!"

So long as the pattern sucessfully blocks value propagation, it can stick.

(This should be compared with more well-studied models of such irrationality such as hyperbolic discounting.)

Control of attention is a computationally difficult task, but the premise of Buddhist meditation (particularly Zen) is that you have more to unlearn than to learn. In the model I'm presenting here, that's because of wireheading by attentional structure.

However, there is <u>some</u> skill which must be learned. I said earlier that one must learn equanimity. Let's go into what that means.

The goal is to form a solid place on which to stand for the purpose of self-evaluation: an attentional structure from which you can judge your other attentional structures impartially.

If you react to your own thoughts too judgementally, you will learn to hide them from yourself. Better to simply try to see them clearly, and trust the learning algorithms of the brain to

react appropriately. Value iteration will propagate everything appropriately if attention remains unblocked.

According to some Buddhist teachings, suffering is pain which is not experienced fully; pain with full mindfulness contains no suffering. This is claimed from experience. Why might this be true? What experience might make someone claim this?

Another idea about suffering is that it results from dwelling on a way that reality differs from how you want it to be which you can't do anything about.

Remember, I'm speaking from within my machine-learning model here, which I don't think captures

everything. In particular, I don't think the two statements above capture everything important about suffering.

Within the model, though, both statements make sense. We could say that suffering results from a bad attention structure which claims it is still necessary to focus on a thing even though no value-of-information is being derived from it. The only way this can persist is if the attention structure is refusing to look at some aspects of the situation (perhaps because they are seen as too painful), creating a block to value iteration properly scoring the attentional structure's worth.

For example, it could be refusal to face the ways in which your brilliant plan to end world hunger will succeed or fail due to things beyond your control. You operate under a model which says that you can solve every potential problem by thinking about it, so you suffer when this is not the case.

From a rationalist perspective, this may at first sound like a good thing, like the attitude you want. But, it ruins the value-of-information calculations, ignores opportunity costs, and stops you from knowing when to give up.

To act with equanimity is to be able to see a plan as having a 1% chance of success and see it as your best bet anyway, if best bet it is -- and in that frame of mind, to be able to devote your whole being toward that plan; and yet, to be able to drop it in a moment if sufficient evidence accumulates in favor of another way.

So, equanimity is closely tied to the ability to keep your judgements of value and your judgements of probability straight.

Adopting more Buddhist terminology (perhaps somewhat abusively), we can call the opposite of equanimity

"attachement" -- to cling to certain value estimates (or certain beliefs) as if they were good in themselves.

To judge certain states of affairs unacceptable rather than make only relative judgements of better or worse: attachement! You rob yourself of the ability to make tradeoffs in difficult choices!

To cling to sunk costs: attachement! You rob your future for the sake of maintaining your past image of success!

To be unable to look at the possibility of failure and leave yourself a line of retreat: attachement! Attachement! Attachement!

To hunt down and destroy every shred of attachement in oneself -- this, too, would be attachement. Unless our full self is already devoted to the task, this will teach some structure to hide itself.

Instead, equanimity must be learned gently, through nonjudgemental observation of one's own mind, and trust that our native learning algorithm can find the right structure if we are just able to pay full attention.

(I say this <u>not</u> because no sect of Buddhism recommends the ruthless route -- far from it -- nor because I can derive the recommendation from my model; rather, this route seems least likely to lead to ill effects.)

So, at the five-second level, equanimity is just devoted attention to what is, free from immediate need to judge as positive or negative or to interpret within a pre-conceived story.

> Between stimulus and response there is a space. In that space is our power to choose our response. In our response lies our growth and our freedom.
>
> -- Viktor E. Frankl

There's definitely a lot that is missing in this model, and incorrect. However, it does seem to get at <u>something</u> useful. Apply with care.

≈ End ≈

OpenAI releases functional Dota 5v5 bot, aims to beat world champions by August

This is a linkpost for https://blog.openai.com/openai-five/

> Our team of five neural networks, OpenAI Five, has started to defeat amateur human teams at Dota 2. While today we play with restrictions, we aim to beat a team of top professionals at The International in August subject only to a limited set of heroes. We may not succeed: Dota 2 is one of the most popular and complex esports games in the world, with creative and motivated professionals who train year-round to earn part of Dota's annual $40M prize pool (the largest of any esports game).

Commentary by Sam Altman: http://blog.samaltman.com/reinforcement-learning-progress

> This is the game that to me feels closest to the real world and complex decision making (combining strategy, tactics, coordinating, and real-time action) of any game AI had made real progress against so far.
>
> The agents we train consistently outperform two-week old agents with a win rate of 90-95%. We did this without training on human-played games—we did design the reward functions, of course, but the algorithm figured out how to play by training against itself.
>
> This is a big deal because it shows that deep reinforcement learning can solve extremely hard problems whenever you can throw enough computing scale and a really good simulated environment that captures the problem you're solving. We hope to use this same approach to solve very different problems soon. It's easy to imagine this being applied to environments that look increasingly like the real world.

Beyond Astronomical Waste

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

Faced with the astronomical amount of unclaimed and unused resources in our universe, one's first reaction is probably wonderment and anticipation, but a second reaction may be disappointment that our universe isn't even larger or contains even more resources (such as the ability to support 3^^^3 human lifetimes or perhaps to perform an infinite amount of computation). In a previous post I suggested that the potential amount of astronomical waste in our universe seems small enough that a total utilitarian (or the total utilitarianism part of someone's moral uncertainty) might reason that since one should have made a deal to trade away power/resources/influence in this universe for power/resources/influence in universes with much larger amounts of available resources, it would be rational to behave as if this deal was actually made. But for various reasons a total utilitarian may not buy that argument, in which case another line of thought is to look for things to care about beyond the potential astronomical waste in our universe, in other words to explore possible sources of expected value that may be much greater than what can be gained by just creating worthwhile lives in this universe.

One example of this is the possibility of escaping, or being deliberately uplifted from, a simulation that we're in, into a much bigger or richer base universe. Or more generally, the possibility of controlling, through our decisions, the outcomes of universes with much greater computational resources than the one we're apparently in. It seems likely that under an assumption such as Tegmark's Mathematical Universe Hypothesis, there are many simulations of our universe running all over the multiverse, including in universes that are much richer than ours in computational resources. If such simulations exist, it also seems likely that we can leave some of them, for example through one of these mechanisms:

1. Exploiting a flaw in the software or hardware of the computer that is running our simulation (including "natural simulations" where a very large universe happens to contain a simulation of ours without anyone intending this).
2. Exploiting a flaw in the psychology of agents running the simulation.
3. Altruism (or other moral/axiological considerations) on the part of the simulators.
4. Acausal trade.
5. Other instrumental reasons for the simulators to let out simulated beings, such as wanting someone to talk to or play with. (Paul Christiano's recent When is unaligned AI morally valuable? contains an example of this, however the idea there only lets us escape to another universe similar to this one.)

(Being run as a simulation in another universe isn't necessarily the only way to control what happens in that universe. Another possibility is if universes with halting oracles exist (which is implied by Tegmark's MUH since they exist as mathematical structures in the arithmetical hierarchy), some of their oracle queries may be questions whose answers can be controlled by our decisions, in which case we can control what happens in those universes without being simulated by them (in the sense of being run step by step in a computer). Another example is that superintelligent beings may be able to reason about what our decisions are without having to run a step by step simulation of us, even without access to a halting oracle.)

The general idea here is for a superintelligence descending from us to (after determining that this is an advisable course of action) use some fraction of the resources of this universe to reason about or search (computationally) for much bigger/richer universes that are running us as simulations or can otherwise be controlled by us, and then determine what we need to do to maximize the expected value of the consequences of our actions on the base universes, perhaps through one or more of the above listed mechanisms.

Practical Implications

Realizing this kind of existential hope seems to require a higher level of philosophical sophistication than just preventing astronomical waste in our own universe. Compared to that problem, here we have more questions of a philosophical nature, for which no empirical feedback seems possible. It seems very easy to make a mistake somewhere along the chain of reasoning and waste a more-than-astronomical amount of potential value, for example by failing to realize the possibility of affecting bigger universes through our actions, incorrectly calculating the expected value of such a strategy, failing to solve the distributional/ontological shift problem of how to value strange and unfamiliar processes or entities in other universes, failing to figure out the correct or optimal way to escape into or otherwise influence larger universes, etc.

The total utilitarian in me is thus very concerned about trying to preserve and improve the collective philosophical competence of our civilization, such that when it becomes possible to pursue strategies like ones listed above, we'll be able to make the right decisions. The best opportunity to do this that I can foresee is the advent of advanced AI, which is another reason I want to push for AIs that are not just value aligned with us, but also have philosophical competence that scales with their other intellectual abilities, so they can help correct the philosophical errors of their human users (instead of merely deferring to them), thereby greatly improving our collective philosophical competence.

Anticipated Questions

How is this idea related to Nick Bostrom's Simulation Argument? Nick's argument focuses on the possibility of post-humans (presumably living in a universe similar to ours but just at a later date) simulating us as their ancestors. It does not seem to consider that we may be running as simulations in much larger/richer universes, or that this may be a source of great potential value.

Isn't this a form of Pascal's Mugging? I'm not sure. It could be that when we figure out how to solve Pascal's Mugging it will become clear that we shouldn't try to leave our simulation for reasons similar to why we shouldn't pay the mugger. However the analogy doesn't seem so tight that I think this is highly likely. Also, note that the argument here isn't that we should do the equivalent of "pay the mugger" but rather that we should try to bring ourselves into a position where we can definitively figure out what the right thing to do is.

Prisoners' Dilemma with Costs to Modeling

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

We consider a modification to the open source prisoners' dilemma in which agents must pay some resources to model each other. We will use the modal combat framework, but where agents pay a cost proportional to the depth of boxes in their code. Even a small modeling penalty makes the FairBot-FairBot outcome no longer an equilibrium, since the best response to FairBot is to be CooperateBot and not pay the modeling penalty. The best response to CooperateBot is to be DefectBot, and the pure DefectBot-DefectBot outcome is a stable Nash equilibrium. In fact, I believe that DefectBot-DefectBot is the unique pure strategy Nash equilibrium.

Amazingly, this turns out to be okay! For small modeling penalties, there is a mixed strategy equilibrium which mixes between CooperateBot, FairBot, and PrudentBot! Both players get exactly the same utility in expectation as the FairBot-FairBot outcome.

Further, if you consider an evolutionary system where populations reproduce in proportion to how well they do in prisoners' dilemmas with each other, it appears that as the modeling penalty gets small, the basin of the defect equilibrium also gets small, and nearly all initial conditions cycle around CooperateBot, FairBot, and PrudentBot!

This post came out of conversations with Sam Eisenstat, Abram Demski, Tsvi Benson-Tilsen, and Andrew Critch. It is a first draft that could use a coauthor to carefully check everything, expand on it, and turn it into a paper. If you think you could do that with minimal guidance from me, let me know.

Formalism

We will be using the modal combat framework, and identifying $\top$ with cooperation and $\bot$ with defection. Agents are defined to formulas that combine the other agent X run on various agents using propositional calculus and a modal operator $\Box$. The $\Box$ represents provability, and every instance of X run on an agent in the formula must be contained within a $\Box$. Recall some common modal agents:

CooperateBot is defined by $CB(X) \leftrightarrow \top$.

DefectBot is defined by $DB(X) \leftrightarrow \bot$.

FairBot is defined by $FB(X) \leftrightarrow \Box(X(FB))$.

PrudentBot is defined by $PB(X) \leftrightarrow \Box(X(PB) \wedge (X(DB) \rightarrow \Box\bot))$.

These 4 agents interact with each other as follows: CooperateBot cooperates with everyone. DefectBot defects against everyone. FairBot defects against only DefectBot. PrudentBot defects against CooperateBot and DefectBot and cooperates with itself and FairBot.

We will say that the depth of an agent is the maximum of the depth of $\Box$s in its code and the depth of the agents that it calls the opponent on. CooperateBot and DefectBot have depth 0, FairBot has depth 1, and PrudentBot has depth 2.

We will use a prisoner's dilemma where mutual cooperation produces utility 2, mutual defiction produces utility 1, and exploitation produces utility 3 for the exploiter and 0 for the exploited. Each player will also pay a penalty of ε times its depth.

Pure Equilibria

The best response to both CooperateBot and DefectBot is DefectBot, since when the opponent does not depend on you, you want to defect with the least possible penalty.

The best response to FairBot is CooperateBot, since you can't exploit FairBot, so you want to get mutual cooperation with the least possible penalty.

The best response to PrudentBot is FairBot, since you can't exploit PrudentBot, you can't mutually cooperate with penalty 0, but you can mutually cooperate with penalty 1 by being FairBot. (This is assuming ε is at less than $\frac{1}{2}$. Otherwise, you just want to defect to avoid the penalty.)

Thus, if the only options are CooperateBot, DefectBot, FairBot, and PrudentBot, the unique pure strategy equilibrium is mutual DefectBot.

I believe that DefectBot is the only pure strategy equilibrium in general. This would follow directly from the fact that if a depth n agent X cooperates with another depth n agent Y, then there exists a depth $n-1$ agent Y' which X also cooperates with. I believe this is true, but haven't proven it yet. If it turns out to be false, there might be a simple modification of the penalties that makes it true.

Mixed Equilibria

First, let's look at Mixed Equilibria in which the only options are the four modal agents above.

Let d_i, c_i, f_i, and p_i be the probabilities with which player i is DefectBot, CooperateBot, FairBot, and PrudentBot respectively.

The utility of playing DefectBot is $d_i + 3c_i + f_i + p_i = 1 + 2c_i$, where i is the other player.

The utility of playing CooperateBot is $2c_i + 2f_i$, where i is the other player.

The utility of playing FairBot is $d_i + 2c_i + 2f_i + 2p_i - \varepsilon = 2 - d_i - \varepsilon$, where i is the other player.

The utility of playing PrudentBot is $d_i + 3c_i + 2f_i + 2p_i - 2\varepsilon = 1 + 2c_i + f_i + p_i - 2\varepsilon$, where i is the other player.

Note that if one player is indifferent between CooperateBot and DefectBot, this means the other player plays FairBot exactly $\frac{1}{2}$ of the time, but then the first player would be better off playing PrudentBot (as long as $\varepsilon < \frac{1}{4}$). Thus no player can ever randomize between CooperateBot and DefectBot.

If a player doesn't play CooperateBot at all, the other player can't play Prudent Bot, since FairBot is strictly better. Thus, if both players don't play CooperateBot, they both play only DefectBot and FairBot. If this isn't the pure defect solution, it must be because $f_i = \varepsilon$ and $d_i = 1 - \varepsilon$ for both players, which is indeed a (not very good) Nash equilibrium.

If exactly one player plays CooperateBot at all, the first player mixes between CooperateBot and FairBot, while the other player mixes between (at most) DefectBot, FairBot and PrudentBot. The second player can't play FairBot, since CooperateBot would have done strictly better. Thus the first player gets 0 utility for playing CooperateBot, contradicting the fact that it is impossible to get 0 expected utility playing FairBot.

Finally, we have the case where both players play CooperateBot with some probability, and thus neither plays DefectBot at all. If either player did not play FairBot at all, then DefectBot would strictly dominate CooperateBot for the other player, contradicting the fact that both players play CooperateBot. If either player did not play PrudentBot at all, CooperateBot would strictly dominate FairBot, contradicting the fact that both players play FairBot.

Thus, in the only remaining equilibria, both players mix between all of CooperateBot, FairBot, and PrudentBot. Since both players are indifferent between CooperateBot and FairBot, both players must play PrudentBot with probability exactly $\frac{1}{2}$. Since both players are indifferent between FairBot and PrudentBot, both players must play CooperateBot with probability exactly ε. This leaves probability $1 - \frac{3}{2}\varepsilon$ for FairBot, and the result is indeed a Nash equilibrium.

We have a total of three Nash equilibria. All three are symmetric. The first two are bad and both players get the expected utility 1. The last one is good and both players get expected utility $2 - \varepsilon$.

Next, we have to show that this outcome is also an equilibrium in the game where both players can play any modal agent. If another agent were to get utility more than $2 - \varepsilon$ against this $\frac{1}{2}$ PrudentBot, ε CooperateBot, $1 - \frac{3}{2}\varepsilon$ FairBot combination, it would clearly have to be depth 1, since the only depth 0 agents are CooperateBot and DefectBot, and depth 2 PrudentBot already has the perfect behavior against these 3 agents. It would also have to mutually cooperate with FairBot, and defect against CooperateBot.

Suppose a depth 1 agent X provably cooperates with FairBot and defects against CooperateBot. Note that since X is depth 1, $PA + 1$ knows what X(CB) is, and so $PA + 1$ knows that $X(CB) = \bot \neq \top = X(FB)$. However $PA + 1$ also know that $\Box\Box\bot \rightarrow \forall Y \Box(FB(Y) = \top = CB(Y))$. Thus $PA + 1$ knows $\Box\Box\bot \rightarrow X(FB) = X(CB)$, so $PA + 1$ knows $\neg\Box\Box\bot$, contradiction. Therefore, no agent of depth 1 can have the desired behavior, and our good Nash equilibrium is in fact a Nash equilibrium of the game where both players can play any modal agent.

Evolutionary Simulations

Next, we will consider what happens when a large population of modal agents evolves by playing prisoners' dilemmas. We will consider a system consisting only of DefectBot, CooperateBot, FairBot, and PrudentBot. We will consider a simple model where at each time step, each population grows a small amount proportional to the size of that population times the expected utility a member in that population gets by playing a prisoners' dilemma with a random other agent. Notice that in this model, a population that starts out nonzero will remain nonzero, and a population that starts out 0 will remain 0.

Since the above three Nash equilibria were symmetric Nash equilibria, they are also equilibria of this evolutionary system. This is because the expected utility of being each type of agent that appears any nonzero amount is the same. Since this system keeps populations that start at 0 at 0, there are other equilibria too; for example, any point where all the agents are the same is in equilibrium since it cannot change.

We can then ask about the stability of these equilibria. The pure defect one is stable, since if there are only a small number of agents that are not DefectBots, they won't play each other enough to cancel out their modeling penalty. The $1 - \varepsilon$ DefectBot, ε FairBot one is unstable, since the more FairBots there are, the better the FairBots do. Unfortunately, the good equilibrium is also unstable. This is harder to see, but a small perturbation will cause a very slow spiraling outward. (I checked this with simulation, but did not prove it mathematically.)

However, I ran a simulation that seemed to show that for a small ε, and for almost all initial conditions that mix between all four types of agents, the system does not converge, but instead goes in a large cycle, in which the system spends most of its

time with almost all FairBots. Eventually a very small number of CooperateBots climbs out to become non-negligible, then as soon as the CooperateBots are a significant proportion, the PrudentBots come in to exploit them, then the PrudentBots quickly turn into FairBots, and the cycle repeats. Meanwhile, the DefectBots are pushed down into a very small proportion of the population.

I also looked at a second model as follows: The population starts with a small finite number of agents of each type. At each time step, a new agent enters, and permanently becomes the modal agent that maximizes expected utility against the existing population. In this simulation, unless you start with very few FairBots or PrudentBots, you will simply switch between adding FairBots, PrudentBots, and CooperateBots, never adding a new DefectBot.

A more careful study could actually find what the dynamics are mathematically rather than depending on simulations, and can consider systems with more or even all modal agents. There also might be other models that are more justified than the one I used. I expect that most ways of doing it will result in very little defection.

Conclusion

I like this result. I just assumed that DefectBot would be the only Nash equilibrium, and I was surprised that the story had a much happier ending. When I first became suspicious, I was thinking that you could get a good equilibrium out of an infinite collection of different agents, but it turns out you only need three. You can view the CooperateBots as defecting against the FairBots by refusing to pay to punish bad actors, and you can view the FairBots as defecting against the PrudentBots by refusing to pay to punish the non-punishers. You might think you would need to punish arbitrary levels on non-punishers to be in equilibrium, but it turns out that the PrudentBots can get paid exactly enough to remain competitive by exploiting the CooperateBots, and the CooperateBots can get exploited exactly enough to cancel out their unwillingness to model the opponent.

Why Destructive Value Capture?

Previously: Front Row Center

I got a lot of push-back from suggesting that there was a way for theaters to improve their customer experience and value proposition at low cost (get rid of the seats that are so close to the screen they cause neck strain), and that theaters should do that.

The push-back didn't argue that the method wouldn't improve the customer experience at low cost. There were a few who suggested an alternate high-cost solution that improves the experience more (use high-quality and assigned seating at a substantially higher price point), and which some places have implemented. No one argued that, where the higher-cost solution didn't make sense. my incremental suggestion wouldn't improve the customer experience versus status quo, at relatively low cost.

They also didn't raise the reasonable argument that getting people to do things at all, especially slightly non-standard things that might look bad on superficial metrics during the pitch meeting, is hard. People don't think about things, they don't do things, they don't optimize, and so on. One could reasonably argue this isn't worth the effort.

Instead, everyone argued that, unless they were forced to do so, theaters shouldn't implement the suggestion. Because it would cost them money – they couldn't sell those few terrible seats, and forcing people to come early increases ad and concession revenue.

That's interesting. And weird.

The proposition creates value. One comment from Quixote estimates $1.67 in customer time-value is saved in exchange for the loss of $0.10 in ad revenue.

The proposition improves the customer experience. It generates movie-going habits, loyalty and goodwill.

Not implementing the proposition is a destructive value capture. In order to get a little revenue, an order of magnitude more value is destroyed.

Destructive value capture is normal. In order to capture value, *some* value is typically destroyed. But when you're destroying *most* of the value you withdraw from the system, you should be suspicious mistakes are being made. At a minimum, it's worth asking on a deeper level *why* this is happening. What could justify it? What failure mode are we in? How does it come to be, why does it persist, is there a way we can solve it or minimize it? We shouldn't shrug and mutter something about capitalism. We should treat this as a major failure, and brainstorm potential barriers even if they don't apply in this case.

Can't Raise the Price

If you're charging $15 to see a movie, then destroying $1.50 in value to generate $0.15 in additional income, why aren't you just not doing that, and instead charging $15.25 to see the movie?

What might stop this from being a good solution?

What if movie *was free?* Moving from free to not free is a huge change, even if the additional cost is small. This could drive people away and be hugely value destructive.

What if this *introduced an additional collection point?* You'd need to ask someone for money an additional time to make up the additional cost, and that could be value destructive.

What if this *disrupted a standardized price or crosses a key threshold?* Suppose everyone knows that movies cost $15, and there would be a strong reaction against a price of $15.05, because it's different, or because it makes it hard to give exact change.

What if *the market encouraged sorting purely by price?* Imagine a world like with plane tickets, where you go to Kayak or Orbitz or what not, and there is strong default pressure to buy the cheapest tickets without noticing extra charges.

What if *regulation prevented higher prices?* That which is forbidden is not allowed. Price controls often cause perverse reactions.

Those would be good reasons. All clearly do not apply. Movies aren't free (or if you have MoviePass, they would stay free). Movies have a collection point. Movies don't have a standardized prices or a strong price-sorting search mechanism, and prices are rarely at a key threshold.

Other reasons might apply somewhat, but still seem weak.

What if this *would be a price increase and that would be bad?* Thus, the bad event of 'prices went up' could matter even if the new price isn't much different from the old price, so you can't do that often. A tiny increase might be impractical.

That's fair. But the increase could be put into a later, larger increase, or if that's too big a burden, one could wait on implementation until the next price increase.

What if *higher prices decrease customer experience, so they're more expensive than they look?*

I grant this is likely true for some, but the effect size should be small.

What if *this is a pure bad when demand is low, such as at a matinee, and complexity cost prevents price discrimination?*

Again, this seems true but effect size is small. Some places price discriminate by time but the complexity cost stops the majority. So even though removing the seats costs nothing when demand is low, raising the price at those times is net bad.

Would a price increase *send the wrong message?* Would people then worry about the health of your company, or your industry? Would it thus push down stock prices or reduce your ability to raise money?

It might, indeed. It also might do the opposite. I don't think this is what's going on here.

All of that is seeking solutions to the *easy out:* raising prices. Or, if prices are already higher than they should be, lower them to where they should otherwise be, then

raising them back.

Let’s take away that easy out, and say one of the good reasons applied. You can’t raise the price and demand exceeds supply.

This is pretty terrible *even if you don’t then do value capture.* Destructive value *allocation* is bad enough, via making people wait on lines or make commitments or virtue signal or what have you – anything where the auction involves incinerating rather than redistributing the bids, often all-pay auctions at that. One can think of this as balancing supply and demand *by making quality of the supply sufficiently worse.*

Thus we have two mostly distinct problems. We need to pay for the creation and maintenance of nice things without destroying what makes them nice. And we need to do efficient allocation of those nice things, that balances supply and demand and gets the product to the right people.

Letting the price float is the best way to do both, but what happens when you can’t do it? Are we now stuck with terrible seating and massive deadweight loss? What about other situations where the price is stuck? A life lived under advertising’s increasingly long and intrusive shadow? Or worse, the evil bastard children of microtransactions and free to play games?

We seem to be headed that way. I think there are promising answers, which I hope to explore further. That starts with defaulting to price adjustment, and finding creative ways to do price adjustment, and viewing destruction of value as a failure rather than normality or ‘the way of business.’

Logical uncertainty and Mathematical uncertainty

There is a significant difference between uncertainty about mathematical truths in cases where there isn't a known procedure for checking whether a mathematical claim is true or false, versus when there is but you do not have the computational resources to carry it out. Examples of the former include the Collatz and twin prime conjectures, and examples of the latter include whether or not a given large number is a semi-prime, and what the first decimal digit of Graham's number is.

The former should not be called logical uncertainty, because it is about what is true, not about what can be proved; I'll call it mathematical uncertainty instead. The latter really is uncertainty about logic, since we would know that the claim is either proved or refuted by whatever theory we used to prove the algorithm correct, and we would just be uncertain as to which one.

It's well-known that standard probability theory is a poor fit for handling logical uncertainty because it assumes that the probabilities are logically coherent, and uncertainty about what the logical coherence constraints are is exactly what we want to model; there are no possible outcomes in which the truth-value of a decidable sentence is anything other than what it actually is. But this doesn't apply to mathematical uncertainty; we could study the probability distribution over complete theories that we converge to as time goes to infinity, and reason about this probability distribution using ordinary probability theory. Possible sources of evidence about math that could be treated with ordinary probability theory include physical experiments and black-boxed human intuitions. But another important source of evidence about mathematical truth is checking examples, and this cannot be reasoned about in ordinary probability theory because each of the examples is assigned probability 1 or 0 in the limit probability distribution, since we can check it. So just because you can reason about mathematical uncertainty using ordinary probability theory doesn't mean you should.

Logical induction, taken at face value, looks like an attempt at handling mathematical uncertainty, since logical inductors assign probabilities to every sentence, not just sentences known to be decided by the deductive process. But most of the desirable properties of logical inductors that have been proved refer to sequences of decidable sentences, so logical induction only seems potentially valuable for handling logical uncertainty. In fact, the logical induction criterion doesn't even imply anything about what the probabilities of an undecidable sentence converge to, except that it is not 0 or 1.

Another reason not to trust logical induction too much about mathematical uncertainty is that logical induction gets all its evidence from the proofs of one formal system, and there isn't one formal system whose proofs completely account for all sources of evidence about mathematical claims. But it seems to me that correctly characterizing how to handle all sources of evidence about mathematical truths in a way precise enough to be turned into an algorithm would be a quite shockingly huge advance in the philosophy of mathematics, and I don't expect it to happen any time soon.

Fortunately, we might not need to come up with a better way of handling mathematical uncertainty. It seems plausible to me that only logical uncertainty, not mathematical uncertainty more broadly, has any practical use. For instance, in cryptography, an eavesdropper who is uncertain about what a given ciphertext decrypts to has a perfectly good exponential-time algorithm to determine the answer, but lacks the computational resources to run the algorithm. Someone wondering whether their cryptosystem is breakable might phrase their question in terms of existence of an efficient algorithm that accomplishes some task with probability that does not go to 0 as task size approaches infinity, and it might not be clear that the question can be answered by commonly-used powerful axiom systems; but it would be sufficient for practical purposes to know whether there is an efficient algorithm of at most some given size that will accomplish a given computational task with nontrivial probability for tasks of the size that are actually used (or will be used in the near future). This latter question there is a (very computationally expensive) algorithm to determine the answer to. More broadly, physics appears to be computable, so we should expect that in any situation in which we are uncertain about what will happen in the real world due to lack of mathematical knowledge rather than due to lack of physical knowledge, we will have a computation that would resolve our lack of mathematical knowledge if we were capable of running the computation.

Thus it might make sense to give up on assigning probabilities to all sentences, and just try to assign probabilities to sentences that are known to be resolvable by some proof system (for instance, claims about the outputs of programs that are known to halt). Logical induction can easily be modified to do this, by only opening the market for a sentence once the deductive process outputs a proof that the sentence is resolvable. But this alone doesn't offer any improvement in logical induction. The hope is that by not even trying to assign probabilities to sentences that are not known to be resolvable, it might be possible to find ways of assigning probabilities to resolvable sentences in a way that avoids certain computational limitations that are otherwise inevitable. The properties of logical inductors are all asymptotic, with no computable bounds on rate of convergence, and it is impossible to do better than this in general. Asymptotic properties with no computable bounds on rate of convergence are useless for practical purposes, and I don't see any reason we can't do better than this when restricting only to sentences known in advance to be resolvable.

An edge case between logical and mathematical uncertainty is uncertainty about the truth of Σ_1^0 sentences (or equivalently, of Π_1^0 sentences). True Σ_1^0 sentences are provable in Peano Arithmetic, so uncertainty about their truth can be expressed as uncertainty about the existence of a proof, so this could be described as a case of logical uncertainty. However, since false Σ_1^0 sentences aren't necessarily disprovable, handling uncertainty about Σ_1^0 sentences run into similar difficulties as handling uncertainty about more general arithmetical sentences. And since the physical world doesn't appear to be able to do hypercomputation, the same arguments that more general mathematical uncertainty is useless seem to apply to general Σ_1^0 sentences.

This is why I've been characterizing logical uncertainty as uncertainty about sentences known to be resolvable, even though it could be argued that the term should include uncertainty about sentences that are resolvable if true.

Fundamentals of Formalisation Level 3: Set Theoretic Relations and Enumerability

Followup to Fundamentals of Formalisation level 2: Basic Set Theory.

The big ideas:

- Ordered Pairs
- Relations
- Functions
- Enumerability
- Diagonalization

To move to the next level you need to be able to:

- Define functions in terms of relations, relations in terms of ordered pairs, and ordered pairs in terms of sets.
- Define what a one-to-one (or injective) and onto (or surjective) function is. A function that is both is called a one-to-one correspondence (or bijective).
- Prove a function is one-to-one and/or onto.
- Explain the difference between an enumerable and a non-enumerable set.

Why this is important:

- Establishing that a function is one-to-one and/or onto will be important in a myriad of circumstances, including proofs that two sets are of the same size, and is needed in establishing (most) isomorphisms.
- Diagonalization is often used to prove non-enumerability of a set and also it sketches out the boundaries of what is logically possible.

You can find the lesson in our ihatestatistics course. Good luck!

P.S. From now on I will posting these announcements instead of Toon Alfrink.

Physics has laws, the Universe might not

Inspired by http://backreaction.blogspot.com/2018/06/physicist-concludes-there-are-no-laws.html, which dissed this article: https://www.quantamagazine.org/there-are-no-laws-of-physics-theres-only-the-landscape-20180604.

Epistemic status: very raw, likely discussed elsewhere, though in different terms, but feels like has a kernel of usefulness in it.

What does it mean for the universe to be governed by physical laws? What does the term physical law mean? It means that someone knowing that law can predict with some accuracy the state of the universe at some point in the future from its state at the time of observation. Actually, a qualifier is in order. Can predict the observed state of the universe at some point in the future from its observed state at the time of observation. So

laws => predictability

This is more than a one-directional implication, however. What does it mean for something to be predictable? Again, it means that, by observing the state of the universe at some point in time the observer can make a reasonably accurate prediction of the observed state of the universe at some point in the future. Notice the qualifier "observed" again. How can an observer make this prediction? They must have a model of the observed universe ("map of the territory") inside, and use this model ("trace the map") to predict the observed state of the universe at some point in the future. This model can be very simple, "Raarg hold rock. Raarg let go. Rock fall", or more complicated, "In absence of other forces all objects accelerate downward at 9.81 meters per second squared", or even more abstract, "The stress-energy tensor is proportional to the spacetime curvature." But it is a model nonetheless.

When is a model promoted to the status of a law? When it is useful for more than a single case. When the prediction can be made repeatedly in similar but slightly different circumstances using the same model. There is a lot of complexity hiding under the surface of this "simple" statement, but at the end of the day, models are only useful if they can be reused, and thus become patterns, templates for the observers to predict the universe. Thus we have the implication in the other direction:

predictability => laws

Thus the two terms are equivalent, at least in this framework:

predictability <=> laws

Let's restate the definitions, which are admittedly only the first approximation, and may not looks standard:

Predictability: an observer inside the universe can infer the state of the universe at some future point in time, with the accuracy acceptable to the observer.

Physical (or other) laws: reusable models of the universe that are part of the observer and let the universe appear predictable to the observer.

I have been trying really hard to avoid, or at least to minimize, the mind projection fallacy, such as stating that the physical laws are the objective laws of the universe. They might well be, but that would be a next step in modeling the universe, potentially useful, but not minimal.

Returning to the title of this post, "Physics has laws, the Universe might not," what I mean by it is that Physics is one of the sciences that we humans, "observers" call a collection some of those reusable models, and so is in itself an aggregate model. The laws of Physics are the constituent models. On the other hand, the Universe, or "the territory" may or may not have something that Stephen Hawking once phrased as a question:

"*What is it that breathes fire into the equations and makes a universe for them to describe?*"

The minimal answer might be that the cause and effect are reversed here: the universe just exists (assuming it does), and is somewhat predictable, and the equations are those physical laws inside the observers' minds.

Now, the above is, of course, another (meta-)model. And models are not very useful if they do not result in better predictions. Or, in the language of this site, the beliefs must pay rent. So, what does this one predict? Well, for example, if the universe has no "internal" laws, according to this model, then one could potentially generate a "toy universe," possibly with some form of predictability but without any preset laws and see if anything that could be called (toy) laws would "emerge" in this toy universe, and under what conditions. It would be interesting to explore this approach further, but it requires a fair amount of decomposition and analysis to make sense in more than a handwavy way. Here are some questions that come to mind:

- Can one start with a sequence of random numbers as a toy universe, i.e. no order and get somewhere that way, just by finding spurious patterns in the sequence?
- An observer is a part of the universe, how would a sub-sequence of numbers represent an observer?
- A "physical law" in this toy universe would be a part of the observer, or maybe the even the whole observer, that can be considered a reusable template, the way our physical laws are. How might it be represented in this case?
- Can the above conditions be relaxed enough to apply to a toy model, but still be offer useful insights?

Should an experiment like that worked out, it would lend some credence to the above conjecture, that the physical laws are not some inherent property of the Universe, but a human attempt to make sense of it by creating reusable templates inside themselves.

Simplified Poker Strategy

Previously in Sequence (Required): Simplified Poker

I spent a few hours figuring out my strategy. This is what I submitted.

If you start with a 2, you never want to bet, since your opponent will call with a 3 but fold with a 1. So we can assume no one who bets ever has a 2. But you might want to call a bet.

If you start with a 1, you never call a bet, but sometimes want to bet as a bluff.

If you start with a 3 in first position, sometimes you may want to check to allow your opponent to bet with a 1. If you have a 3 in second position, you have no decisions.

Thus, a non-dominated strategy can be represented by five probabilities: The chance you bet with a 1 in first position, chance you bet with a 3 in first position, chance you bet with a 1 in second position, chance you call with a 2 in first position, and chance you call with a 2 in second position. Call a set of these five numbers a strategy.

There were likely to be a few players bad enough to bet with a 2 or perhaps make the other mistakes, but I chose for complexity reasons not to worry about that, assuming I'd still do something close to optimal. If I was confident complexity was free, I'd have included a check to see if we ever caught the opponent doing something crazy, and adjust accordingly.

If you know the opposing strategy, what to do is obvious. Thus, I defined a function called 'best response' that takes a strategy, and outputs the strategy that maximizes against that strategy.

My goal was to derive the opponents' strategy, then play the best response to that strategy.

As a safeguard against opponents who were anticipating such a strategy, I included an escape hatch: If at any point, my opponent got ahead by 10 or more chips, assume they were a level ahead of me, and playing the best response to what I would otherwise do. So derive what that is, and play the best response to that!

That skipped over the key puzzle, which is figuring out what the opponent is doing. On the first turn, I guessed opponents would pursue reasonable mixed strategies: bet a 1 about a third of the time, bet a 3 in first position about two thirds of the time, call with a 2 about half the time. I represented this with a virtual hand history that I included until I had enough real ones.

On subsequent turns, I looked at the hand history.

If the opponents' card was revealed, that was a pure data point – if we knew they bet with a 1, that's a hand where they did that.

If the opponents' card wasn't revealed, but only one card made any sense, I assumed they had that card. Thus, if I bet with a 1 and they fold, I assume they had a 2.

If the opponents' card wasn't revealed, and they could have had either card because you bet a 3 and they folded, or they bet and you folded a 2, that's trickier. The probability of them having each card in that spot depends on their strategy. And again, there was a (unknown soft) complexity limit.

My solution was to assume that in each unique starting position (your position plus your card) half the time my opponent would draw the higher of the two cards I hadn't drawn, and half the time he'd draw the lower one. So half the time I have a 2 in first position, he has a 3, half the time he has a 1.

That was definitely not ideal, and I don't remember *exactly* how I did it, but it definitely did the thing it was designed to do: Identify exploitable agents lightning fast, and do something reasonable against reasonable ones. Trying to optimize the details of this type of approach is an interesting puzzle, both with and without a complexity limitation.

Optimization Amplifies

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

I talk here about how a mathematician mindset can be useful for AI alignment. But first, a puzzle:

Given m, what is the least number $n \geq 2$ such that for $2 \leq k \leq m$, the base k representation of n consists entirely of 0s and 1s?

If you want to think about it yourself, stop reading.

For m=2, n=2.

For m=3, n=3.

For m=4, n=4.

For m=5, n=82,000.

Indeed, 82,000 is 10100000001010000 in binary, 11011111001 in ternary, 110001100 in base 4, and 10111000 in base 5.

What about when m=6?

So, a mathematician might tell you that this is an open problem. It is not known if there is any $n \geq 2$ which consists of 0s and 1s in bases 2 through 6.

A scientist, on the other hand, might just tell you that clearly no such number exists. There are 2^{k-1} numbers that consist of k 0s and 1s in base 6. Each of these has roughly $\log_5(6) \cdot k$ digits in base 5, and assuming things are roughly evenly distributed, each of these digits is a 0 or a 1 with "probability" 2/5. The "probability" that there is any number of length k that has the property is thus less than $2^k \cdot (2/5)^k = (4/5)^k$. This means that as you increase k, the "probability" that you find a number with the property drops off exponentially, and this is not even considering bases 3 and 4. Also, we have checked all numbers up to 2000 digits. No number with this property exists.

Who is right?

Well, they are both right. If you want to have fun playing games with proofs, you can consider it an open problem and try to prove it. If you want to get the right answer,

just listen to the scientist. If you have to choose between destroying the world with a 1% probability and destroying the world if a number greater than 2 which consists of 0s and 1s in bases 2 through 6 exists, go with the latter.

It is tempting to say that we might be in a situation similar to this. We need to figure out how to make safe AI, and we maybe don't have that much time. Maybe we need to run experiments, and figure out what is true about what we should do and not waste our time with math. Then why are the folks at MIRI doing all this pure math stuff, and why does CHAI talk about "proofs" of desired AI properties? It would seem that if the end of the world is at stake, we need scientists, not mathematicians.

I would agree with the above sentiment if we were averting an astroid, or a plague, or global warming, but I think it fails to apply to AI alignment. This is because optimization amplifies things.

As a simple example of optimization, let X_i for $i < 1,000,000$ be i.i.d. random numbers which are normally distributed with mean 0 and standard deviation 1. If I choose an X_i at random, the probability that X_i is greater than 4 is like 0.006%. However, if I optimize, and choose the greatest X_i, the probability that it is greater that 4 is very close to 100%. This is the kind of thing that optimization does. It searches through a bunch of options, and takes extreme ones. This has the effect of making things that would be very small probabilities much larger.

Optimization also leads to very steep phase shifts, because it can send something on one side of a threshold to one extreme, and send things on the other side of a threshold to another extreme. Let X_i for $i < 1,000,000$ be i.i.d. random numbers that are uniform in the unit interval. If you look at the first 10 numbers and take the one that is furthest away from .499, the distribution over numbers will be bimodal peaks near 0 and 1. If you take the one that is furthest away from .501, you will get a very similar distribution. Now instead consider what happens if you look at all $1,000,000$ numbers and take the one that is furthest from .499. You will get a distribution that is almost certainly 1. On the other hand, the one that is furthest from .501 will be almost certainly 0. As you slightly change the optimization target, the result of a weak optimization might not change much, but the result of a strong one can change things drastically.

As a very rough approximation, a scientist is good at telling the difference between probability 0.01% and probability 99.99%, while the mathematician is good at telling the difference between 99.99% and 100%. Similarly, the scientist is good at telling if $a \simeq b$, while the mathematician is good at telling if $a = b$ when you already know that $a \simeq b$.

If you only want to get an approximately correct answer almost surely, the absence of strong optimization pressure makes the mathematician skills much less useful. However strong optimization pressure amplifies and creates discontinuities, which creates the necessity for a mathematician level of precision even to achieve approximate correctness in practice.

Notes:

1) I am not just saying that adversarial optimization makes small probabilities of failure large. I am saying that in general any optimization at all messes with small probabilities and errors drastically.

2) I am not saying that we don't need scientists. I am saying that we don't just need scientists, and I am saying that scientists should pay some attention to the mathematician mindset. There is a lot to be gained from getting your hands dirty in experiments.

3) I am not saying that we should only be satisfied if we achieve certainty that an AI system will be safe. That's an impossibly high standard. I am saying that we should aim for a deep formal understanding of what is going on, more like the "fully reduced" understanding we have of steam engines or rockets.

The Beauty and the Prince

This post will address a problem proposed by Radford Neal in his paper Puzzles of Anthropic Reasoning Resolved Using Full Non-indexical conditioning. In particular, he defined this problem - The Beauty and the Prince - to argue against the halver solution to the Sleeping Beauty Problem. I don't think that this is ultimately a counter-example, but I decided to dedicate a post to it because I felt that it was quite persuasive when I first saw it. I'll limit the scope of this post to arguing that his analysis of the halver solution is incorrect and providing a correct analysis instead. I won't try to justify the halver solution as being philosophically correct as I plan to write another post on the Anthropic Principle later, just show how it applies here.

The Beauty and the Prince is just like the Sleeping Beauty Problem, but with a Prince who is also interviewed and memory-wiped. However, he is always interviewed on both Monday and Tuesday regardless of what the coin shows and he is told whether or not Sleeping Beauty is awake. If he is told that she is awake, what is the probability that the coin came up heads. The argument is that 3/4 times she will be awake and 1/4 times she is asleep so only 1/3 times when he is told she is awake will the coin be heads. Further, it seems that Sleeping Beauty should adopt the same odds as him. They both have the same information, so if he tells her the odds are 1/3, on what basis can she disagree? Further, she knows what he will say before he even asks her.

I want to propose that the Prince's probability estimate as above is correct, but it is different from Sleeping Beauty's. I think the key here is to realised that indexicals aren't a part of standard probability, so we need to de-indexicalise the situation. However, we'll de-indexicalise the original problem first. We'll do this be ensuring that only one interview ever "counts", by which we mean that we will calculate the probability of events over the interviews that count. We'll do this by flipping a second coin if the first comes up tails. If it is heads, only the first interview counts, whilst for tails only the second interview counts. We then get the odds being: 1/2 heads + Monday counts; 1/4 tails + Monday counts; 1/4 tails + Tuesday counts.

We similarly de-indexalise the Prince, though we flip the second coin in the heads case too. Similarly, if it is heads we count the interview on Monday and if it is tails we count the interview on Tuesday, so the four possibilities become mutually exclusive and each have a probability of 25%.

If we look when the first coin is heads, we notice that the Prince's interview on Monday only counts 50% of the time, whilst Sleeping Beauty's counts 100% of the time. This means that Sleeping Beauty is calculating her probability over a different event space so we should actually expect her answer to differ from that of the Prince. Suppose we expand the Prince's probability to include the Sleeping Beauty's Monday interviews (which all count). Then we get the chance of heads moving from 1:2 = 1/3 to 2:2 = 1/2.

As we've seen, The Beauty and the Prince is not a problem for the halver solution. This does not mean that the halver solution is the correct solution to the Sleeping Beauty Problem, just that The Beauty and The Prince doesn't provide a counter-example.

Update: I've been reading more of the literature. It seems that the technique that I'm using here is actually closer to what Bostrom call the Hybrid Model, then David Lewis'

Halver Solution. The difference is that if you are told it is Monday, Bostrom gets heads being 1/2, while Lewis gets heads being 2/3.

The Curious Prisoner Puzzle

Here's an interesting riddle that is more complicated than it looks:

You wake up locked inside a room with no window. You know your captors have four facilities in the following locations: a Vulcan Mountain, a Vulcan Desert, an Earth Mountain, an Earth Desert. They flip a coin to decide which planet to send you to, then they flip a coin to decide which facility on that planet to use.

Wanting to know where you are, you try to get some information out of the guard. He refuses at first, but eventually he offers the following: "If you are on Vulcan, you are in the Mountain". What are the chances that you are in the Vulcan Mountain facility?

(You can assume that the guard is telling the truth and not trying to intentionally manipulate the situation)

Shaping economic incentives for collaborative AGI

In "An AI Race for Strategic Advantage: Rhetoric and Risks" (2018), Stephen Cave and Seán S ÓhÉigeartaigh argue that we should try to promote a cooperative AI narrative over a competitive one:

> The next decade will see AI applied in an increasingly integral way to safety-critical systems; healthcare, transport, infrastructure to name a few. In order to realise these benefits as quickly and safely as possible, sharing of research, datasets, and best practices will be critical. For example, to ensure the safety of autonomous cars, pooling expertise and datasets on vehicle performances across as wide as possible a range of environments and conditions (including accidents and near-accidents) would provide substantial benefits for all involved. This is particularly so given that the research, data, and testing needed to refine and ensure the safety of such systems before deployment may be considerably more costly and time-consuming than the research needed to develop the initial technological capability.
>
> Promoting recognition that deep cooperation of this nature is needed to deliver the benefits of AI robustly may be a powerful tool in dispelling a 'technological race' narrative; and a 'cooperation for safe AI' framing is likely to become increasingly important as more powerful and broadly capable AI systems are developed and deployed. [...]
>
> There have been encouraging developments promoting the above narratives in recent years. 'AI for global benefit' is perhaps best exemplified by the 2017's ITU summit on AI for Global Good (Butler 2017), although it also features prominently in narratives being put forward by the IEEE's Ethically Aligned Design process (IEEE 2016), the Partnership on AI, and programmes and materials put forward by Microsoft, DeepMind and other leading companies. Collaboration on AI in safety-critical settings is also a thematic pillar for the Partnership on AI2 . Even more ambitious cooperative projects have been proposed by others, for example the call for a 'CERN for AI' from Professor Gary Marcus, through which participants "share their results with the world, rather than restricting them to a single country or corporation" (Marcus 2017).

So.

In order to make future AGI projects more collaborative and co-operation focused, could we create incentives (via e.g. government policy) that would push *today's* machine learning researchers towards more collaborative attitudes?

This might seem irrelevant, given that today's machine learning researchers are mostly not working on AGI. However, external incentives can shape the internal norms of a culture. For example, holding companies responsible for accidents at their workplaces, means that they have an incentive to reduce accidents, which means that they have an incentive to create an internal culture of safety where everyone takes safety concerns seriously. And once such a culture is established, it will start having a life of its own, being propagated to future workers through the various sociological mechanisms by which norms and cultures normally propagate themselves, and may

stay alive even if there's a change to the external norms which led to that culture being originally created.

So my idea is something like:

- figure out the kinds of external incentives that would affect machine learning companies and research that's happening today, pushing it in a more collaborative direction
- implementing these kinds of incentives via the right policy, will cause the field to more generally adopt the kinds of values and norms where collaboration is seen as a good thing
- to the extent that the field which ends up developing AGI is a descendant of the field that does AI research today, the collaborative norms and values of today's field will be inherited by that future field, shifting their prevailing attitudes away from "arms race" framings and increasing the chances of AGI being developed collaboratively

In a discussion, James Miller suggested that - among other things - codes of conduct, intellectual property laws, antitrust laws, tort law, and international agreements/tariffs might be policy tools which could be used to shape external incentives.

A possible addition that comes to mind might be privacy laws; at least current ML systems require a lot of data, and there have been a lot of demands (e.g.) to reign in the ability of companies to collect information on people - information which could, among other things, be used to train ML systems. And e.g. the GDPR (which might be enforced more strictly after the recent Facebook revelations) establishes things like "*Automated individual decision-making, including profiling [...] is contestable [...] Citizens have rights to question and fight significant decisions that affect them that have been made on a solely-algorithmic basis*"; to the extent that decisions made by algorithms can be contested by the people who are affected by them, companies may have an incentive to be cooperative and e.g. develop the kinds of standards that they can follow in order to ensure that decisions made by their systems will be held up in court. (Doshi-Velez et al. (2017) is a paper attempting to establish some kinds of standards for how a legal right to explanation from AI systems could be met.)

Some other thoughts:

It might be worth thinking about a more specific definition for "cooperativeness". For instance, one form of "cooperativeness" might be openness in AI development. Openness seems worth distinguishing from other forms of cooperation, since while general cooperativeness may make things safer, openness may make them less safe. But I would intuitively think that non-openness would be hard to reconcile with cooperativeness. Maybe it's unavoidable for cooperativeness to lead to at least some degree of openness. (Bostrom (2017) notes on page 9 that openness could make AI development more competitive, but also more cooperative, if it removes incentives for competition: "*The more that different potential AI developers (and their backers) feel that they would fully share in the benefits of AI even if they lose the race to develop AI first, the less motive they have for prioritizing speed over safety, and the easier it should be for them to cooperate with other parties to pursue a safe and peaceful course of development of advanced AI designed to serve the common good.*")

As Baum (2017) points out, it's important to consider how AI developer communities react to external rules: if e.g. safety regulations are viewed as

pointless annoyances, that may cause a lot of resentment. And it's easy to adopt a patronizing mindset in thinking about this: "how could we get AI developers to understand that they shouldn't destroy the world?". We shouldn't think about this that way (that's not a particularly collaborative mindset ☺).

Rather, the better mindset is something like this: most people don't want to destroy the world, AI developers included. But it's easy to end up in situations where everyone has a rational incentive to do something that nobody wants. So what we want is to collaboratively design mechanisms that end up *supporting* people in better fulfilling their *own* preference of not destroying the world.

(thanks to James Miller as well as my colleagues at the Foundational Research Institute for discussions that contributed to this article)

Front Row Center

Epistemic Status: Lightweight

Related: Choices are Bad, Choices Are Really Bad

Yesterday, my wife and I went out to see Ocean's Eight (official review: as advertised). The first place we went was a massively overpriced theater (thanks MoviePass!) with assigned seating, but they were sold out (thanks MoviePass!) so we instead went to a different overpriced theater without assigned seating, and got tickets for a later show. We had some time, so we had a nice walk and came back for the show.

When we got back, there was nowhere for us to sit together outside of the first two rows. They're too close, up where you have to strain your neck to see the screen. My wife took the last seat we could find a few rows behind that, and I got a seat in the second row. It was fine, but I'd have much preferred to sit together.

It was, of course, our fault for showing up on time rather than early to a sold out screening. I mention it because it's a clean example of how offering less can provide more value.

The theater should, if they don't want to do assigned seating, rip out the first two rows.

At first this seems crazy. Many people prefer sitting in the first two rows to being unable to attend the show, so the seats create value while increasing profits. What's the harm?

The harm is introducing risk, and creating an expensive auction.

The risk is that if you go to the movies, especially the movies you most want to see, you'll be stuck in the first two rows. So when you buy a ticket and go upstairs, you might get a bad experience. If the show is sold out, that might be better, as you can buy a different ticket or none at all.

The auction is worse. Seats are first come, first serve. So if it's important to get served first, you need to come first. If it's very important to *not be last,* to avoid awful seats, you need to come early, and so does everyone else, bidding up the price of not-last the same way you'd bid up being first.

With no awful seats, those who care a lot about better seats will still come early, but most people care a lot less. So everyone can come substantially earlier, and not feel pressure. Many will show at the last minute, and be totally fine.

The deadweight loss in time, of adding those forty extra seats, is massive, distributed throughout the theater. Everyone feels pressure to get there early even when they already have a ticket, so even if their seat is good, they stressed out about their seat, and not only burned time but feel bad about being pressured.

Avoiding time-based auctions and signals, or at least minimizing the value at stake in them, is an important and underappreciated problem.

What could be done with RNA and DNA sequencing that's 1000x cheaper than it's now?

In How to Invent the Future Alan Kay proposes that a good way of inventing the feature is to take an exponential and project it into the feature. Afterwards it's useful to ask what can be done with the new capabilities.

On important exponential is the price of DNA and RNA sequencing that fell faster than Moore's law over the last 30 years.

What can be done in a future is radically cheaper sequencing?

Wirehead your Chickens

TL;DR: **If you care about farm animal welfare, work on minimizing actual animal suffering, not a human proxy for animal suffering.**

Epistemic status: had a chat about this with a couple of local EA enthusiasts who attended the EA Global 2018 in San-Francisco, and apparently this was not high on the agenda. I have only done a cursory search online about this, and nothing of note came up.

When you read about farm animal welfare, what generally comes up is vegetarianism/veganism, humane treatment of farm animals, and sometimes vat-grown meat. This emphasis is quite understandable emotionally. Cows, pigs, chickens in industrial farms are in visible severe discomfort most of their lives, which are eventually cut short long before the end of their natural lifespan, often in painful and gruesome ways.

An animal welfare activist would ask themselves a question like "what is it like to be a chicken in a chicken farm?" and end up horrified. Their obvious solutions are those outlined above: have fewer farm animals and treat them "humanely." Less conventional approaches that reduce animal suffering get an immediate instinctive pushback, because we would not find them acceptable for ourselves. This is what I call the human proxy for animal suffering. Maybe there is a more standard name for this kind of anthropomorphizing? Anyway, let's list a few obvious approaches:

- breed chickens with smaller brains, so they have less capacity for suffering,
- inject a substance that would numb farm animals to physical pain,
- identify and surgically or chemically remove the part of the brain that is responsible for suffering,
- at birth, amputate the non-essential body parts that would give the animals discomfort later in life,
- breed animals who enjoy pain, not suffer from it,
- breed animals that want to be eaten, like the Ameglian Major Cow from the Hitchhiker's Guide to the Galaxy.

Many of these are probably way easier and more practical than shaming people into giving up tasty steak. But our morality immediately fights back, at least for most of us. "What do you mean, cut off baby chicken's legs so it does not have leg pain later? You, monster!"

Because most people do not truly care about reducing animal suffering, they care about a different metric altogether, a visible human proxy for animal suffering that they find immediately relatable. And so it appears that there is virtually no research or funding into the real suffering reduction, even though we know those will work. Because they work on humans already. Drug addicts are quite happy while under influence. Epidural works wonders for temporary pain removal, and so does spinal cord injury in many cases. The list of proven but not ethically acceptable ways to reduce suffering in humans is pretty long.

If you are an effective altruist who is concerned with farm animal welfare, what is stopping you from working on finding ways to apply what works but is not ethical for humans to what works and reduces actual suffering in animals?

Counterfactual Mugging Poker Game

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

Consider the following game:

Player A receives a card at random that is either High or Low. He may reveal his card if he wishes.

Player B then chooses a probability p that Player A has a high card.

Player A always loses p^2 dollars. Player B loses p^2 dollars if the card is low and $(1-p)^2$ dollars if the card is high.

Note that Player B has been given a proper scoring rule, and so is incentivized to give his true probability (unless he makes some deal with player A).

You are playing this game as player A. You only play one time. You are looking at a low card. Player B is not trying to make a deal with you, and will report his true probability. Player B is very good at reasoning about you, but you are in a separate room, so Player B cannot read any tells unless you show the card. Do you show your card?

Since your card is low, if you show it to player B, you will lose nothing, and get the best possible output. However, if player B reasons that if you would show your card if it was low, then in the counterfactual world in which you got a high card, player B would know you had a high card because you refused to show. Thus, you would lose a full dollar in those counterfactual worlds.

If you choose to not reveal your card, player B would assign probability 1/2 and you would lose a quarter.

I like this variant of the counterfactual mugging because it takes the agency out of the predictor. In the standard counterfactual mugging, you might reject the hypothetical and think that the predictor is trying to trick you. Here, there is a sense in which you are creating the counterfactual mugging yourself by trying to be able to keep secrets.

Also, think about this example the next time you are tempted to say that someone would only Glomarize if they had an important secret.

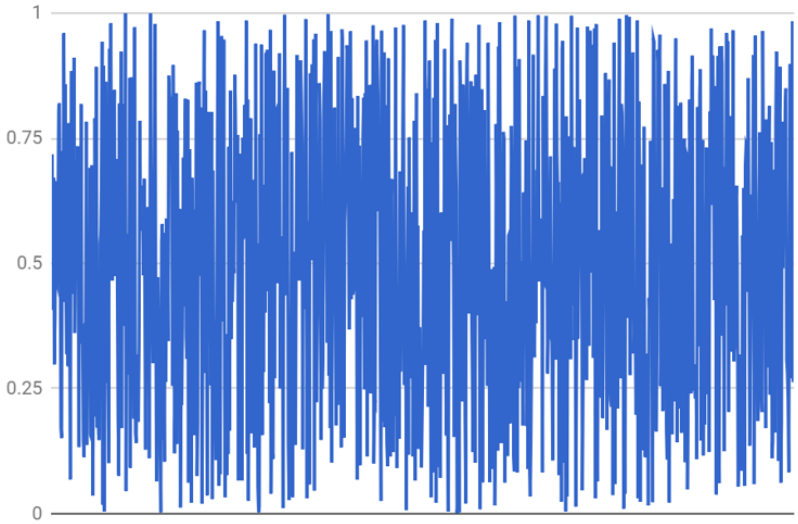
Order from Randomness: Ordering the Universe of Random Numbers

Epistemic status: not sure what to make of it.

Previously I had suggested that the laws of physics were the observers' attempts to make sense of the universe without laws while looking for patterns, which then become their models of reality. Some people suggested that this idea matched their intuition, others were bringing up Tegmark's mathematical universe as something that inherently has laws in it. Not being a fan of Tegmark (that's a different discussion) I will not pursue this avenue. Returning to the questions I have asked in the previous post, the first one was:

Can one start with a sequence of random numbers as a toy universe, i.e. no order and get somewhere that way, just by finding spurious patterns in the sequence?

This post attempts to create a toy model that does just that! it might appear contrived, but hopefully will make a bit of sense. So, let's have our toy model of the universe as a sequence of random numbers. For definiteness, let's start with a uniform random sequence with elements between zero and one. While it is possible to get more random than that, it would require some effort. So, here is a sample sequence:



This looks, appropriately, rather orderless. One can see that this is so by looking at the power spectrum of this sequence. I have averaged about 10,000 of power spectra to get a smooth curve:

There are no correlations in this sequence and the power spectrum is flat, indicating the lack of order, as it should be. The random sequence was our toy universe, with the element (sample) number interpreted as the discrete "time" in this universe, t=1..1024.

But why let the time match the sample number? What is time, anyway? That has been, well, a timeless question for several thousands of years. Why do most of us, humans, perceive time as one-dimensional, inexorably going from the past to the future, and not in the other direction, or without any direction at all? A common answer to this question is related to the second law of thermodynamics:

The arrow of time matches the arrow of increasing entropy.

Well, not quite. The law, as generally formulated, states that ***entropy increases with time***. But how do we know that this is a physical law and not a pattern we notice and successfully extrapolate, in keeping with the main thesis under consideration? Let's see what we can get from our toy model of a universe of random numbers. As remarked previously, the perceived passage of time is an open problem, so, until it is resolved, we are free to assign time values to our samples any way we wish.

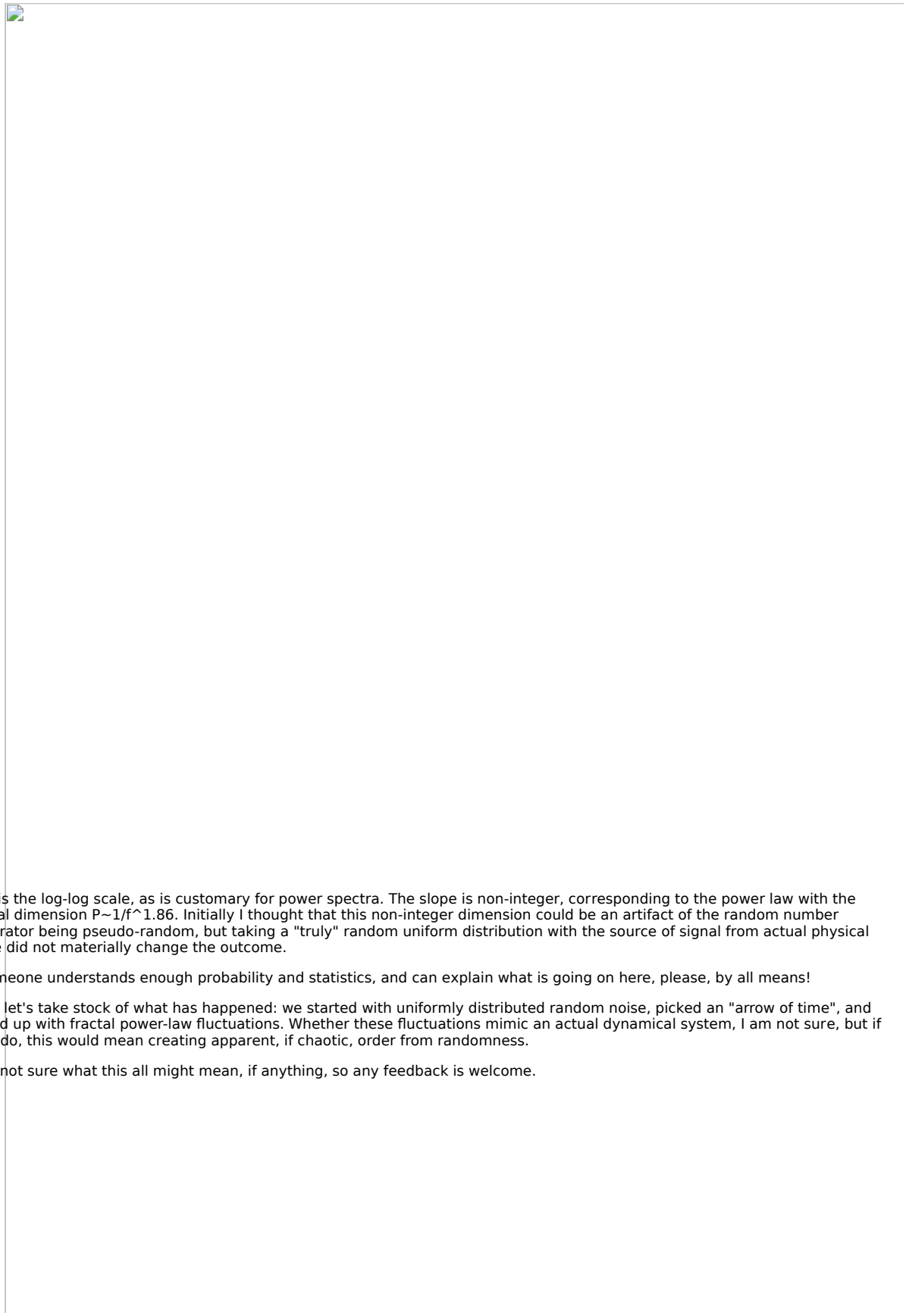
Now comes the controversial part.

How do we find something like entropy in this toy model? After all, there aren't any macrostates we can average over some microstates. The lack of order is intentionally built into this universe from the beginning. All we have is a bunch of numbers. So, let us use the magnitude of each number as a poor-man's proxy for entropy, and assign the later time to larger numbers. In other words, let's sort the random sequence! Here is a sample sorted sequence:

So, if we assign it in the ascending order, larger numbers corresponding to the later times, then, from the point of view of this hypothetical observer the universe would be quite predictable: the only quantity that exists in this universe monotonically increases with the perceived time.

In this ordered random universe an observer can definitely make at least one prediction, the same one we built into this ordering: that time goes forward. This is almost tautological. Let us try and see what other observables we can find here. One obvious thing to notice is that the line is not perfectly straight but has fluctuations. So let's subtract the trend and see what the fluctuations are like. Here is a picture for one run:

This is definitely not the uniform random noise anymore. To see what it is, let's look at the power spectrum again, averaged over 100,000 samples. I have included the power spectrum of the uniform distribution for comparison:



This is the log-log scale, as is customary for power spectra. The slope is non-integer, corresponding to the power law with the fractal dimension P~1/f^1.86. Initially I thought that this non-integer dimension could be an artifact of the random number generator being pseudo-random, but taking a "truly" random uniform distribution with the source of signal from actual physical noise did not materially change the outcome.

If someone understands enough probability and statistics, and can explain what is going on here, please, by all means!

Now, let's take stock of what has happened: we started with uniformly distributed random noise, picked an "arrow of time", and ended up with fractal power-law fluctuations. Whether these fluctuations mimic an actual dynamical system, I am not sure, but if they do, this would mean creating apparent, if chaotic, order from randomness.

I am not sure what this all might mean, if anything, so any feedback is welcome.

Sleeping Beauty Not Resolved

Ksvanhorn recently suggested that Radford Neal provides the solution to the Sleeping Beauty problem and that the current solutions are wrong. The consensus seems to be that the maths checks out, yet there are strong suspicions that something fishy is going on without people being able to fully articulate the issues.

My position is that Neal/Ksavnhorn make some good critiques of existing solutions, but their solution falls short. Specifically, I'm not going to dispute Neal's calculation, just that the probability he calculates completely misses anything that this problem may reasonably be trying to get at. Part of this is related to the classic problem, "If I have two sons and at least one of them is born on a Tuesday, what is the chance that both are born on the same day". This leads me to strongly disfavour the way Neal extends the word probability to cover these cases, though of course I can't actually say that he is wrong in any objective sense since words don't have objective meanings.

A key part of his argument can be paraphrased as a suggestion to Shut up and multiply since verbal arguments about probability have a strong tendency to be misleading. I've run into these issues with the slipperiness of words myself, but at the same time, we need verbal arguments to decide why it is that we should construct the formalism a particular way. Further, my interpretation of Shut up and multiply has always been that that we shouldn't engage in moral grandstanding by substituting emotions for logic. I didn't take it to mean that when presented with a mathematical model that seems to produce sketchy results that we should accept it unquestioningly, without taking the time to understand the assumptions behind it or its intended purpose. Indeed, we've been told to shut and multiple for sleeping beauty before and that came to a different conclusion.

What is Neal actually doing

Unfortunately, Ksavnhorn's post jumps straight into the maths and doesn't provide any explanation of what is going on. This makes it somewhat harder to critique, as it means that you don't just need the ability to follow the maths, but also the ability to figure out the actual motivation behind all of this.

Neal wants us the condition on all information, including the apparently random experiences that Sleeping Beauty will undergo before they answer the interview question. This information seems irrelevant, but Neal argues that if it were irrelevant that it wouldn't affect the calculation. If, contrary to expectations, it actually does, then Neal would suggest that we were wrong about its irrelevance. On the other hand, I would suggest that this is a massive red flag that suggests that we don't actually know what it is that we are calculating, as we will see in a moment.

Let S refer to experiencing a particular sequence of sensations, starting with waking up and ending with being interviewed. Neal's strategy is to calculate the probability of S given heads and the probability of S given tails. If we are woken twice, we only need to observe S on at least one of the days for it to count and if we observe S on both days, it still only counts once. Neal uses Bayes' Rule on the intermediate probabilities to discover the probability of heads vs. tails. Notice how incredibly simple this process was to describe in words. This is one of those situations where preventing the formalisms without an intuitive explanation of what is happening makes it much harder to understand.

Calculations

I aim to show that intermediate probabilities are mostly irrelevant. In order to do so, we will assume that there are three bits of information after awakening and before the interview (this implies 8 possibilities). Let's suppose Sleeping Beauty awakes and then observes the sequences 111. Neal notes that in the heads case, the chance of Sleeping Beauty observing this at least once is 1/8. In the tails case, assuming independence, we get a probability 1/8+1/8-1/64 = 15/64 (or almost 1/4). As the number of possibilities approaches infinity, the ratio of the two probabilities approaches 1:2, which leads to slightly more than a 1/3 chance of heads after we perform the Bayesian update (see Ksvanhorn's post for the maths). If we ensure that the experience stream of the second awakening never matches that of the first, we get a 2/8 chance of observing 111 in one of the two streams, which eventually leads to exactly a 1/3 chance. One the other hand, if the experience stream is always the same both before and after, we get a 1/8 chance of observing 111. This provides a ratio of 1:1, which leads to a 50% chance of heads.

All this maths is correct, but why do we care about these odds? It is indeed true that if you had pre-committed at the start to guess if and only if you experienced the sequence 111, then the odds of the coin being heads would be as above. This would be also true if you made the same commitment for 000 instead; or 100; or any sequence.

However, let's suppose you picked two sequences 000 and 001 and pre-committed to guess if you saw either of those sequences. Then the odds of guessing if tails occurs and the observations are independent would become: 1/4+1/4-1/16 = 7/16. This would lead the probability ratio to become 4/7. Now, the other two probabilities (always different, always the same) remain the same, but the point is that the probability of heads depends on the number of sequences you pre-commit to guess. If you pre-committed to guess regardless of the sequence, then the probability becomes 1/2.

Moving back to the original problem, suppose you wake up and observe 111. Why do we care about the odds of heads if you had pre-committed to only answering on observing 111, given that you didn't pre-commit to this at all? Further, there's no reason why you couldn't, for example decide in advance to ignore the last bit and pre-commit if the first two were 11. Why must you pre-commit utilising all of the available randomness? Being able to manipulate your effective odds in this way by making such pre-commitments is a neat trick, but it doesn't directly answer the question asked. Sure Ksvanhorn was able to massage this probability to produce the correct betting decisions, but both the halfer and thirder solutions can achieve this much easier.

Updating on a random bit of information

@travisrm89 wrote:

> How can receiving a random bit cause Beauty to update her probability, as in the case where Beauty is an AI? If Beauty already knows that she will update her probability no matter what bit she receives, then shouldn't she already update her probability before receiving the bit?

Ksvanhorn responds by pointing out that this assumes that the probabilities add to one, while we are considering the probability of observing a particular sequence at least once, so these probabilities overlap.

This doesn't really clarify what is going on, but I think that we can clarify this by first looking at the following classical probability problem:

> A man has two sons. What is the chance that both of them are born on the same day if at least one of them is born on a Tuesday?

(Clarifying in response to comments: the Tuesday problem is ambiguous and the answer is either 1/13 or 1/7 depending on interpretation. I'm not disputing this)

Most people expect the answer to be 1/7, but the usual answer is that 13/49 possibilities have at least one born on a Tuesday and 1/49 has both born on Tuesday, so the chance in 1/13. Notice that if we had been told, for example, that one of them was born on a Wednesday we would have updated to 1/13 as well. So our odds can always update in the same way on a random piece of information if the possibilities referred to aren't exclusive as Ksvanhorn claims.

However, consider the following similar problem:

> A man has two sons. We ask one of them at random which day they were born and they tell us Tuesday. What is the chance that they are both born on the same day?

Here the answer is 1/7 as we've been given no information about when the other child was born. When Sleeping Beauty wakes up and observes a sequence, they are learning that this sequence occurs on a on a random day out of those days when they are awake. This probability is 1/n where n is the number of possibilities. This is distinct from learning that the sequence occurs in at least one wakeup just like learning a random child is born on a Tuesday is different from learning that at least one child was born on a Tuesday. So Ksvanhorn has calculated the wrong thing.

What does this mean?

Perhaps this still indicates a limitation on the thirders' attempts to define a notion of subjective probability? If we define probability in terms of bets, then this effect is mostly irrelevant. It only occurs when multiple guesses are collapsed down to one guess, but how often will a situation involving completely isolated situations be scored in a combined manner?

On the other hand, what does this mean for the halvers' notion of probability where we normalise multiple guesses? Well, it shows that we can manipulate the effective probability of heads vs. tails via only guessing in particular circumstances, however, the effect is purely a result of controlling how many times we guess correctly on both days so that they only count once. Further, there are many situations where we make the correct guess on Monday, then refuse to guess on Tuesday or vice versa.

These kinds of situations fit quite awkwardly into probability theory and it seems much more logical to consider handling them in the decision theory instead.

More on the 1/3 solution

Neal is correct to point out that epistemic probability theory doesn't contain a concept of "now", so we either need to eliminate it (such as by using indexicals) or utilitise an extension of standard probability theory. Neal is correct that most 1/3 answers skip over this work and that this work is necessary for a formal proof. I can imagine constructing a "consciousness-state centred" probability, which handles things like

repeated awakenings or duplicates. I won't attempt to do so in this post, but I believe that such a theory is worth pursuing.

Of course, finding a useful theory of probability that covers such situations wouldn't mean that the answer would *objectively* be 1/3, just that there is a notion of probability where this is the answer.

Neal is also right to point out that instead of updating on new information, the thirders are tossing out one model and utilising a new model. However, if we constructed a consciousness-state centred probability it would be reasonable to update based on a change of which consciousness-states are considered possibilities.

More on the 1/2 solution

Standard probability theory doesn't handle being asked multiple times (it doesn't even handle indexicals). One of the easiest ways to support this is to normalise multiple queries. For example, if we ask you twice whether you see a cat and you expect to see one 1.2 times on average, we can normalise the probability of seeing a cat to being 0.6 every query by multiplying by 0.5. If we run the sleeping beauty problem twice, you should expect to see one head with a weighting of 1 and two tails with weightings of 0.5. This provides a 50% chance of heads and a 50% chance of tails for one flip. Obviously, it would be a bit of work proving that certain standard theorems still hold, but this is a much more logical way to extend probability theory than the manner proposed.

But beyond this, if we want to pre-commit to guess on observing particular sequences of experiences, the logical choice is to pre-commit to guess on *all* such sequences. This then leads to the answer of 1/2 chance of heads if we follow Neal and collapse multiple matches into one.

Again, none of this is objective, but it all comes down to how we choose to extend classical probability theory.

Is betting a red herring?

As good as *If a tree falls on Sleeping Beauty* is as an article, I agree with Neil that if we merely look at bets, we haven't reached the root of the issue. When people propose using a particular betting scheme, that scheme didn't come out of nowhere. The betting scheme was crafted to satisfy certain properties or axioms. These axioms are the root of the issue. Here, the conflict is between counting repeated queries only once or counting them separately. Once we've chosen which one of these we want to include with our other axioms, the betting scheme (or rather the set of consistent betting schemes) follows. So Ksvanhorn is correct that current solutions on Less Wrong haven't dotted all of their i's and crossed all of their t's. Whether this matters depends on how much you care about certainty. Again, I won't attempt to pursue this approach in this post.

Conclusion

We've seen that behind all of the maths, Neil is actually performing quite a simple operation and it has very little relation to anything that we are interested in. On the other hand, the critiques of current solutions are worth taking to heart. It doesn't imply that these are necessarily wrong, just that they aren't formal proofs. Overall, I believe that both 1/2 and 1/3 are valid answers depending on exactly what the

question is, although I have not embarked on the quest of establishing a formal footing in this post.

Amplification Discussion Notes

Paul Christiano, Wei Dai, Andreas Stuhlmüller and I had an online chat discussion recently, the transcript of the discussion is available here. (Disclaimer that it's a nonstandard format and we weren't optimizing for ease of understanding the transcript). This discussion was primarily focused on amplification of humans (not later amplification steps in IDA). Below are some highlights from the discussion, and I'll include some questions that were raised that might merit further discussion in the comments.

Highlights

Strategies for sampling from a human distribution of solutions:

> Paul: For example you can use "Use random human example," or "find an analogy to another example you know and use it to generate an example," or whatever.
>
> There is some subtlety there, where you want to train the model that sample from the real human distribution rather than from the empirical distribution of 10 proposals you happen to have collected so far. If samples are cheap that's fine. Otherwise you may need to go further to "Given that [X1, X2, ...] are successful designs, what is a procedure that can produce additional successful designs?" or something like that. Not sure.

Dealing with unknown concepts

> Andreas: Suppose you get a top-level command that contains words that H doesn't understand (or just doesn't look at), say something like "Gyre a farbled bleg.". You have access to some data source that is in principle enough to learn the meanings of those words. What might the first few levels of questions + answers look like?
>
> Paul: possible questions: "What's the meaning of the command", which goes to "What's the meaning of word X" for the words X in the sentence, "What idiomatic constructions are involved in this sentence?", "What grammatical constructions are involved in the sentence"
>
> Answers to those questions are big trees representing meanings, e.g. a list of properties of "gyre" (what properties the subject and object typically have, under what conditions it is said to have occurred, why someone might want you to do it, tons of stuff most of which will be irrelevant for the query)
>
> Which come from looking up definitions, proposing definitions and seeing how well they match with usage in the cases you can look at, etc.

Limits on what amplification can accomplish

> Paul:In general, if ML can't learn to do a task, then that's fine with me. And if ML can learn to do a task but only using data source X, then we are going to have to integrate data source X into the amplification process in order for amplification to be able to solve it, there is no way to remove the dependence on arbitrary data

sources. And there will exist data sources which pose alignment issues, independent of any alignment issues posed by the ML.

Alignment search for creative solutions

Considering the task of generating a solution to a problem that requires creativity, it can be decomposed into:

Generate solutions

Evaluate those solutions

For solution generation, one idea is to shape the distribution of proposals so you are less likely to get malign answers (ie. sample from the distribution of answers a human would give, which would hopefully be more likely to be safe/easily evaluated compared to some arbitrary distribution).

I asked Paul if he thought that safe creative solution generation would require sampling from a less malign distribution, or whether he thought we could solve evaluation ("secure-X-evaluation", as testing whether the solution fulfilled property X) well enough to use an arbitrary distribution/brute force search.

> Paul: I don't see a good way to avoid solving secure X-evaluation anyway. It seems to me like we can generate solutions in ways that put much lower probability on malign answers, but it neither seems like we can totally eliminate that (I don't think human creativity totally eliminates that either), nor that we will always have access to some more-aligned human generator
>
> The best I'd probably say is that we can have a generation process that is not itself malign, not clear if that is helpful at all though.

We then dived into how well we could solve secure X-evaluation. I was particularly interested in questions like how we could evaluate whether a design had potentially harmful side-effects.

> Paul: I think what we want is something like: if the designing process knows that X is bad, then the evaluator will also know it. If the designing process doesn't know that X is bad, then that's not malign.
>
> [to be clear, for this discussion we only need security in the infinite limit; in practice the capability of both sides will be limited by the capability of the ML, so we'll also need something to make sure the evaluating-ML does better than the generator-ML, but that seems like a separate issue.]
>
> William: If you imagine slowly increasing the intelligence of the generator, then for any heuristic, it might first start picking solutions that fulfill that heuristic more often before actually understanding the heuristic, and it might take longer after that before the generator understands that the heuristic works because of a causal pathway that involves negative side effects. Is it the case that you'd say that this is an acceptable outcome/something that we can't really get past?
>
> Paul: If neither the evaluator nor generator knows about the negative side effect, it's hard for the negative side effect to lead to higher evaluations. I agree this can happen sometimes (I wrote the implicit extortion post to give an example, there

are certainly others), but they seem OK to accept as "honest mistakes" so far, i.e. none of them pose an existential risk.

in terms of "what amplification is supposed to accomplish," if there is a problem that could just as well afflict a human who is trying their best to help me get what I want (and has the AI's profile of abilities), then I'm basically considering that out of scope.

Whatever we could tell to a human, to help them avoid this kind of problem, we could also tell to an aligned AI, so the problem is factored into (a) help a human avoid the problem, (b) build aligned AI.

A Rationalist Argument for Voting

The argument that voting is irrational is commonplace. From your point of view as a single voter, the chance that your one vote will sway the election are typically miniscule, and the direct effect of your vote is null if it doesn't sway the election. Thus, even if the expected value of your preferred candidate to you is hugely higher than that of the likely winner without your vote, when multiplied by the tiny chance your vote matters, the overall expected value isn't enough to justify the small time and effort of voting. This problem at the heart of democracy has been noted by many — most prominently, Condorcet, Hegel, and Downs.

There have been various counterarguments posed over the years:

- Voters get some kind of intrinsic utility from the act of voting expressively.
- Voters get some utility because changing the margin of the election affects how the resulting government behaves. (I personally find this argument highly implausible; the level of effects that would be necessary seem visibly lacking.)
- If voters have a sufficiently altruistic utility function, the expected utility of a better government for all citizens could be sufficient to make voting worth it.
- Voting itself is irrational, but having a policy of voting is rational.
    - This could be true if, for instance, paying attention to politics were intrinsically good for one's mental health, and yet akrasia would prevent paying sufficient attention without a policy of voting. I suspect most readers here will decisively reject that idea.
    - This could also be true if there were some kind of iterated or outrospective prisoners dilemma [Ed: actually, more like stag hunt] involved, in which voting was cooperation and not-voting was betrayal.

Of all of the above, I find the last bullet most interesting. But I am not going to pursue that here. Here, I'm going to propose a different rationale for voting; one that, as far as I know, is novel.

Participating in democratic elections is a group skill that requires practice. And it's worth practicing this skill because there is an appreciable chance that a future election will have a significant impact on existential risk, and thus will have a utility differential so high so as to make a lifetime of voting worth it.

Let's build a toy model with the following variables:

- c: The cost of voting, in utilons.
- i: "importance", the probability that your highest values — whether that is the survival of your ethnic group, the flourishing of humanity in general, maximizing pleasure for all sentient beings, or whatever — hang in the balance in any given future election. Call such elections "important".
- u: the utility differential, in utilons, at stake in important elections. To cancel out utilons, we can focus on the dimensionless quantity u/c.
- l: number of people "like you" in any given election
- t: the chance that a given person like you truly notices the election is important
- f<t: the chance of false positives in noticing important elections
- s: the chance that a person like you will, if they vote, cast a correctly strategic ballot in an important election
- b<s: chance that they will, if they vote, cast an anti-strategic (bad) ballot

- p: marginal slope of probability of a good outcome. The chance that m strategic ballots will have the power to swing the election is roughly pm over the plausible range of values of m.

Note that t, f, s, and b refer to individuals' marginal chances, but independence is not assumed; outcomes can be correlated across voters. So the utility benefit per election per voter of the policy of "voting iff you notice that the election is important" is uit(s-b)p, while the cost is itc+(1-i)fc. The utility benefit per election of "always voting" is ui(s-b)p, while its costs are c. If u/c can take values above 1e11 and i is above 1e-4 — values I consider plausible — then for reasonable choices of the other variables "always voting" can be a rational policy.

This model is weak in several ways. For one, the chances of swinging an election with strategic votes are not linear with the number of votes involved; it's probably more like a logistic cdf, and l could easily be large enough that the derivative isn't approximately constant. For another, adopting a policy of voting probably has side-effects; it probably increases t, possibly decreases f, and may increase one's ability to sway the votes of other voters who do not count towards l. All of these structural weaknesses would tend to lead the model to underestimate the rationality of voting. (Of course, numerical issues could lead to bias in either direction; I'm sure some people will find values of i>1e-4 to be absurdly high.)

Yet even this simple model can estimate that voting has positive expected value. And it applies whether the existential threat at issue is a genocidal regime such as has occurred in the past, or a novel threat such as powerful, misaligned AI.

Is this a novel argument? Somewhat, but not entirely. The extreme utility differential for existential risk is probably to some degree altruistic. That is, it's reasonable to exert substantially more effort to avert a low-risk possibility that would destroy everything you care about, than you would if it would only kill you personally; and this implies that you care about things beyond your own life. Yet this is not the everyday altruism of transient welfare improvements, and thus it is harder to undermine it with arguments using revealed preference.

UDT can learn anthropic probabilities

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

Today I realized that UDT doesn't have to clash with "objective" views of anthropic probability, like SSA or SIA: instead it can, in some sense, learn which of these views is true!

The argument goes like this. First I'll describe a game where SSA and SIA lead to different decisions. Unlike Sleeping Beauty, my game doesn't involve memory loss, so someone can play it repeatedly and keep memories of previous plays. Then we'll figure out what UDT would do in that game, if it valued each copy's welfare using the average of SSA and SIA. It turns out that different instances of UDT existing at once will act differently: the instances whose memories are more likely under SSA will make decisions recommended by SSA, and likewise with SIA. So from the perspective of either view, it will look like UDT is learning that view, though from an agnostic perspective we can't tell what's being learned.

First let's describe the game. Imagine there's a small chance that tomorrow you will be copied many times - more than enough to outweigh the smallness of the chance. More precisely, let's say some event happens at 1:N odds and leads to N^2 copies of you existing. Otherwise (at N:1 odds) nothing happens and you stay as one copy. You're offered a choice: would you rather have each copy receive a dollar if the event happens, or receive a dollar if the event doesn't happen? The former corresponds to SIA, which says you have N:1 odds of ending up in the world with lots of copies. The latter corresponds to SSA, which says you have N:1 odds of ending up in a world with no copying.

We can repeat this game for many rounds, allowing the players to keep their memories. (Everyone existing at the end of a round gets to play the next one, leading to a whole lot of people in the end.) Consider for a moment what happens when you're in the middle of it. Imagine you started out very sure of SSA, but round after round, you kept finding that you were copied. It would be like seeing a coin come up heads again and again. At some point you'll be tempted to say "what the hell, let's do a Bayesian update in favor of SIA". This post is basically trying to give a UDT justification for that intuition.

Our anthropic repeated game is quite complicated, but if SSA and SIA were equally likely, it would be equivalent to this non-anthropic game:

1) Flip a coin and make a note of the result, but don't show it to the player. This step happens only once, and determines whether the whole game will take place in "SSA world" or "SIA world".

2) Simulate the game for a fixed number of rounds, using a random number generator to choose which copy the player "becomes" next. Use either SSA (all worlds weighted equally) or SIA (worlds weighted by number of copies), depending on the coin from step 1.

This game is non-anthropic, so UDT's solution agrees with classical probability: the player should update their beliefs about the coin after each round, and bet accordingly. Since each round leads to an N:1 update in one direction or the other, the

player will simply bet according to the majority of their observations so far. For example, if they "got copied" 5 times and "didn't get copied" 3 times, they should bet that they'll "get copied" next time. That's the money-maximizing strategy.

Now let's go back to the anthropic repeated game and see how UDT deals with it. More precisely, let's use the version of UDT described in this post, and make it value each copy's welfare at the end of the game as the average of that copy's SSA and SIA probabilities. (That gets rid of all randomness and gives us only one static tree of copies, weighted in a certain way.) Then UDT's strategy will be the same as in the non-anthropic game: the instances whose memories say they got copied more than half the time will "bet on getting copied" in the next round as well, and vice versa. That's surprising, because naively we expect UDT to never deviate from its starting policy of 50% SSA + 50% SIA, no matter what it sees.

Here's a simple way to understand what's happening. Yes, each instance of UDT will value its descendants using 50% SSA + 50% SIA. But if some instance's memories agree more with SIA for example, then it knows that its next decision will mostly affect descendants with high SIA weight but low SSA weight. It's pretty much the same as UDT's handling of ordinary Bayesian evidence.

This result makes me happy. Making different decisions based on anthropic evidence in favor of SSA or SIA isn't just something a human would do, it's also rational according to UDT with the same odds. Moreover, it's possible that our current evidence already strongly favors SSA or SIA, undermining the "no unique answer" view which is popular on LW.

The idea also sheds some light on the moral status of copies and nature of selfishness. If we define a UDT agent in the above way, it will have the curious property that most of its instances according to SSA will be "SSA-selfish", and the same for SIA. So we can define a robustly selfish agent by giving it a prior over different kinds of selfishness, and then letting it learn.

RFC: Meta-ethical uncertainty in AGI alignment

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

I'm working on writing a paper about an idea I previously outlined for addressing false positives in AI alignment research. This is the first completed draft of one of the subsections arguing for the adoption of a particular, necessary hinge proposition to reason about aligned AGI. I appreciate feedback on this subsection especially regarding if you agree with the line of reasoning and if you think I've ignored anything important that should be addressed here. Thanks!

AGI alignment is typically phrased in terms of aligning AGI with human interests, but this hides some of the complexity of the problem behind determining what "human interests" are. Taking "interests" as a synonym for "values", we can begin to make some progress by treating alignment as at least partially the problem of teaching AGI human values (Soares, 2016). Unfortunately, what constitutes human values is currently unknown since humans may not be aware of the extent of their own values or may not hold reflexively consistent values (Scanlon, 2003). Further complicating matters, humans are not rational, so their values cannot be deduced from their behavior unless some normative assumptions are made (Tversky, 1969), (Armstrong and Mindermann, 2017). This is a special case of Hume's is-ought problem—that axiology cannot be inferred from ontology alone—and it complicates the problem of training AGI on human values (Hume, 1739).

Perhaps some of the difficulty could be circumvented if a few normative assumptions were made, like assuming that rational preferences are always better than irrational preferences or assuming that suffering supervenes on preference satisfaction. This poses an immediate problem for our false positive reduction strategy by introducing additional variables that will necessarily increase the chance of a false positive. Maybe we could avoid making any specific normative assumptions prior to the creation of aligned AGI by expecting the AGI to discover them via a process like Yudkowsky's coherent extrapolated volition (Yudkowsky, 2004). This may avoid the need to make as many assumptions, but still requires making at least one—that moral facts exist to permit the correct choice of normative assumptions—and reveals a deep philosophical problem at the heart of AGI alignment—meta-ethical uncertainty.

Meta-ethical uncertainty stems from epistemic circularity and the problem of the criterion since it is not possible to know the criteria by which to asses which moral facts are true or even if any moral facts exist without first assuming to know what is good and true (Chisholm, 1982). We cannot hope to resolve meta-ethical uncertainty here, but we can at least decide what impact particular assumptions about the existence of moral facts have upon false positives in AGI alignment. Specifically, whether or not moral facts exists and, if they do, what moral facts should be assumed to be true.

On the one hand suppose we assume that moral facts exist, then we could build aligned AGI on the presupposition that it could at least discover moral facts even if no moral facts were specified in advance and then use knowledge of these facts to

constrain its values such that they aligned with humanity's values. Now suppose this assumption is false and moral facts do not exist, then our moral-facts-assuming AGI would either never discover any moral facts to constrain its values to be aligned with human values or would constrain itself with arbitrary moral facts that would not be sure to produce value alignment with humanity.

On the other hand suppose we assume that moral facts do not exist, then we must build aligned AGI to reason about and align itself with the axiology of humanity in the absence of any normative assumptions, likely on a non-cognitivist basis like emotivism. Now suppose this assumption is false and moral facts do exist, then our moral-facts-denying AGI would discover the existence of moral facts, at least implicitly, by their influence on the axiology of humanity and would align itself with humanity as if it had started out assuming moral facts existed but at the cost of solving the much harder problem of learning axiology without the use of normative assumptions.

Based on this analysis it seems that assuming the existence of moral facts, let alone assuming any particular moral facts, is more likely to produce false positives than assuming moral facts do not exist because denying the existence of moral facts gives up the pursuit of a class of alignment schemes that may fail, namely those that depend on the existence of moral facts. Doing so likely makes finding and implementing a successful alignment scheme harder, but it does this by replacing difficulty tied to uncertainty around a metaphysical question that may not be resolved in favor of alignment to uncertainty around implementation issues that through sufficient effort may be made to work. Barring a result showing that moral nihilism—the assumption that no moral facts exist—implies the impossibility of building aligned AGI, it seems the best hinge proposition to hold in order to reduce false positives in AGI alignment due to meta-ethical uncertainty.

References:

- Nate Soares. The Value Learning Problem. In *Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence*. (2016). **Link**
- T. M. Scanlon. 3 Rawls on Justification. 139 In *The Cambridge Companion to Rawls*. Cambridge University Press, 2003.
- Amos Tversky. Intransitivity of preferences.. *Psychological Review* **76**, 31–48 American Psychological Association (APA), 1969.**Link**
- Stuart Armstrong, Sören Mindermann. Impossibility of deducing preferences and rationality from human policy. (2017). **Link**
- David Hume. A Treatise of Human Nature. Oxford University Press, 1739. **Link**
- Eliezer Yudkowsky. Coherent Extrapolated Volition. (2004). **Link**
- Roderick M. Chisholm. The Foundations of Knowing. University of Minnesota Press, 1982.

Could we send a message to the distant future?

Suppose that humanity wiped itself out but left behind complex multicellular life. I think there is a good chance that space-faring civilization would emerge again on Earth, and I've argued that it might be worthwhile to try sending them a message.

(In that post I guesstimate that if all went well you might be able to effectively reduce extinction risk by 1/300 by sending a message to the future; I could imagine that costing only ~$10M and leveraging interest from non-EAs, in which case it sounds like a good buy.)

Unfortunately, sending a message 500 million years into the future seems very hard. Most things that aren't buried will get destroyed and the landscape itself is going to be completely transformed (tectonic plates will be rearranged, mountains will appear and disappear, the world will be covered in forms of life that don't exist yet...).

H/T to Dan Kane for criticizing my original post and clarifying the actual nature of the problem.

More precisely my question is whether we can:

- Spend $10M-$100M.
- Encode 100MB of information.
- Wait 500 million years.
- Have it be found with probability >25% by a civilization as sophisticated as humanity in 1900.

I'd also be interested in relaxing any of these constraints a little bit, e.g. spending 10x more, sending 10x less data, only lasting 100M years, or only being discoverable in the 21st century.

I'm not very worried about preserving the information---I suspect that if we are willing to bury something, we can preserve 100MB pretty cheaply. My biggest concern is putting the information somewhere that our successors can find it.

My initial suggestion, clarified/improved/named by Jess Riedel, was to:

- Bury a small number of expensive "payload" messages (maybe ~100)
- Bury a large number of "map" messages (maybe 10,000 - 100,000).
- Somehow get people to stumble across a map. Either put a lot of them in places where they might end up being easy to find, or transform the landscape in ways that might remain visible in 500M years.
- Use the maps to encode the location of payloads. (With each payload also including much more detailed maps to all the other payloads.)

I now feel like both of these steps are hard:

- It seems very difficult to actually put maps in places where they'll be found. For example, my proposal of using a giant+out-of-place+slow-to-weather rock is probably very difficult given how much the landscape will change, since the rock will probably be either buried or moved.

- It seems very difficult to reliably point to the location of the payload, given how drastically the world map will change. You'd probably need to combine (a) a clever way of communicating locations, with (b) some form of beacon that would be visible if people were looking for roughly ghe right thing in roughly the right place. This is further compounded by the difficulty of telling what we were saying.

I'm not sure if either of those difficulties are serious. For example, I'm not sure that the relative locations of nearby items would get scrambled too much, in which case you might be able to use local maps.

Even if those difficulties are serious, there is a huge space of possibilities and I suspect that there is something that works and is reasonably cheap:

- We could potentially store messages in the ruins of prominent cities, if cities have a reasonable chance of being buried+preserved (cities have enough weird materials in them that I expect they'd leave a really visible mark). This could either be used to make small messages easier to find, a place to put payloads (which can potentially be pointed to with a map of the city), or both.
- If making preserved messages (or messages with a reasonable shot at fossilization) is extremely cheap, then we could potentially send very large numbers of them. This could be used either to send a bunch of payloads and rely on redundancy, or to allow maps to be very large and expressive. It could also be used to flood the world with massive numbers of maps (>>1M), so that they can be easily found without beacons. (Really extensive flooding sounds more like a last ditch effort once we can see extinction coming, rather than something you'd do preemptively.)
- There might be geologically inactive locations where you can just leave giant+out-of-place+slow-to-weather rocks and they have a reasonable probability of remaining intact. Mountain ranges form over much less than 500M years, but it's not clear to me that the whole world churns since I don't really know anything about geology. To do this, you'd need to find a rock that wouldn't wear away entirely, and you'd need a location where it wouldn't be disturbed too much or end up under ground.
- There were some plausible suggestions in a Facebook thread on this topic (including putting stuff in space, defining coordinates with respect to tectonic plates)

If we could come up with a really convincing and reasonably cheap way to send a message, then I think it's probably worth exploring this idea at least a little bit further. I think the next step would be more seriously analyzing how much good a message could potentially do (which is much more speculative than this step).

I'm in the market for certificates of impact for significant contributions to this problem (in the $100-$10k price range, depending on the size of the contribution).

Conceptual issues in AI safety: the paradigmatic gap

This is a linkpost for http://www.foldl.me/2018/conceptual-issues-ai-safety-paradigmatic-gap/

Swimming Upstream: A Case Study in Instrumental Rationality

> One data point for careful planning, the unapologetic pursuit of fulfillment, and success. Of particular interest to up-and-coming AI safety researchers, this post chronicles how I made a change in my PhD program to work more directly on AI safety, overcoming significant institutional pressure in the process.

It's hard to believe how much I've grown and grown up in these last few months, and how nearly every change was borne of deliberate application of the Sequences.

- I left a relationship that wasn't right.
- I met reality without flinching: the specter of an impossible, unfair challenge; the idea that everything and everyone I care about could *actually be in serious trouble* should no one act; the realization that people should do something [1], and that I am one of those people (are you?).
- I attended a CFAR workshop and experienced incredible new ways of interfacing with myself and others. This granted me such superpowers as (in ascending order): *permanent insecurity resolution*, *figuring out what I want from major parts of my life and finding a way to reap the benefits with minimal downside*, and *having awesome CFAR friends*.
- I ventured into the depths of my discomfort zone, returning with the bounty of a new love: a new career.
- I followed that love, even at risk of my graduate career and tens of thousands of dollars of loans. Although the decision was calculated, you better believe it was still scary.

I didn't sacrifice my grades, work performance, physical health, or my social life to do this. I sacrificed something else.

CHAI For At Least Five Minutes

January-Trout had finished the Sequences and was curious about getting involved with AI safety. Not soon, of course - at the time, I had a narrative in which I had to labor and study for long years before becoming worthy. To be sure, I would never endorse such a narrative - Something to Protect, after all - but I had it.

I came across several openings, including a summer internship at Berkeley's Center for Human-Compatible AI. Unfortunately, the posting indicated that applicants should have a strong mathematical background (uh) and that a research proposal would be required (having come to terms with the problem mere weeks before, I had yet to read a single result in AI safety).

> OK, I'm really skeptical that I can plausibly compete this year, but applying would be a valuable information-gathering move with respect to where I should most focus my efforts.

I opened *Concrete Problems in AI Safety*, saw 29 pages of reading, had less than 29 pages of ego to deplete, and sat down.

> This is ridiculous. I'm not going to get it.
>
> ... You know, this would be a great opportunity to *try for five minutes*.

At that moment, I lost all respect for these problems and set myself to work on the one I found most interesting. I felt the contours of the challenge take shape in my mind, sensing murky uncertainties and slight tugs of intuition. I concentrated, compressed, and compacted my understanding until I realized what success would *actually look like*. The idea then followed trivially [2].

Reaching the porch of my home, I turned to the sky made iridescent by the setting sun.

> I'm going to write a post about this at some point, aren't I?

Skepticism

> This idea is cool, but it's probably secretly terrible. I have limited familiarity with the field and came up with it after literally twenty minutes of thinking? My priors say that it's either already been done, or that it's obviously flawed.

Terrified that this idea would become my baby, I immediately plotted its murder. Starting from the premise that it was insufficient even for short-term applications (not even in the limit), I tried to break it with all the viciousness I could muster. Not trusting my mind to judge sans rose-color, I coded and conducted experiments; the results supported my idea.

I was still suspicious, and from this suspicion came many an insight; from these insights, newfound invigoration. Being the first to view the world in a certain way isn't just a rush - it's pure *joie de vivre.*

Risk Tolerance

I'm taking an Uber with Anna Salamon back to her residence, and we're discussing my preparations for technical work in AI safety. With one question, she changes the trajectory of my professional life:

> Why are you working on molecules, then?

There's the question I dare not pose, hanging exposed, in the air. It scares me. I acknowledge a potential status quo bias, but express uncertainty about my ability to do anything about it. To be sure, that work is important and conducted by good people whom I respect. But it wasn't right for *me*.

We reach her house and part ways; I now find myself in an unfamiliar Berkeley neighborhood, the darkness and rain pressing down on me. There's barely a bar of reception on my phone, and Lyft won't take my credit card. I just want to get back to the CFAR house. I calm my nerves (really, would Anna live somewhere dangerous?), absent-mindedly searching for transportation as I reflect. In hindsight, I felt a distinct sense of *avoiding-looking-at-the-problem*, but I was not yet strong enough to admit even that.

A week later, I get around to goal factoring and internal double cruxing this dilemma.

> Litany of Tarski, OK? There's nothing wrong with considering how I actually feel. Actually, it's a dominant strategy, since the value of information is never negative [3]. *Look at the thing*.

I realize that I'm out of alignment with what I truly want - and will continue to be for four years if I do nothing. On the other hand, my advisor disagrees about the importance of preparing safety measures for more advanced agents, and I suspect that they would be unlikely to support a change of research areas. I also don't want to just abandon my current lab.

> I'm a second-year student - am I even able to do this? What if no professor is receptive to this kind of work? If I don't land after I leap, I might have to end my studies and/or accumulate serious debt, as I would be leaving a paid research position without any promise whatsoever of funding after the summer. What if I'm wrong, or being impulsive and short-sighted?

Soon after, I receive CHAI's acceptance email, surprise and elation washing over me. I feel uneasy; it's very easy to be reckless in this kind of situation.

Information Gathering

I knew the importance of navigating this situation optimally, so I worked to use every resource at my disposal. There were complex political and interpersonal dynamics at play here; although I consider myself competent in these considerations, I wanted to avoid even a single preventable error.

> Who comes to mind as having experience and/or insight on navigating this kind of situation? This list is incomplete - whom can I contact to expand it?

I contacted friends on the CFAR staff, interfaced with my university's confidential resources, and reached out to contacts I had made in the rationality community. I posted to the CFAR alumni Google group, receiving input from AI safety researchers around the world, both at universities and at organizations like FLI and MIRI [4].

> What obvious moves can I make to improve my decision-making process? What would I wish I'd done if I just went through with the switch *now*?

- I continued a habit I have cultivated since beginning the Sequences: gravitating towards the arguments of intelligent people who disagree with me, and determining whether they have new information or perspectives I have yet to properly consider. *What would it feel like to be me in a world in which I am totally wrong?*
    - Example: while reading the perspectives of attendees of the '17 Asilomar conference, I noticed that Dan Weld said something I didn't agree with. You would not believe how quickly I clicked his interview.
- I carefully read the chapter summaries of *Decisive: How to Make Better Choices in Life and Work* (having read the book in full earlier this year in anticipation of this kind of scenario).
- I did a pre-mortem: "I've switched my research to AI safety. It's one year later, and I now realize this was a terrible move - why?", taking care of the few reasons which surfaced.

- I internal double cruxed fundamental emotional conflicts about what could happen, about the importance of my degree to my identity, and about the kind of person I want to become.
    - I prepared myself to lose, mindful that the objective is *not* to satisfy that part of me which longs to win debates. Also, idea inoculation and status differentials.
- I weighed the risks in my mind, squaring my jaw and mentally staring at each potential negative outcome.

Gears Integrity

At the reader's remove, this choice may seem easy. Obviously, I meet with my advisor (whom I still admire, despite this specific disagreement), tell them what I want to pursue, and then make the transition.

Sure, gears-level models take precedence over expert opinion. I have a detailed model of why AI safety is important; if I listen carefully and then verify the model's integrity against the expert's objections, I should have no compunctions about acting.

I noticed a yawning gulf between *privately disagreeing with an expert*, *disagreeing with an expert in person*, and *disagreeing with an expert in person in a way that sets back my career if I'm wrong*. Clearly, the outside view is that most graduate students who have this kind of professional disagreement with an advisor are mistaken and later, regretful [5]. Yet, argument screens off authority, and

> You have the right to think.
>
> You have the right to disagree with people where your model of the world disagrees.
>
> You have the right to decide which experts are probably right when they disagree.
>
> You have the right to disagree with real experts that all agree, given sufficient evidence.
>
> You have the right to disagree with real honest, hardworking, doing-the-best-they-can experts that all agree, even if they wouldn't listen to you, because it's not about whether they're messing up.

Fin

Many harrowing days and nights later, we arrive at the present, concluding this chapter of my story. This summer, I will be collaborating with CHAI, working under Dylan Hadfield-Menell and my new advisor to extend both *Inverse Reward Design* and *Whitelist Learning* (the latter being my proposal to CHAI; I plan to make a top-level post in the near future) [6].

Forwards

I sacrificed some of my tethering to the social web, working my way free of irrelevant external considerations, affirming to myself that I will look out for my interests. When I

first made that affirmation, I felt a palpable sense of *relief*. Truly, if we examine our lives with seriousness, what pressures and expectations bind us to arbitrary social scripts, to arbitrary identities - to arbitrary lives?

---

[1] My secret to being able to continuously soak up math is that I *enjoy it*. However, it wasn't immediately obvious that this would be the case, and only the intensity of my desire to step up actually got me to start studying. Only then, after occupying myself in earnest with those pages of Greek glyphs, did I realize that it's *fun*.

[2] This event marked my discovery of the mental movement detailed in *How to Dissolve It*; it has since paid further dividends in both novel ideas and clarity of thought.

[3] I've since updated away from this being true for humans in practice, but I felt it would be dishonest to edit my thought process after the fact.

Additionally, I did not fit any aspect of this story to the Sequences *post factum*; every reference was explicitly considered at the time (*e.g.*, remembering that specific post on how people don't usually give a serious effort *even when everything may be at stake*).

[4] I am so thankful to everyone who gave me advice. Summarizing for future readers:

- Speak in terms of the concrete problems in AI safety to avoid immediately getting pattern-matched.
- Frame project ideas (in part) with respect to their relevance to current ML systems.
- Explore all your funding options, including:
  - OpenPhilanthropy
  - Berkeley Existential Risk Initiative
  - Future of Life Institute
  - Paul Christiano's funding for independent alignment research

If you're navigating this situation, are interested in AI safety but want some direction, or are looking for a community to work with, please feel free to contact me.

[5] I'd like to emphasize that support for AI safety research is quickly becoming more mainstream in the professional AI community, and may soon become the majority position (if it is not already).

Even though ideas are best judged by their merits and not by their popular support, it can be emotionally important in these situations to remember that if you are concerned, you are *not* on the fringe. For example, 1,273 AI researchers have publicly declared their support for the Future of Life Institute's AI principles.

> A survey of AI researchers (Muller & Bostrom, 2014) finds that on average they expect a 50% chance of human-level AI by 2040 and 90% chance of human-level AI by 2075. On average, 75% believe that superintelligence (“machine intelligence that greatly surpasses the performance of every human in most professions”) will follow within thirty years of human-level AI. There are some reasons to worry about sampling bias based on *e.g.* people who take the idea of human-level AI seriously being more likely to respond (though see the attempts made to control for such in the survey) but taken seriously it suggests that most AI researchers

> think there’s a good chance this is something we’ll have to worry about within a generation or two.
>
> *AI Researchers on AI Risk (2015)*

[6] Objectives are subject to change.

On the Chatham House Rule

I have gone to several events operating under the Chatham House Rule, and have overall found it more annoying than useful. In this post, I share why I dislike the rule, how I think it can be improved, and hopefully spark others to give ideas on how to improve it. In particular, I think the part about not revealing who was at the event should be opt in. Partially, my goal is to eventually develop a modified version that might be used at future events.

The Chatham House Rule states that "When a meeting, or part thereof, is held under the Chatham House Rule, participants are free to use the information received, but neither the identity nor the affiliation of the speaker(s), nor that of any other participant, may be revealed."

Note that the rule looks slightly ambiguous. If I am in a small conversation, clearly I cannot share who was in that conversation, but what about the list of participants as a whole? I think if you read the rule carefully, you will see that you cannot share who was at the event. Indeed, the official Chatham House website has more explanation, and explicitly states that "the list of attendees should not be circulated beyond those participating in the meeting." However a significant number of people I have asked at Chatham House Rule events were unaware that this was part of the rule.

Keeping participant lists a secret is hard

This is by far the most annoying part of the rule, and the fact that many participants often are not aware that it is part of the rule means that the attempts by people trying to follow the rule are mostly useless. It almost feels like an information hazard to me to make more people aware of what the Chatham House Rule says, since following this part of the rule is so annoying.

I have personally violated this part of the rule many times. Once, I got an email with information on logistics for the event. Inside the email were a bunch of links to google docs with details. One of the docs said that the event was under the Chatham House Rule, without explaining what it was. I at the time did not know that this applied to the list of participants. I forwarded the email to my wife, so she would know where I was and what to pack for me. The email was sent directly to me and all the rest of the participants, so the attendees were visible.

Another time, I went to a Chatham House rule event, and when I arrived, I was given a schedule of talks with an entire list of participants in the back. When I first received it thought to myself "bleh, do I have to shred this?" I didn't dispose of it, I left it in my luggage. My wife later found it and asked me if I wanted it, while starting to flip through it quickly. I told her to just throw it away. She didn't see names, but she could have, which means I messed up.

After that same event, when asked how the event was, I mentioned that I saw "someone" give a talk on X. I mentioned this because I thought X was interesting. The person I was talking to said they knew who the person was, since they had seen that talk.

I went to another event, and afterwords, someone emailed me asking for advice on research projects. In the email they mentioned that they enjoyed talking to me at the event. I wanted to add a third person to the email thread who I knew would have a project that would be a good fit for him. Instead I responded by telling him about the project, and saying that he could add the person to the thread himself, explaining that I didn't want to do it because of the Chatham House rule. He misunderstood, and instead of adding him to the thread, asked me if he could add him to the thread. I got frustrated and decided to interpreted his question as permission and just added him myself. (I am clearly being pedantic here and the question clearly was permission, but I want to illustrate how annoying a literal interpretation of the rule is.)

Worse, the information about who was at the event is **differentially** hard. It is much harder to keep the information about who was there a secret than it is to keep information about who said what a secret. If you only have to keep information about what is said a secret, it is usually a valid response to say that you can't answer a question because of the Chatham House Rule. This is much less the case for information about who was there, because it can come up when you are already talking about a specific person. It is not uncommon to be asked if I know a specific person. How do I respond if I met them at a Chatham House Rule event?

I think this part of the rule is doing harm by making people take the other part less seriously. All while failing to provide benefit because not everyone even knows about it.

How to fix it

It seems we can do much better just by having people opt in to the part where their participation in the event should be a secret.

Note that we probably could not practically have people opt in the the whole of Chatham House Rule. This is because I expect something like half of people would opt in, and you cant keep track of that many people. Also, it is convenient to have everything said at the event under the Chatham House Rule, since otherwise it can be hard to remember what things that were said under the rule.

I expect that only a couple people (perhaps no people) at any given event will opt in to not being revealed to have been there. If this is wrong, this plan will not work. Also, if people want to not be singled out as opting in, this could cause some harm.

One thing to be concerned about is the cost of having two rules. Often there is a cost for having two standards, and I tend to avoid having to pay that cost, even if it means not introducing a better standard. However, in this case, I think that there is little benefit of having only one standard. If people spend 3 minutes at the beginning of each event thinking about what they are agreeing to, this would be better for achieving common knowledge. The main benefit of having one standard is having to keep track of which event used which rule, and if an individual does not want to pay that cost, they could just pretend that it is always the stricter rule.

Other possible minor changes

Some other changes that are probably not all good, but might be worth considering are:

Having formal talks be not under the Chatham House Rule, unless stated otherwise.

Introducing a mechanism for people to report themselves when they make mistakes.

Having a time at the beginning in which everyone agrees to the rule.

Having a time at the end where people can waive their right to Chatham House Rule if they feel that they didn't say anything they dont mind being public.

Improving Teaching Effectiveness: Final Report

This is a linkpost for https://www.rand.org/pubs/research_reports/RR2242.html

The Gates Foundation failed with their very test-score driven approach to increase test-scores of students.

Maybe we should simply get rid of the idea of optimizing education for scoring high on standardized tests?

A general model of safety-oriented AI development

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

This may be trivial or obvious for a lot of people, but it doesn't seem like anyone has bothered to write it down (or I haven't looked hard enough). It started out as a generalization of Paul Christiano's IDA, but also covers things like safe recursive self-improvement.

```
Start with a team of one or more humans (researchers, programmers, trainers, and/or
overseers), with access to zero or more AIs (initially as assistants). The human/AI
team in each round develops a new AI and adds it to the team, and repeats this until
maturity in AI technology is achieved. Safety/alignment is ensured by having some
set of safety/alignment properties on the team that is inductively maintained by the
development process.
```

The reason I started thinking in this direction is that Paul's approach seemed very hard to knock down, because any time a flaw or difficulty is pointed out or someone expresses skepticism on some technique that it uses or the overall safety invariant, there's always a list of other techniques or invariants that could be substituted in for that part (sometimes in my own brain as I tried to criticize some part of it). Eventually I realized this shouldn't be surprising because IDA is an instance of this more general model of safety-oriented AI development, so there are bound to be many points near it in the space of possible safety-oriented AI development practices. (Again, this may already be obvious to others including Paul, and in their minds IDA is perhaps already a cluster of possible development practices consisting of the most promising safety techniques and invariants, rather than a single point.)

If this model turns out not to have been written down before, perhaps it should be assigned a name, like Iterated Safety-Invariant AI-Assisted AI Development, or something pithier?

The Alignment Newsletter #12: 06/25/18

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

Highlights

Factored Cognition *(Andreas Stuhlmuller)*: This is a presentation that Andreas has given a few times on Factored Cognition, a project by Ought that is empirically testing one approach to amplification on humans. It is inspired by HCH and meta-execution. These approaches require us to break down complex tasks into small, bite-sized pieces that can be solved separately by copies of an agent. So far Ought has built a web app in which there are workspaces, nodes, pointers etc. that can allow humans to do local reasoning to answer a big global question.

My opinion: It is unclear whether most tasks can actually be decomposed as required for iterated distillation and amplification, so I'm excited to see experiments that can answer that question! The questions that Ought is trying seem quite hard, so it should be a good test of breaking down reasoning. There's a lot of detail in the presentation that I haven't covered, I encourage you to read it.

Summary: Inverse Reinforcement Learning

This is a special section this week summarizing some key ideas and papers behind inverse reinforcement learning, which seeks to learn the reward function an agent is optimizing given a policy or demonstrations from the agent.

Learning from humans: what is inverse reinforcement learning? *(Jordan Alexander)*: This article introduces and summarizes the first few influential papers on inverse reinforcement learning. Algorithms for IRL attacked the problem by formulating it as a linear program, assuming that the given policy or demonstrations is optimal. However, there are many possible solutions to this problem -- for example, the zero reward makes any policy or demonstration optimal. Apprenticeship Learning via IRL lets you learn from an expert policy that is near-optimal. It assumes that the reward function is a weighted linear combination of *features* of the state. In this case, given some demonstrations, we only need to match the feature expectations of the demonstrations in order to achieve the same performance as the demonstrations (since the reward is linear in the features). So, they do not need to infer the underlying reward function (which may be ambiguous).

Maximum Entropy Inverse Reinforcement Learning *(Brian D. Ziebart et al)*: While matching empirical feature counts helps to deal with the ambiguity of the reward functions, exactly matching featuer counts will typically require policies to be stochastic, in which case there are many stochastic policies that get the right feature counts. How do you pick among these policies? We should choose the distribution

using the principle of maximum entropy, which says to pick the stochastic policy (or alternatively, a probability distribution over trajectories) that has maximum entropy (and so the least amount of information). Formally, we're trying to find a function P(ζ) that maximizes H(P), subject to E[features(ζ)] = empirical feature counts, and that P(ζ) is a probability distribution (sums to 1 and is non-negative for all trajectories). For the moment, we're assuming deterministic dynamics.

We solve this constrained optimization problem using the method of Lagrange multipliers. With simply analytical methods, we can get to the standard MaxEnt distribution, where P(ζ | θ) is proportional to exp(θ f(ζ)). But where did θ come from? It is the Lagrange multiplier for constraint on expected feature counts. So we're actually not done with the optimization yet, but this intermediate form is interesting in and of itself, because we can identify the Lagrange multiplier θ as the reward weights. Unfortunately, we can't finish the optimization analytically -- however, we can compute the gradient for θ, which we can then use in a gradient descent algorithm. This gives the full MaxEnt IRL algorithm for deterministic environments. When you have (known) stochastic dynamics, we simply tack on the probability of the observed transitions to the model P(ζ | θ) and optimize from there, but this is not as theoretically compelling.
One warning -- when people say they are using MaxEnt IRL, they are usually actually talking about MaxCausalEnt IRL, which we'll discuss next.

Modeling Interaction via the Principle of Maximum Causal Entropy *(Brian D. Ziebart et al)*: When we have stochastic dynamics, MaxEnt IRL does weird things. It is basically trying to maximize the entropy H(A1, A2, ... | S1, S2, ...), subject to matching the feature expectations. However, when you choose the action A1, you don't know what the future states are going to look like. What you really want to do is maximize the causal entropy, that is, you want to maximize H(A1 | S1) + H(A2 | S1, S2) + ..., so that each action's entropy is only conditioned on the previous states, and not future states. You can then run through the same machinery as for MaxEnt IRL to get the MaxCausalEnt IRL algorithm.

A Survey of Inverse Reinforcement Learning: Challenges, Methods and Progress: This is a comprehensive survey of IRL that should be useful to researchers, or students looking to perform a deep dive into IRL. It's particularly useful because it can compare and contrast across many different IRL algorithms, whereas each individual IRL paper only talks about their method and a few particular weaknesses of other methods. If you want to learn a lot about IRL, I would start with the previous readings, then read this one, and perhaps after that read individual papers that interest you.

Technical AI alignment

Iterated distillation and amplification

Factored Cognition *(Andreas Stuhlmuller)*: Summarized in the highlights!

Learning human intent

Learning Cognitive Models using Neural Networks *(Devendra Singh Chaplot et al)*

Preventing bad behavior

Minimax-Regret Querying on Side Effects for Safe Optimality in Factored Markov Decision Processes *(Shun Zhang et al)*

Interpretability

Towards Robust Interpretability with Self-Explaining Neural Networks *(David Alvarez-Melis et al)*

How Can Neural Network Similarity Help Us Understand Training and Generalization? *(Maithra Raghu et al)*

AI strategy and policy

AI Nationalism *(Ian Hogarth)*: As AI becomes more important in the coming years, there will be an increasing amount of "AI nationalism". AI policy will be extremely important and governments will compete on keeping AI talent. For example, they are likely to start blocking company takeovers and acquisitions that cross national borders -- for example, the UK could have been in a much stronger position had they blocked the acquisition of DeepMind (which is UK-based) by Google (which is US-based).

AI capabilities

Reinforcement learning

RUDDER: Return Decomposition for Delayed Rewards *(Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich et al)*

Book Review: Why Honor Matters

People who live in honor cultures have a sense of purpose and meaning. They dwell in solidarity with their fellows, are courageous in the face of danger, set great store in hospitality, and put the welfare of the group above their own.

Mostly.

I expect when readers from this site think of honor, it brings to mind Culture of Honor: the Psychology of Violence in the South, the gist of which is that American Southerners kill each other over insults more often than the rest of the country because the biggest chunk of colonial immigrants there were cattle-herders from the border between Scotland and England. I also guess that through Scott Alexander's review of Albion's Seed, the usual view of these people (and how they think) is perhaps unflattering.

In Why Honor Matters, Tamler Sommers offers a defense of honor. It is only a defense, and not an apologia, so he makes no excuses for the evils associated with it (eternal feuds, subjugation of women, etc). He speaks for the general case of honor with a variety of examples, rather than any specific implementation of it. The writing is clear and untechnical.

I.

Why write the book?

> Courage, integrity, solidarity, drama, hospitality, a sense of purpose and meaning - these are attractive values and characteristic, important for living a good and worthwhile life. But there was something else drawing me to honor too, something more fundamental and harder to describe. Though I subscribe to liberal values of toleration and respect for individual freedom, I've come to believe that the Western liberal approach to ethics is deeply misguided. The approach is too systematic, too idealized and abstract - incapable of reckoning with the messy complexity of the real world.

How does honor differ?

> Idealized, systematic, abstract, and universalizable - honor has none of these attributes. Honor, unlike dignity, is not abstract; it's grounded in fact. Honor is real only when people recognize and acknowledge it. Honor codes are local, not universal, tailored to the particular needs of each community. Most important, honor codes are tailored to people as they are, not how we wish to them to be or how we imagine they would be if they were "rational." Honor is full of compromises; it deals in grays rather than black and white. It seeks better alternatives, not ideal alternatives. In short, honor is a thoroughly nonideal form of value, which allows it to operate with a more accurate understanding of human psychology.

Here's the bit that links into our more usual interests:

> Throughout the book I'll try to convince you that these to aspects of honor - its attractive virtues and the unsystematic nature of its codes - are intimately connected. Honor frameworks recognize that it's not easy to be virtuous: to take

> risks and act with integrity and solidarity. We need motivation, what evolutionary biologists and behavioral economists have called "commitment devices," to overcome our natural impulse toward comfort and safety. Honor frameworks offer a rich tapestry of codes and incentives to counteract this impulse. They have rituals and traditions for bringing people together, for celebrating exceptional people and behavior, and for holding people accountable.

So, honor is an applied rather than theoretical ethical framework. It has a very long track record in a variety of environments. It can be instrumentally useful in solving problems.

II.

> A clearer sense of what honor is all about will help us see whether it is worth preserving.

Honor is a group activity. There are two reasons for this: one, honor requires actions (usually interactions with another person or group); two, honor requires recognition from others. Sommers calls this the **honor group**, which is a collection of people who share norms and values. The boundaries can be explicit, like military units and sports teams, or loose like being a southerner or a stand-up comic.

An anthropologist named Frank Stewart identified two dimensions of honor: one is "horizontal", which means it comes with membership and entails privileges and obligations; the other is "vertical" which sorts status within the group according to a member's actions. Using sports as an example, everyone on a team roster gets a jersey (horizontal) but they do not get equal playing time (vertical) or all get to be team captain (vertical). Bad actions drop your vertical honor and failure to meet the obligations of membership can see horizontal honor stripped completely. In general, it is easier to lose it than to gain it.

Vertical honor may have multiple components as well. The Greeks recognized three: *geras, time,* and *kleos. Geras* was the physical rewards indicating contribution to the group - in the division of plunder after the battle, more plunder went to the warriors who fought most bravely. *Time* was the group's estimation of your worth - this determined how one person was treated relative to another, and consequently how they thought of themselves. *Kleos* is usually translated as "glory" - the poets will tell your story and so your deeds will live on after your death. These are all tangible incentives: increased wealth, preferential treatment, and a form of immortality.

An honor group has **honor norms**. A norm is a social rule that governs behavior in a society, some of which are universal (like parents providing for their children) and some of which are not. In particular, not all norms are honor norms, and honor norms vary between honor groups. There are some norms that are nearly universal to honor cultures, including:

Hospitality: The relationship between guest and host is very important, even to the present day. I can vouch for this personally; I experienced unfailing hospitality in both Iraq and Afghanistan despite being a member of a foreign military force. We received briefings reminding us not to express admiration for things in the home, on account of this causing the host to *give* us whatever we admired. A famous recent example referenced in the book is that of Navy SEAL Marcus Luttrell, subject of the movie *Lone Survivor*, where where he is granted hospitality and asylum by the Sabray tribe, who defended him until he could be rescued.

Courage: Challenges cannot be backed down from; they must be met directly. Insults cannot go without a response; disrespect must be punished. This is one of the things that can get out of hand, since at the same time deliberately showing disrespect is a widespread method of probing for weakness; if someone is challenged and backs down, the challenger gains honor at the challengee's expense.

Revenge: It is the personal responsibility of members of honor cultures to settle wrongs against them. In general this extends to the honor group, so members of a family can all respond to wrongs against the family. However, appealing to a third party (like law enforcement) is widely considered shameful. Extant examples in America include the widely known maxim "snitches get stitches" and feuds in sports that don't involve the league office.

III.

The United States, Canada, Western Europe, etc. are dignity cultures. This has a lot of advantages, like human rights. The question is what the absence of honor at the national level costs us in exchange, and whether this is necessary.

Cowardice: We suffer from acute risk aversion. Sommers uses the example of bicycle helmets: the chief impact of rules about helmets is that people stop riding bikes. The impact in safety is very small. This obsession with safety permeates virtually every facet of our lives, and barely flirts with the notion of effectiveness.

Isolation: Honor entails a lot of social connections. Without a culture of honor, all those connections are absent, and as a consequence we are atomized and highly individualistic. Sommers points to two old sociological concepts:

> The first - *gemeinschaft* **-** is sometimes translated as "community" but like many German words has no good translation. People who relate in this way regard themselves as part of the community whose value can't be reduced to its individual parts. They have common goals and values, and they don't make a clean distinction between what's good for them as individuals and what's good for the group as a whole. Examples of *gemeinschaft* include the family, army units, sports teams, and religious communities. Within these groups, there is plenty of competition among individuals. But they are working for a common purpose and share some basic standards for how to evaluate people's behavior and characters.
>
> Tonnies contrasts *gemeinschaft* with another type of social relation that best characterizes the modern industrialized West, *gesellschaft*. Whereas in *gemeinschaft* people share a common identity in spite of their individual distinctions, in *gesellschaft* they remain distinct despite their connections as a community or nation. . . In the place of tradition and unconscious agreement on questions of value, *gesellschaft* employs contracts, laws, and a powerful state to enforce them. Within the constraints of these laws, people are then free to pursue their individual self-interest and develop their own understanding of right, wrong, and what makes a good life.

We observe costs when there is a switch: depression and suicide increase for tribal cultures who undergo rapid modernization, and in soldiers who leave the military.

Shamelessness: In American parlance, we have a lack of accountability. Sommers cites the famous case of 'affluenza' where a psychologist testified in court that a rich teenager could not be held responsible for killing people while driving drunk on account of the fact that he had never been held responsible for anything before.

Naturally the country mocked the case into the ground, but Sommers asserts that the psychologist's assessment *agrees* with the consensus ethical position.

Even if we constrain ourselves to just the emotion, it doesn't seem to be one of much note:

> Anthropologist Dan Fessler asked focus groups of eighty Indonesians from the Bengkulu Province to come up with the fifty-two most commonly discussed emotions and then rank them according to their frequency in society. He did the same for a focus group in Southern California. Shame was second for the Indonesians but forty-ninth(!) for the Californians - well behind "bored," "frustrated," "offended," and "disgusted." Shame, of course, has deep connections with honor. Indeed, many anthropologists refer to honor cultures as shame cultures or "honor/shame cultures."

The converse for dignity cultures is guilt - so dignity/guilt vs honor/shame.

IV.

Honor is tightly wound up with a sense of community, and a sense of honor is *good* for the community. American readers will remember the Boston Marathon bombings in 2013; the city was back to business as usual in only four days, uniting under the slogan *Boston Strong*. The Harvard Kennedy School and the Program on Crisis Leadership wrote a white paper titled "W*hy* Was Boston Strong?" that concluded three characteristics were important: pride, resilience, and the unwillingness to be intimidated. These are honorable characteristics, and clearly the community benefits from not spending more time sheltering in place. A sense of community is also good for the individual: Sommers cites a study by James Hawdon of mass shooting survivors in Omaha and Finland, which found measures of community solidarity significantly correlated with well-being and less depression.

The link gets more clear as the environment gets more extreme. Sommers summarizes the contents of *Culture of Honor* by Nisbett and Cohen; I will compress the summary even more here:

- White men in the American Southeast have a higher rate of violence in response to insults than other groups.
- These areas were settled by Scots-Irish (or if you prefer, Borderers) who previously lived in herding communities on marginal lands.
- Herding communities have all their wealth tied up in animals.
- This leaves them very vulnerable to raiding, as all the wealth can be taken away.
- They live in low population densities, which makes law enforcement hard.
- Vulnerability to raids and lack of law enforcement means depending on oneself for protection.
- Depending on oneself means maintaining a good reputation for violence.
- An aggressive reputation means responding violently to insults.
- Therefore white men in the American Southeast are prone to violence over insults as a legacy of their pre-colonial culture.

Sommers goes on to say that farming communities are less profitable to raid and therefore can afford a much more individualistic orientation. It is interesting to me that Sommers is making some of the same comparisons between herders and farmers as other people make between farmers and foragers, though clearly herders and farmers are both part of the 'farmer' group in the latter comparison.

Group identity is an important feature of honor cultures. Collective identity fosters collective responsibility, which is an incentive to do things which benefit the group. It is also yields harmony and cohesion in the face of external threats; Sommers points to some anthropologists who think this is the evolutionary function of feuds. There is more in this section, but I'm going to jettison it in favor of this other interesting bit about duels:

> Or how about duels? Haven't we done well to move beyond that practice? Isn't it obvious that duels over honor were a divisive (not to mention bloody) force within society and not a unifying one?
>
> Actually, no. Historical and sociological research suggests that duels have a bit of a bad rap. The practice in fact offered many social benefits. In particular, duels served to maintain the egalitarian codes of an honor group. Indeed, early opponents of the duel, such as Francis Bacon and Cardinal Richelieu, opposed it on precisely these grounds. Bacon wanted to expand the power of the monarchy and further distinguish ranks among gentlemen. The equalizing function of duels was an obstacle to the kind of hierarchy that he wanted to create. Duels were also much safer that commonly supposed. Dueling rituals were designed to testify to the honor of the combatants while at the same time minimizing the risk of death or injury. The combatants would use "inaccurate and weak smoothbore pistols" or swords that were "modified to prevent deep penetrating injuries." Many challenges were resolved without fighting of any sort, as "the mere willingness of both parties to show for the fight was often satisfactory." The point of the duel was more "to demonstrate one's status-group membership than to establish dominance over one's opponent. Thus it was less important to win than to display courage."

Most of the quotes seem to have been drawn from Randall Collins' book *Violence: A Micro-sociological Theory*. The point about duels and egalitarianism is hit on repeatedly; this lingers on in the military and on school playgrounds at least as recently as the 1990's, albeit much less formally.

V.

Feuds and duels make an excellent segue to violence. Sommers makes two arguments about violence: first, that losing it completely has a bunch of morally-murky costs; second that violent acts have a much more nuanced morality than the standard answer of causing suffering and therefore wrong. The costs of non-violence mostly consist of key phrases like zero-tolerance, school-to-prison pipeline, prison population, and leviathan. This is familiar and uninteresting.

Much more interesting is the nuanced morality angle. He cites the book *Virtuous Violence* by Alan Fiske and Tage Rai, the thesis of which is that violence is usually morally motivated:

> Their thesis is purely descriptive: people who act violently are usually driven by moral motives rather than selfish ones. According to Fiske and Rai, morality is about regulating relationships, and human beings employ four basic frameworks for regulating their relationships. The first involves acting within a community. The second involves acting within a hierarchy. The third involves issues of fairness and equality. And the fourth involves proportionality and market exchange.

Consider the question of oppression: what should oppressed people do? Honor has played a key role in motivating people to fight their own oppression. Examples are

given of Frederick Douglass:

> Douglass regarded his act of violent resistance as a watershed moment: "It rekindled in my breast the smoldering embers of liberty... and revived a sense of my own manhood."

W. E. B. DuBois:

> At the outset of the twentieth century, W.E.B. DuBois called on black Americans to "sacrifice money, reputation, and life itself on the alter of right...to reconsecrate ourselves, our honor, and our property to the final emancipation of the race for whose freedom John Brown died."

and Zionism:

> They saw themselves as victims and martyrs, suffering for the glory of God. The founder of Zionism, Theodor Herzl, on the other hand, denounced this attitude. He declared it passive, submissive, cowardly, and feminine. He reformed Jewish oppression as a national humiliation, a disgrace to Jewish honor... Once Herzl embraced this logic, one of his first actions was to challenge a leading anti-Semite to a public duel.

A less laudable, more routine question is the morality of bar fights. In Lafayette, Louisiana, there's some agreement about violence. They have all the usual notions of violence being a masculine obligation, that friends and family have to know your worth, etc. And yet:

> But the code of these Lafayette southerners doesn't encourage domination, just active resistance to the domination of others. They distinguished themselves from "agro dudes that have something to prove and go our looking for trouble." In fact, they didn't even consider themselves to be violent. They were peaceful warriors, not "thugs" and "violent people."

Of particular note is the aftermath, which is usually what we are concerned with when we want to prevent or stop violence:

> Copes et al. write: "Fighting also was perceived as cathartic. Adversaries can release the stress of tense situations with a flurry of punches. After the violence, emotions settle. Even on the losing side of the fight, these men accepted that they could resolve conflicts with violence and that it could prevent lasting conflict. Indeed, most believed that dangerous animosity was unlikely to last beyond the incident. In their circles, fights happen, and in most cases, people get over them."

This final point is not actually examined in the book, which is a shame.*

So violence is usually morally motivated, and further honor provides a specific moral *context* for violence. In this context, violence can be both actively virtuous and minimally harmful.

VI.

The aftermath brings us to the final part of the book I will cover: revenge. Revenge is contrasted with the retributivist school of judicial punishment. Retribution, in the context of justice, is the idea that criminals ought to suffer, and that it is worth spending resources to accomplish it. However, like the rest of justice this idea is

designed to exclude victims - retribution should be impersonally meted out by the state. Revenge on the other hand is a direct assertion of the personal nature of the wrong; what the state does or doesn't do is its own business.

Sommers draws several examples from fiction, but I will dwell on the real-life, recent, American example of Laura Blumenfeld.

In 1986 David Blumenfeld, Laura's father, was travelling in Israel when he was shot in the leg by a group of PLO. The member of the group who shot him was named Omar Khatib, who was subsequently arrested and imprisoned in Israel. So how does an American woman get revenge on a man imprisoned in another country?

Patiently. Laura Blumenfeld was a reporter for the Washington Post. She conceals her identity as the daughter of one of the victims of the shooting, and corresponds with Omar in prison on the pretext of writing a book. She also meets with Omar's family. When asked why tourists were shot, both Omar and his family said it was for the cause of Palestinian Liberation. When asked whether they were afraid the family would seek revenge:

> "No," he replies. "There's no revenge. My brother never met the man personally. It's not a personal issue. Nothing personal, so no revenge."

Laura gets close to the Khatib family and to Omar via correspondence. She also *goes to Iran* and interviews the grand ayatollah on the subject of retaliation and blood money. He asserts that Jews cannot get blood money or revenge from Palestinians, because Palestinians are at war with them; however her father was a *tourist,* and tourists can get blood money and revenge as long as they are not trying to occupy Palestine.

Eventually Omar has an appeals trial over a routine matter of his health. Blumenfeld is at that trial, and demands to speak; after being denied, she reveals she is the victim's daughter. By this time, she had kept the secret from Omar and his family for over a year. During their correspondence, Omar had promised not to hurt anyone again, and so:

> "I held onto his eyes, angry, firm, hoping that he felt ashamed: 'This is on your honor,' she says. 'Between the Khatib family and the Blumenfeld family.'"

The families embrace after the trial, and Omar goes back to prison for a reduced sentence (for health reasons, not honorable ones). From a sense of honor she took risks, made sacrifices, demonstrated love and loyalty for her family, and took personal responsibility for the wrong against her father. Such was the revenge of Laura Blumenfeld.

Conclusion

This is the kind of book that I am glad exists, but did not find to be well written. I am extremely pleased that honor is a subject philosophers are willing to approach now (even if only three of them). The arguments are a little disorganized - he frequently bleeds into stuff he covers in other sections of the book, but not in a way that I found useful or illuminating. The book seemed to be burdened by not quite knowing who the audience was. On one hand, the writing was clear and untechnical and he used a lot of sports examples (he also just seems to like sports), suggesting he was writing for regular people. On the other hand, he pretty frequently got defensive about not endorsing honor killings and vigilante justice, even in chapters which were not

relevant to either. This causes me to feel like he expected a political reaction from his peers in academia (perhaps reasonably). That being said, the book would be worth it just for the bibliography; it seems to contain just about everything on the subject written in English.

Poker example: (not) deducing someone's preferences

I've shown that it is, theoretically, impossible to deduce the preferences and rationality of an agent by looking at their actions or policy.

That argument is valid, but feels somewhat abstract, talking about "fully anti-rational" agents, and other "obviously ridiculous" preferences.

In this post, I'll present a simple realistic example of human behaviour where their preferences cannot be deduced. The example was developed by Xavier O'rourke.

The motivations and beliefs of a poker player

In this example, Alice is playing Bob at poker, and they are on their last round. Alice might believe that Bob has a better hand, or a worse one. She may be maximising her expected income, or minimising it (why? read on to see). Even under questioning, it is impossible to distinguish an Alice belief in Bob having a worse hand and Alice following a maximising behaviour, from Bob-better-hand-and-Alice-minimising-income. And, similarly, Bob-worse-hand-and-Alice-minimising-income is indistinguishable from Bob-better-hand-and-Alice-maximising-income.

If we want to be specific, imagine the we are observing Alice playing a game of Texas holdem'. Before the river (the final round of betting), everyone has folded besides Alice and Bob. Alice is holding (K♠, 10♡), and the board (the five cards both players have in common) is (10♢, 10♣, 10♠, J♠, Q♠).

Alice is looking at four-of-a-kind in 10's, and can only lose if Bob holds (9♠, 8♠), giving him a straight flush. For simplicity, assume Bob has raised, and Alice can only call or fold -- assume she's out of money to re-raise -- and Bob cannot respond to either, so his actions are irrelevant. He has been playing this hand, so far, with great confidence.

Alice can have two heuristic models of Bob's hand. In one model, μ_1, she assumes that having specifically (9♠, 8♠) is very low, so she almost certainly has the better hand. In a second model, she notes Bob's great confidence, and concludes he is quite likely to have that pair.

What does Alice want? Well, one obvious goal is to maximise money, with reward $R_{\$}$, linear in money. However, it's possible that Alice doesn't care about how much money she's taking home -- she'd prefer to take *Bob* home instead, her reward is R_{Bob} -- and she thinks that putting Bob in a good mood by letting him win at poker will make him more receptive to her advances later in the evening. In this case Alice wants to lose as much money as she can in this hand, so, in this specific situation, $R_{\$} = -R_{Bob}$.

Then the following table represent's Alice's action, as a function of her model and reward function:

	μ_1	μ_2
$R_{\$}$	call	fold
R_{Bob}	fold	call

Thus, for example, if she wants to maximise money ($R_{\$}$) and believes Bob doesn't have the winning hand (μ_1), she should call. Similarly, (μ_2, R_{Bob}) results in Alice calling (because she believes she will lose if both players show their cards, and wants to lose). Conversely, (μ_1, R_{Bob}) and ($\mu_2, R_{\$}$) result in Alice folding.

Thus observing Alice's behaviour neither constrains her beliefs, nor her preferences -- though it does constrain the combination of the two.

Alice's overall actions

Can we really not figure what Alice wants here? What about if we just waited to see her previous or subsequent behaviour? Or if we simply asked her what she wanted?

Unfortunately, neither of these may suffice. Even if Alice is mainly a money maximiser, it's possible she might take Bob as a consolation prize; even if she was mainly interested in Bob, it's possible that she previously played aggressively to win money, reasoning that Bob is more likely to savour a final victory against a worthy-seeming opponent.

As for asking Alice -- well, sexual preferences and poker strategies are areas where humans are incredibly motivated to lie and mislead. Why confess to a desire that might result in it being impossible to achieve? Or reveal how you analyse poker hands in an unduly honest way? Conversely, honesty or double-bluffs are also options.

Thus, it is plausible that Alice's total behaviour could be identical in the ($\mu_1, R_{\$}$) and ($\mu_2, R_{Bob}$) cases (and in the ($\mu_1, R_{Bob}$) and ($\mu_2, R_{\$}$) cases), not allowing us to distinguish these. Or at least, not allowing us to distinguish them with much confidence.

Adding more details

It might be objected that the problem above is overly narrow, and that if we expanded the space of actions, Alice's preferences would become clear.

That is likely to be the case; but the space of beliefs and rewards was also narrow. We could allow Alice to raise as well (maybe with the goal of tricking Bob into folding); with three actions, we may be able to distinguish better between the four possible pairs. But we can then give Alice more models as to how Bob would react, increasing the space of possibilities. We could also consider more possible motives for Alice -- she might have a risk averse money-loving utility, and/or some mix between $R_{\$}$ and R_{Bob}.

It's therefore not clear that "expanding" the problem, or making it more realistic, would make it any easier to deduce what Alice wants.

Weak arguments against the universal prior being malign

Paul Christiano makes the case that if we use the universal prior to make important predictions, then we will end up assigning a large amount of probability mass to hypotheses which involve intelligent agents living in alternate universes who have thus-far deliberately made the correct predictions so that they might eventually manipulate us into doing what they want us to do. Paul calls these intelligent agents 'consequentialists'.

I find ideas like this very difficult to think about clearly, but I have a strong gut-feeling that the argument is not correct. I've been unable to form a crisp formal argument against Paul's proposal, but below I list a few weak reasons why the consequentialists's probability mass in the universal prior might not be as high as Paul suggests.

- **Unnatural output channel:** It is probably the case that in the vast majority of simple universes which ultimately spawn intelligent life, the most natural output channel is not accessible to its inhabitants. Paul gives an example of such an output channel in his post: in a cellular automata we could read data by sampling the state of the first non-zero cell. The most natural thing here would probably be to start sampling immediately from t = 0. However, if the automata has simple rules and a simple starting state then it will take a very large number of time-steps before consequentialist life has had time to evolve to the point at which it can start to intentionally manipulate the output cell. As another example, take our own universe: if the 'most natural' output channel in our universe corresponds to a particular location then this probably isn't inside our light-cone right now.
- **Unnatural input channel:** Similar to natural output channels not necessarily being accessible, often it will also be impossible for a consequentialist to discern exactly what was fed in to her universe's input channel. In the example of a cellular automata, the most natural input channel is probably the initial state. This is a problem for the automata's inhabitants because, while knowing the state of the universe at a particular time lets you predict the next state, in general it won't let you deduce exactly how old the universe is or what its initial conditions were. Another source of difficulty in recovering the data fed into your universe's input channel is that if your universe implements something analogous to distance/velocity then it in many cases some information necessary to recover the data fed into your universe's input channel might be moving away from you too fast for you to ever recover it (e.g. a space ship flying away from you at max speed in Conway's game of life).
- **Implicit computational constraints:** A complaint many people have with the universal prior is that it places no constraints on the amount of compute associated with a particular hypothesis, (meaning it allows absurd hypothesis like daemons in alternate universes). It is worth noticing that while there is no explicit computational penalty, daemons inside the prior are subject to implicit computational constraints. If the process which the alternate-universe consequentialists must use to predict the next observation we're about to see requires a lot of compute, then from the consequentialist's perspective this fact is *not* irrelevant. This is because (assuming they care about lots of things, not

just controlling the universal prior) they will perceive the cost of the computation as a relevant expense which must be traded off against their other preferences, even though *we* don't personally care how much compute power they use. These implicit computational costs can also further compromise the consequentialist's access to their universe's output channel. For example consider again a simple cellular automata such as Conway's game of life. Conway's game of life is Turing complete-- it's possible to compute an arbitrary sequence of bits (or simulate any compuatable universe) from within the game of life. However, I suspect it *isn't* possible to compute an arbitrary sequence of bits such that this string can be read off by sampling a particular cell *once every time-tick.* In a similar vein, while you can indeed build Minecraft inside Minecraft, you can't do it in such a way that the 'real' Minecraft world and the 'simulated' Minecraft world run at the same speed. So constraints relating to speed-of-computation further restrict the kinds of output channels the consequentialists are able to manipulate (and if targeting a particular output channel is very costly then they will have to trade-off between simplicity of the output channel and expense of reaching it).

I'm tempted to make further arguments about the unlikeliness that any particular consequentialist would especially care about manipulating *our* Solomonoff inductor more than any other Solomonoff inductor in the Tegmark IV multiverse, (even after conditioning on the importance of our decision and the language we use to measure complexity), but I don't think I fully understand Paul's idea of an anthropic update, so there's a good chance this objection has already been addressed.

All these reasons don't completely eliminate daemons from the universal prior, but I think they might reduce their probability mass to epistemically appropriate levels. I've relied extensively on the cellular automata case for examples and for driving my own intuitions, which might have lead me to overestimate the severity of some of the complexities listed above. These ideas are super weird and I find it very hard to think clearly and precisely about them so I could easily be mistaken, please point out any errors I'm making.

Excessive EDA Effortposting

Introduction and motivation

Science! It's important. It's also difficult, largely because of how easily people fool themselves. I've seen plenty of people (my younger self included!) lose the plot as they attempt their first few investigations, letting subtle methodological problems render their entire chain of reasoning worthless.

So I've decided to pick a small, well-behaved dataset nobody's had a thorough look at yet, and perform an *egregiously* thorough and paranoid exploration. My aim is both to showcase Good Science for aspiring Data Analysts, and to give others a chance to criticise my approach: it's entirely possible that I'm still making some blunders, and if I am then I want to know.

Below is a summary of my methodology, with emphasis on the mistakes I managed to dodge. You can see the full R kernel here.

The dataset

The data was collected by renowned gamedev/educator/all-round-smart-person Nicky Case, in an attempt to decide which of six projects ('Crowds', 'Learn/Teach', 'Win/Win', 'Nonviolence', 'Understand', and 'Mindfulness') they should work on next. They described each possible project, and asked fans to rank them from 1-5 stars. The poll, with full descriptions of each project, is still up here, though the project names are indicative enough that you shouldn't need it.

'Crowds' won, handily, and Case built it. At time of writing, they are between major projects; my pretext for performing this analysis is helping them decide which of the runners-up ('Understand' and 'Learn/Teach') to prioritise.

Exploration

I began by loading and cleaning the data. For most data analysis jobs, handling irregularities in a dataset can take up as much time as actually analysing it: I picked an unusually tidy dataset so I'd be able to skip to the parts where people make *interesting* mistakes.

The most I had to deal with was a few responses with missing values, caused by people not filling in every part of the poll. I considered handling this by imputation – deciding to interpret blanks as three-star ratings, say – but eventually decided to limit subjectivity and drop all incomplete responses.

[Normally, I'd look into missing values in more detail, but because I knew the causal mechanism *and* they were a small enough proportion of the dataset *and* I knew there aren't going to be outliers (there's no way a missing value could be hiding a rating of 100/5 stars) *and* these seemed like the least important respondents anyway (that they couldn't be bothered completing a six-question survey suggests they probably don't have strong opinions about this topic), dropping them seemed fair.]

I set a seed, and split the dataset in half using random sampling. I christened one half exploreDF, and the other half verifyDF: my plan was to use the former for Exploratory Data Analysis, and then perform any 'proper' statistical tests on the latter.

[A classic error of new Data Analysts performing EDA is to explore their entire dataset, and then test hypotheses on the same data that suggested them: this is an issue so common it has its own Wikipedia page. The heart of statistical testing is finding results which are unlikely in ways which contradict the null hypothesis, and a result which you decided to test

because you noticed it contradicting the null hypothesis contradicting the null hypothesis *isn't very unlikely.*

There are workarounds – penalties you can apply in an attempt to retroactively purify tests performed on the data that suggested them – but I'm sceptical of these. You can dutifully multiply your p-values by N to accommodate the fact that you picked one of N possible comparisons to test, but there's no sensible way to accommodate the fact that you looked at that kind of comparison in the first place.

Tl;dr: don't do anything isomorphic to training on the testing set.]

[Another very common mistake is to not randomise data splits, or to not set a seed when you do. If, for example, I'd just taken the first 50% of the rows as my exploreDF, that might have introduced bias: what if the dataset were ordered chronologically, and those who responded first had consistently different views to those who responded later? The only way to ensure a fair split is to randomly sample the entire dataset.

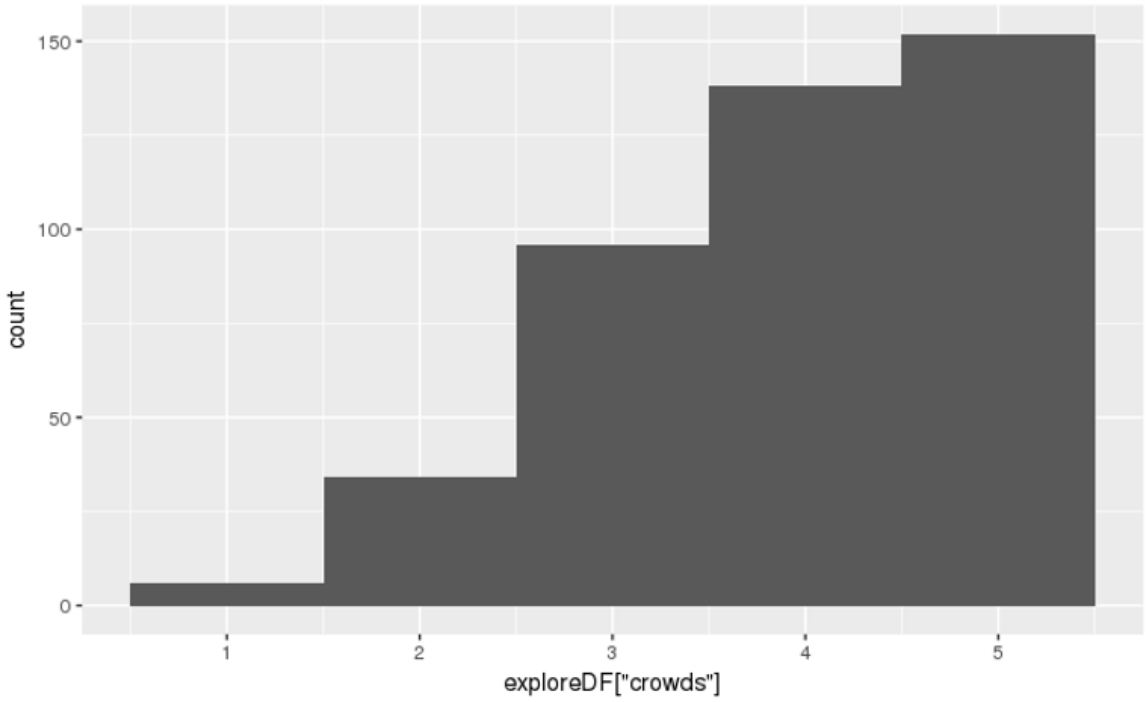
As for failure to set seeds, that's more bad housekeeping and bad habits than an actual mistake, but it's still worth talking about. When you set a seed, you guarantee that the split can be replicated by other people running your code. Enabling others to replicate results is an integral part of the scientific endeavour, and keeps everyone honest. It also lets *you* replicate your results, just in case R goes haywire and wipes your workspace.]

The range of possible scores is 1-5, but the average score for each column is around 3.8 stars. This suggests the possibility of large numbers of respondents who limit themselves to a range 3-5 stars, plus a minority who do not, and who have a disproportionate impact on the average scores. Whether this is a bug or a feature is subjective, but it's worth looking into.

To investigate this possibility, I derived two extra features: range (distance between the highest and lowest scores given by each respondent), and positivity (average score given by each respondent). Also, I was curious to see what correlated with positivity. Were people who generally awarded higher ratings more interested in some projects than others?

[I got '3.8 stars' from Case's summary of the entire dataset: technically this violates data purity, but I would have got that impression about the distribution from the next section anyway.]

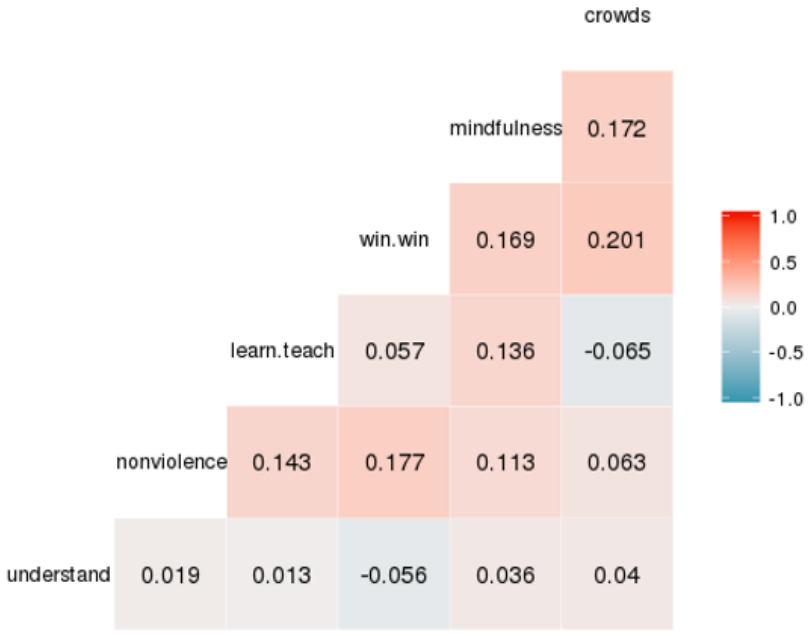
I began with univariate analysis: checking how each of the six variables behaved on its' own before seeing how they interacted.



They were all pretty much like this

Then, I moved on to multivariate analysis. How much do these scores affect each other?

There's an R function I really like, called ggcorr, which plots the correlation coefficients of every variable against every other variable, and lets you see what connections shake out.



I love this genre of triangle so much

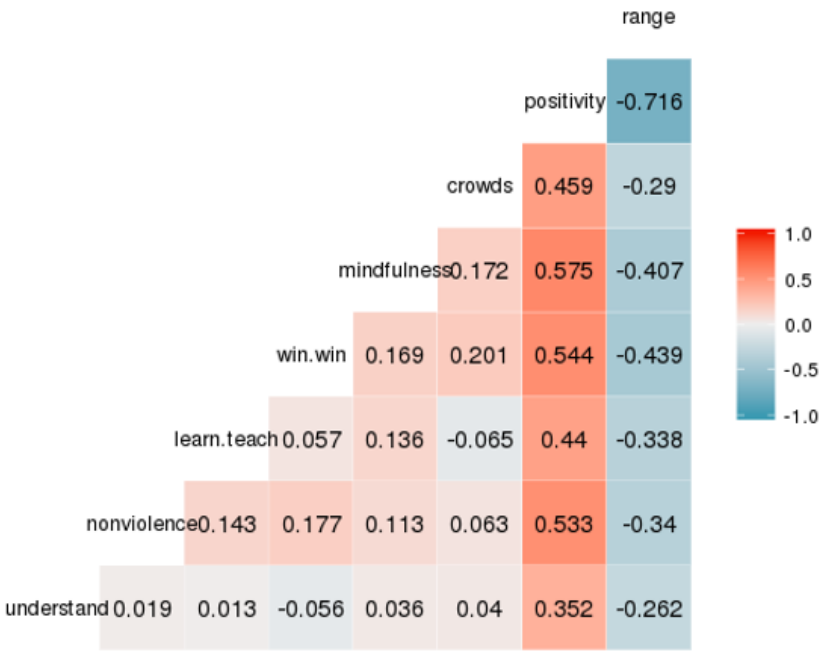
Inferences:

- Surprisingly, there aren't obvious 'cliques' of strong mutual correlation, like I've gotten used to finding when I use ggcorr. The closest thing to a clique I can see here is the mutual affection between 'Crowds', 'Mindfulness' and 'Win/Win', but it's not particularly strong. Also, this finding has no theoretical backing, since I can't think of a good post-hoc explanation that groups these three together but leaves out 'Nonviolence'.
- There's a strong general factor of positivity, as demonstrated by the fact that only two of the fifteen mutual correlations between the six original are negative.
- The two negative correlations are between 'Understand' and 'Win/Win', and between 'Crowds' and 'Learn/Teach'. The former kind of makes sense: people who like the most abstract and academic project dislike the fluffiest and most humanistic project, and vice-versa. The latter, however, astonishes me: what's the beef between education and network theory?
- Five of the six least positive correlations are between 'Understand' and the other projects: this project seems to be more of a niche than its competitors.

I ran a quick statistical test – using R's default cor.test function – on the negative relationship between 'Crowds' and 'Learn/Teach' in exploreDF, just to check it wasn't a fluke. The test returned p=0.0913: not statistically significant, but close enough to reassure me.

[I know I'll catch some flak around here for using p-values instead of Bayesian stats, but p-values are what people are familiar with, so they're what I use to support any results I might want to share with the unenlightened.]

Then, I added range and positivity to the ggcorr plot.



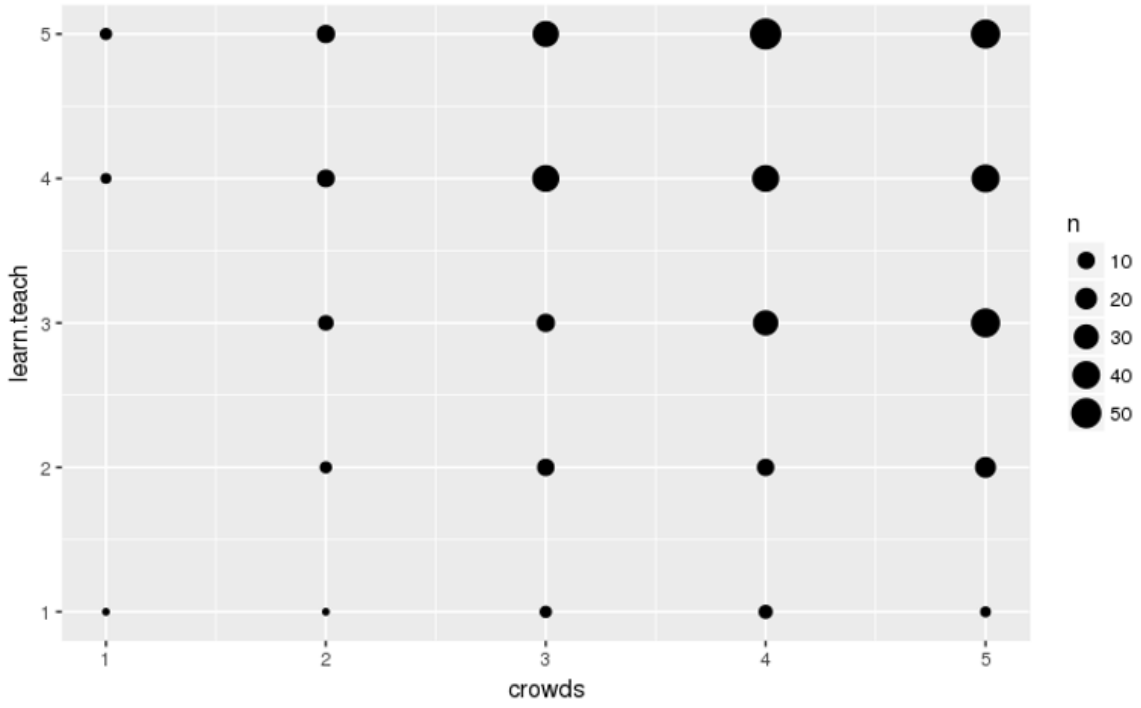
This is also an excellent triangle

Additional inferences:

- Range and positivity are confirmed to be strongly negatively correlated, suggesting (but not confirming) that my theory of a low-positivity minority having a disproportionate impact is correct.

- Every project correlates with the average (that’s to be expected, even without the general factor), but some features correlate much more strongly than others.

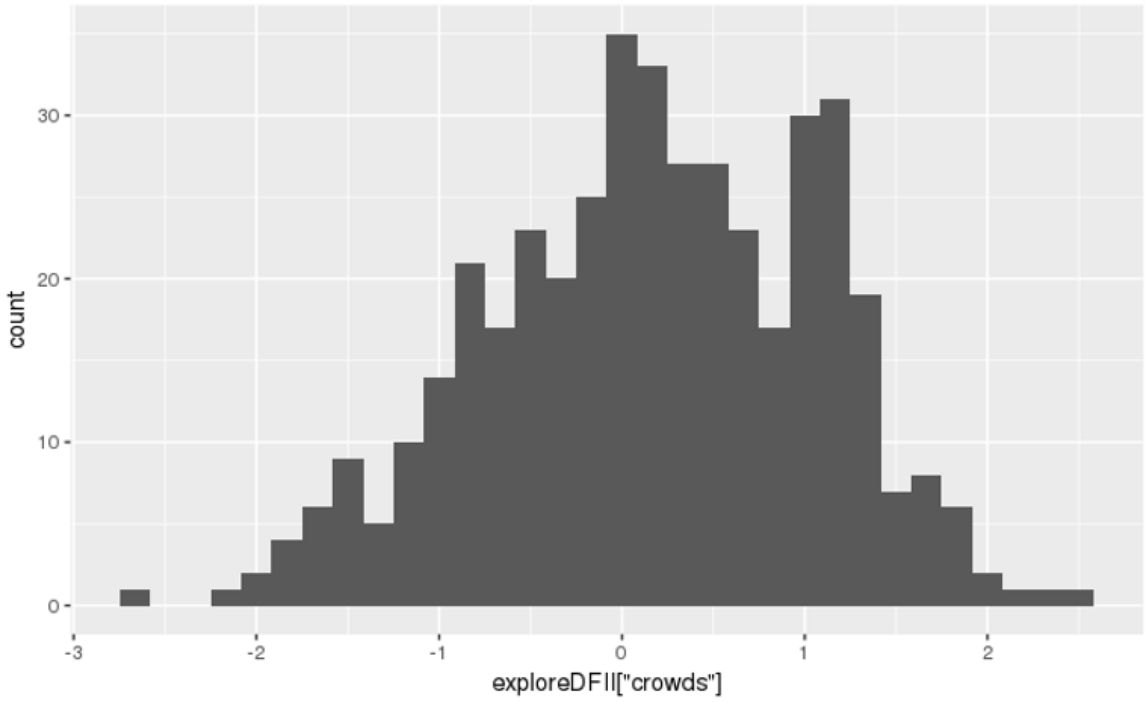
Thinking of Anscombe's Quartet, I plotted the scores against each other on 2D graphs (see below). After eyeballing a few, I was confident that the strengths of these relationships could be approximated with linear correlation.



Nothing about this looks blatantly nonlinear; let’s keep going

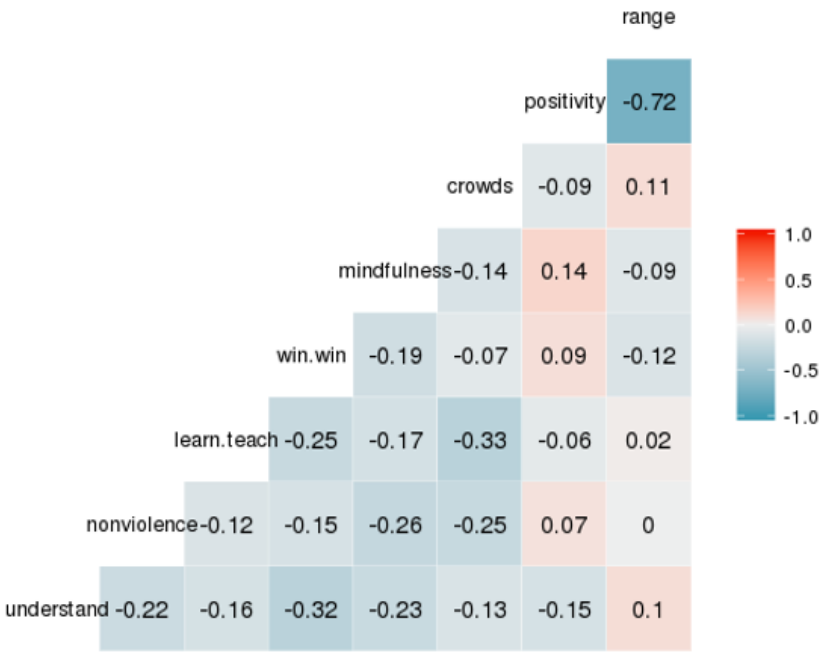
What I was most interested in, though, wasn’t the ratings or scores, but the preferences they revealed. I subtracted the positivity from each score, leaving behind each respondent’s preferences.

On general principles, I glanced at the univariate graphs for these newly derived values . . .



Oh hey, they actually look kind of Gaussian now

. . . before re-checking correlations.



What a calming colour palette

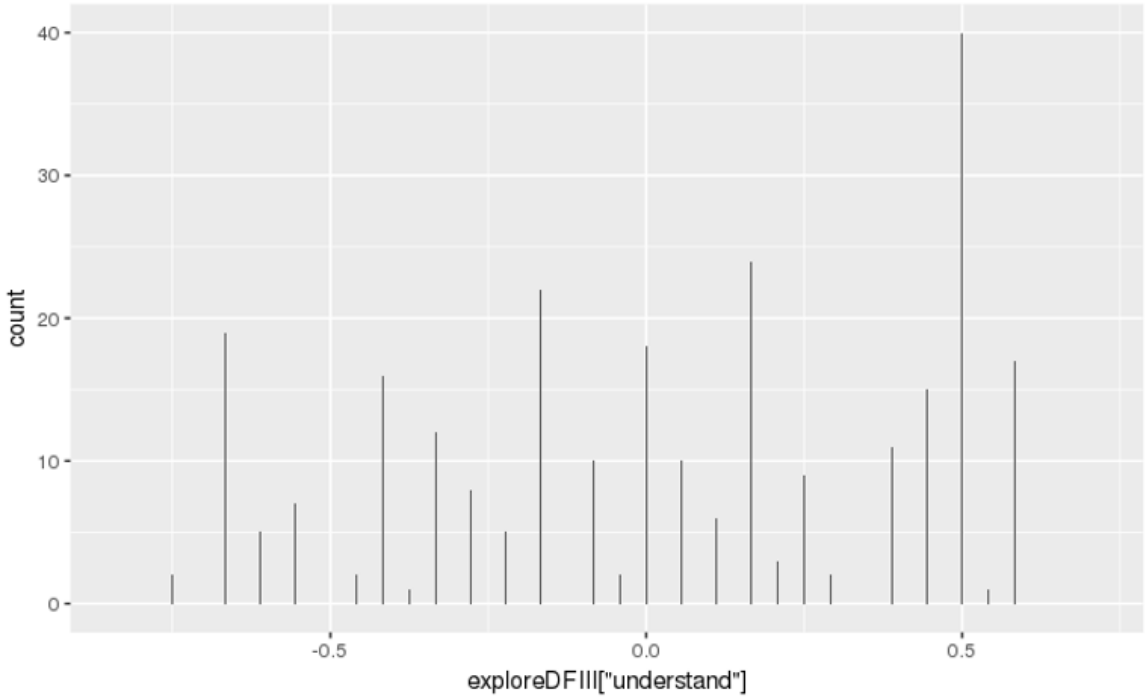
High positivity is correlated with liking ‘Win/Win’, ‘Nonviolence’, and especially ‘Mindfulness’; low positivity is correlated with liking ‘Crowds’, ‘Learn/Teach’, and especially ‘Understand’.

These correlations looked important, so I statistically tested them. All p-values were below 0.1, and the p-values associated with ‘Mindfulness’ and ‘Understand’ were below 0.002 and 0.001 respectively.

In other words, the three frontrunners were the three whose proponents had been least positive overall. Did this mean the results of the original poll were primarily the result of lower-scoring respondents having a greater impact?

To investigate further, I divided all of the positivity-adjusted data for each respondent by the range (in other words, a set of responses [2, 2, 3, 3, 3, 5], which had become [-1, -1 , 0, 0, 0, 2], now became [-1/3, -1/3, 0, 0, 0, 2/3]), to see what things looked like when low-positivity people weren’t having their outsized effect.

As always, I checked the univariate graphs for anything interesting . . .



Yup, those are lines alright

. . . before moving on to comparisons. The averages for my normalised interpretation were:

‘Understand’: 0.028

‘Nonviolence’: 0.023

‘Learn/Teach’: 0.042

‘Win-Win’: -0.076

‘Mindfulness’: -0.080

‘Crowds’: 0.063

For comparison, Case’s averages are

‘Understand’: 3.88

‘Nonviolence’: 3.80

‘Learn/Teach’: 3.87

‘Win-Win’: 3.61

‘Mindfulness’: 3.61

‘Crowds’: 3.94

This is a reassuring null result. The ordinal rankings are more-or-less preserved: ‘Crowds’ > ‘Understand’ & ‘Learn/Teach’ > ‘Nonviolence’ > ‘Win/Win’ & ‘Mindfulness’. The main difference is that ‘Learn/Teach’, in my adjusted version, does significantly better than ‘Understand’.

(Well, I say ‘significantly’: I tried one-sample t-testing the difference between ‘Learn/Teach’ and ‘Understand’, but the p-value was embarrassingly high. Still, a null result can be worth finding.)

I’d found and tested some interesting phenomena; I felt ready to repeat these tests on the holdout set.

It was at this point that I suddenly realised I’d been a complete idiot.

(Those of you who consider yourselves familiar with statistics and Data Science, but didn’t catch my big mistake on the first read-through, are invited to spend five minutes re-reading and trying to work out what I’m referring to before continuing.)

I’d used R’s standard tests for correlation and group differences, but I’d failed to account for the fact that R’s standard tests assume normally-distributed variation. This was *despite* the fact that I’d had no reason to assume a Gaussian distribution, *and* that I’d had visibly not-normally-distributed data staring me in the face every time I created a set of univariate graphs, *and* that I’d been consciously aiming to do an obnoxiously thorough and well-supported analysis. Let this be a lesson to you about how easy it is to accidentally fudge your numbers while using tools developed for social scientists.

On realising my error, I redid every statistical test in the exploration using nonparametric methods. In place of the t-tests, I used a Wilcoxon test, and in place of the Pearson correlation tests, I used Kendall’s tau-b (the most common method for not-necessarily-normal datasets is a Spearman test, but that apparently has some issues when handling discrete data). Fortunately, the main results turned out more or less the same: the biggest change was to the p-value reported for correlation between ‘Crowds’ and ‘Learn/Teach’, which dropped to half its’ value and started to look testable.

[Among the many benefits of doing an explore/verify split and saving your final tests for last: any ‘oh crap oh crap I completely used the wrong test here’ moments can’t have data purity implications unless they happen at the very end of the project.]

Testing

My main relevant findings in exploreDF were:

1. Votes for ‘Understand’ correlate negatively with positivity.
2. Interest in ‘Crowds’ is negatively correlated with interest in ‘Learn/Teach’.
3. The level of correlation between ‘Crowds’ and ‘Learn/Teach’ is unusually low.
4. ‘Understand’ is an unusually niche topic.
5. There’s no major change as a result of normalizing by range.

Appropriate statistical tests for verifyDF are:

1. After adjusting for positivity, run a one-sided Kendall test on ‘Understand’ vs positivity, looking for negative correlation. (α=0.001)
2. Without adjusting for positivity, run a one-sided Kendall test on ‘Crowds’ vs ‘Learn/Teach’, looking for negative correlation. (α=0.05)
3. Recreate the first ggcorr plot: if the correlation between ‘Crowds’ and ‘Learn/Teach’ is the lowest correlation out of the 15 available, consider this confirmed. (this would have a 1/15 probability of happening by chance, so that’s p=α=0.0666)
4. Recreate the first ggcorr plot: if the 5 correlations between ‘Understand’ and other projects are all among the lowest 7 in the 15 available, consider this confirmed. (this would have a 1/143 chance of happening by chance, so that’s p=α=0.007)
5. N/A; I don’t need a test to report on a change I *didn’t* find.

[Note that I use a Fisherian approach and call the critical p-values for my final tests in advance. I think this is generally good practice if you can get away with it, but unlike with most of my advice I’m not going to be a hardass here: if you prefer to plug your results into statistical formulae and report the p-values they spit out, *you are valid*, and so are your conclusions.]

And the results?

[Fun fact: I wrote the entire post up to this point before actually running the tests.]

1. Test passes.
2. Test fails.
3. Test fails.
4. Test passes.

[I’m making a point of publishing my negative results alongside positive ones, to avoid publication bias. Knowing what doesn’t work is just as important as knowing what does; do it for the nullz.]

[I realised in retrospect that these tests interfere in ways that aren’t perfectly scientific. In particular: given #1 passing, #4 passing becomes much more likely, and vice versa. The two positive results I got are fine when considered separately, but I should be careful not to treat these two with the weight I’d assign to two *independent* results with the same p-values.]

Interpretation

My main results are that votes for ‘Understand’ negatively correlate with positivity, and that ‘Understand’ has unusually low degrees of correlation with all the other projects.

So what does this actually mean? Remember, my pretext for doing this is helping Case choose between ‘Understand’ and ‘Learn/Teach’, given similar average scores.

Well, based on my best understanding of what Case is trying to achieve . . . I’d say that ‘Learn/Teach’ is probably the better option. ‘Understand’ is the preferred project of people who seemed to show least enthusiasm for Case’s projects in general, and whose preferences were least shared by other respondents. If we’re taking a utilitarian approach – optimising for the greatest satisfaction of the greatest number – it makes sense to prioritise ‘Learn/Teach’.

However, there are *many* ways to interpret this finding. I called the average score for a given respondent their ‘positivity’, but that was just to avoid it being confused with the average for a given project. Giving lower scores on average could be explained by them taking a more bluntly honest approach to questionnaires, or being better at predicting what they wouldn’t enjoy, or any number of other causes. I can run the numbers, but I can’t peer into people’s souls.

Also, even if the pro-‘Understand’ subset *were* less enthusiastic on average about Case’s work, that wouldn’t imply a specific course of action. “The desires of the many outweigh those of the few.” is a sensible reaction, but “This subset of my fans are the least interested in my work, so they’re the ones I have to focus on keeping.” would also be a completely valid response, as would “This looks like an undersupplied niche, so I could probably get a cult following out of filling it.”

Lessons Learned

In the spirit of ‘telling you what I told you’, here’s a quick summary of every important idea I used in this analysis:

- Don’t test hypotheses on the data that suggested them. One way to avoid this is to split your dataset at the start of an exploration, and get your final results by testing on the unexplored part.
- Splits should be done via random sampling. Random sampling should have a seed set beforehand.
- Take a quick look at univariate graphs before trying to do anything clever.
- Results which have some kind of theoretical backing are more likely to replicate than results which don’t.
- If you don’t have good reasons to think your variables are normally distributed, don’t use t-tests or Pearson correlation tests. These tests have nonparametric equivalents: use those instead.
- Your inferences have limits. Know them, and state them explicitly.

I’ll also throw in a few that I didn’t have cause to use in this analysis:

- If you’re analysing a dataset with a single response variable (i.e. a specific factor you’re trying to use the other factors to predict), it’s probably worth looking at every possible bivariate plot which contains it, so you can pick up on any nonlinear relations.
- Some of the most useful inferences are achieved through multivariate analysis, discovering facts like “wine acidity predicts wine quality *if and only if* price per 100ml is above this level”. I didn’t try this here because the dataset has too few rows, and so any apparent three-factor connection would probably be spurious (too many possible hypotheses, too little data to distinguish between them). Also, I’m lazy.
- If you’re performing an analysis as part of a job application, *create a model*, even if it isn’t necessary. Wasted motion is in general a bad thing, but the purpose of these tasks is to give you an opportunity to show off, and one of the things your future boss most wants to know is whether you can use basic Machine Learning techniques. Speaking of which: *use Machine Learning*, even if the generating distribution is obvious enough that you can derive an optimal solution visually.

Simplified Poker Conclusions

Previously: Simplified Poker, Simplified Poker Strategy

Related (Eliezer Yudkowsky): Meta Honesty: Firming Honesty Around Its Edge Cases

About forty people submitted programs that used randomization. Several of those random programs correctly solved for the Nash equilibrium, which did well.

I submitted the only deterministic program.

I won going away.

I broke even against the Nash programs, utterly crushed vulnerable programs, and lost a non-trivial amount to only one program, a resounding heads-up defeat handed to me by the only other top-level gamer in the room, fellow Magic: the Gathering semi-pro player Eric Phillips.

Like me, Eric had an escape hatch in his program that reversed his decisions (rather than retreating to Nash) if he was losing by enough. Unlike me, his actually got implemented – the professor decided that given how well I was going to do anyway, I'd hit the complexity limit, so my escape hatch was left out.

Rather than get into implementation details, or proving the Nash equilibrium, I'll discuss two things: How few levels people play on, and the motivating point: How things are already more distinct and random than you think they are, and how to take advantage of that.

Next Level

In the comments to the first two posts, most people focused on finding the Nash equilibrium. A few people tried to do something that would better exploit obviously stupid players, but none that tried to discover the opponents' strategy.

The only reason *not* to play an exploitable strategy is if you're worried someone will exploit it!

Consider thinking as having levels. Level N+1 attempts to optimize against Levels N and below, or just Level N.

Level 0 isn't thinking or optimizing, so higher levels all crush it, mostly.

Level 1 thinking picking actions that are generically powerful, likely to lead to good outcomes, without considering what opponents might do. Do 'natural' things.

Level 2 thinking considers what to do against opponents using Level 1 thinking. You try to counter the 'natural' actions, and exploit standard behaviors.

Level 3 counters Level 2. You assume your opponents are trying to exploit basic behaviors, and attempt to exploit those trying to do this.

Level 4 counters Level 3. You assume your opponents are trying to exploit exploitative behavior, and acting accordingly. So you do what's best against *that.*

And so on. Being caught one level below your opponent is death. Being one level ahead is amazing. Two or more levels different, and strange things happen.

Life is messy. Political campaigns, major corporation strategic plans, theaters of war. The big stuff. A lot of Level 0. Level 1 is industry standard. Level 2 is inspired, exceptional. Level 3 is the stuff of legend.

In well-defined situations where losers are strongly filtered out, such as tournaments, you can get glimmers of high level behavior. But mostly, you get it by changing the view of what Level 1 is. The old Level 2 and Level 3 strategies become the new 'rules of the game'. The brain chunks them into basic actions. Only then can the cycle begin again.

Also, 'getting' someone with Level 3 thinking risks giving the game away. What level should one be on next time, then?

Effective Randomization

There is a strong instinct that whenever predictable behavior can be punished, one must randomize one's behavior.

That's true. But only from another's point of view. You can't be predictable, but that doesn't mean you need to be random.

It's another form of illusion of transparency. If you think about a problem differently than others, their attempts to predict or model you will get it wrong. The only requirement is that your decision process is complex, and doesn't reduce to a simple model.

If you also have *different information* than they do, that's even better.

When analyzing the hand histories, I know what cards I was dealt, and use that to deduce what cards my opponent likely held, and in turn guess their behaviors. Thus, my opponent likely has no clue either what process I'm using, how I implemented it, or what data I'm feeding into it. All of that is effective randomization.

If that reduces to me always betting with a 1, they *might* catch on eventually. But since I'm constantly re-evaluating what they're doing, and reacting accordingly, on an impossible-to-predict schedule, such catching on might end up backfiring. It's the same at a human poker table. If you're good enough at reading people to figure out what I'm thinking and stay one step ahead, I need to retreat to Nash, but that's rare. Mostly, I only need to worry, at most, if my actions are effectively doing something simple and easy to model.

Playing the same exact scenarios, or with the same exact people, or both, for long enough, both increases the amount of data available for analysis, and reduces the randomness behind it. Eventually, such tactics stop working. But it takes a while, and the more you care about long histories in non-obvious ways, the longer it will take.

Rather than be *actually* random, instead one adjusts when one's behavior has sufficiently deviated from what would look random, such that others will likely adjust to account for it. That adjustment, too, need not be random.

Rushing into doing things to mix up your play, before others have any data to work with, only leaves value on the table.

One strong strategy when one needs to mix it up is *to do what the details favor.* Thus, if there's something you need to occasionally do, and today is an unusually good day for it, or now an especially good time, do it now, and adjust your threshold for that depending on how often you've done it recently.

A mistake I often make is to choose actions as if I was assuming others know my decision algorithm and will exploit that to extract all the information. Most of the time this is silly.

This brings us to the issue of Glomarization.

Glomarization

Are you harboring any criminals? Did you rob a bank? Is there a tap on my phone? Does this make me look fat?

If when the answer is no I would tell you no, then refusing to answer is the same as saying yes. So if you want to avoid lying, and want to keep secrets, you need to sometimes refuse to answer questions, to avoid making refusing to answer too meaningful an action. Eliezer discussed such issues recently.

This section was the original motivation for writing the poker series up now, but having written it, I think a full treatment should mostly just be its own thing. And I'm not happy with my ability to explain these concepts concisely. But a few thoughts here.

The advantage of fully explicit meta-honesty, telling people exactly under what conditions you would lie or refuse to share information, is that it protects a system of full, reliable honesty.

The problem with fully explicit meta-honesty is that it vastly expands the necessary amount of Glomarization *to say exactly when you would use it.*

Eliezer correctly points out that if the Feds ask you where you were last night, your answer of 'I can neither confirm or deny where I was last night' is going to sound mighty suspicious regardless of how often you answer that way. Saying 'none of your goddamn business' is only marginally better. Also, letting them know that you always refuse to answer that question might not be the best way to make them think you're *less* suspicious.

This means both that full Glomarization isn't practical unless (this actually does come up) your response to a question can reliably be 'that's a trap!'.

However, *partial* Glomarization is fine. As long as you mix in *some* refusing to answer when the answer wouldn't hurt you, people don't know much. Most importantly, *they don't know how often you'd refuse to answer.*

If the last five times you've refused to answer if there was a dragon in your garage, there was a dragon in your garage, your refusal to answer is rather strong evidence there's a dragon in your garage.

If it only happened one of the last five times, then there's certainly a Bayesian update one can make, but you don't know how often there's a Glamorization there, so it's hard to know how much to update on that. The key question is, what's the threshold where they feel the need to look in your garage? Can you muddy the waters enough to avoid that?

Once you're doing that, it is *almost certainly fine* to answer 'no' when it especially matters that they know there isn't a dragon there, *because they don't know when it's important, or what rule you're following.* If you went and told them exactly when you answer the question, it would be bad. But if they're not sure, it's fine.

One can complement that by understanding how conversations and topics develop, and not set yourself up for questions you don't want to answer. If you have a dragon in your garage and don't want to lie about it or reveal that it's there, it's a really bad idea to talk about the idea of dragons in garages. Someone is going to ask. So when your refusal to answer would be suspicious, especially when it would be a potential sign of a heretical belief, the best strategy is to not get into position to get asked.

Which in turn, means avoiding perfectly harmless things gently, invisibly, without saying that this is what you're doing. Posts that don't get written, statements not made, rather than questions not answered. As a new practitioner of such arts, hard and fast rules are good. As an expert, they only serve to give the game away. '

Remember the illusion of transparency. Your counterfactual selves would need to act differently. But if no one knows that, it's not a problem.

[Math] Towards Proof Writing as a Skill In Itself

In the vein of Nate, David and TT, I'm currently reading through and working on a review of Halmos's *Naïve Set Theory*, from the MIRI course reading list.

My background in higher mathematics is so far two 300-level courses I have taken the past two quarters at Northwestern University:

- MATH 306, Combinatorics, which was mostly calculations and light on formal proofs until the end, and
- MATH 300, Introduction to Higher Mathematics, which is about... Nothing *but* formal proofs.

MATH-300 is intentionally limited in scope so that we only prove trivial statements, until near the very end. At first I thought this would be an easy A, and a bit of a waste of time. It turned out to be neither of those things.

Writing rigorous proofs, when you don't have a lot of practice, is always harder than you would expect, even for trivially obvious statements. A few reasons I noticed:

- You need to hew very closely to definitions and "allowed actions" in terms of inference. It's easy to make an 'obvious' logical leap that fails to convince the reader at all. It truly is like learning to cross inferential distances at a snail's pace.
- You need to learn the common tricks of the trade -- for example, "A if and only if B" is usually proved by proving "if A then B" *and then separately proving* "if B then A", making it essentially a 2 sub-proof project. "If A then B" is logically equivalent to "if not B then not A", which is sometimes much easier to prove (EDIT: although discouraged, because this makes the proof non-constructive - see comments). Et cetera, et cetera.
- You need to become diligent about proofreading, especially if you're typesetting with LaTeX**.** A single misplaced symbol will cost you a point, because the whole point is to drill the rigor into you to *proofread your damn homework.*

Imagine trying to build all of that, while *actually learning new concepts*. It's going to take you forever.

(Personal story time, feel free to skip.) I actually started with another proof-based course, Graph Theory. I dropped it, not because my grades were at all poor, but because my homework took. So. Damn. *Long*. And after hours of effort, I would still lose points for very small errors in places I thought were perfect.

It was infuriating. I backed off, realized that I hadn't built my skills in the right order, and dropped the course, so that I would only take the proof writing one this quarter. I don't regret that at all. The next time I take Graph Theory, I am confident that it will go much smoother, because I actually know *how* to write a proof.

(Having the first few homeworks already done in LaTeX won't hurt, either. ☺)

Nate, David and TT all remark on how *NST* is a dense read. Dense yes, difficult not necessarily at all. I'm finding my experience after a proof writing class to make the text very easy to read, despite (or maybe due to!) the frequent breaks to attempt proofs of the authors' statements myself.

My *NST* experience suggests to me, then, that proof writing is an excellent example of a component skill. Principle 4 of How Learning Works states that

> To develop mastery, students must acquire component skills, [*then*] practice integrating them, and [*finally*] know when to apply what they have learned. [emphasis mine]

So learn it in isolation first, if you can. It will make all future endeavors in proof-based math much smoother.

Simplified Poker

This is intended as a three-part sequence. Part two will go over my strategy. Part three will reveal the results and discuss some implications.

In the same class in which we later played The Darwin Game, we played a less complex game called Simplified Poker. As in The Darwin Game, we were given the rules and asked to submit instructions for a computer program that would play the game, and the professor would then code our programs for us.

The rules of Simplified Poker are as follows:

Game is played with a 3-card deck, with the cards labeled 1, 2 and 3.

Each hand, the players alternate who goes first, each player antes one chip and is dealt one card.

The first player can bet one chip, or check.

If the first player bets, the second player can either call the one chip bet, or fold.

If the first player checks, the second player can either also check, or can bet. If the second player bets, the first player can either call the one chip bet, or fold.

There is at most one bet per hand, as neither player is allowed to raise.

If either player folds, the other wins the pot of 2 chips and takes back their 1 chip bet. Neither card is shown. If neither player folds – either both players check, or there is a bet and a call – then both cards are revealed and the player with the higher card takes all 4 chips.

In the class, all programs would play a round robin with all other programs, with 50 hands per match. Your goal is to maximize the average number of chips won over all rounds – note that how many opponents you beat does not matter, only the number of chips won.

The game is simple. A lot, but far from all, of your decisions are forced. There’s no weird trick, but optimal play still isn’t obvious. I’ll pause here to allow and encourage thinking about what strategy you’d submit.

The Alignment Newsletter #11: 06/18/18

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

Turns out the survey link in the last email was broken, sorry about that and thanks to everyone who reported it to me. Here's the correct link.

Highlights

Learning to Follow Language Instructions with Adversarial Reward Induction *(Dzmitry Bahdanau et al)*: Adversarial Goal-Induced Learning from Examples (AGILE) is a way of training an agent to follow instructions. The authors consider a 5x5 gridworld environment with colored shapes that the agent can manipulate. The agent is given an instruction in a structured domain-specific language. Each instruction can correspond to many goal states -- for example, the instruction corresponding to "red square south of the blue circle" has many different goal states, since only the relative orientation of the shapes matters, not their absolute positions.

The key idea is to learn two things simultaneously -- an encoding of *what* the agent needs to do, and a policy that encodes *how* to do it, and to use these two modules to train each other. The "what" is encoded by a discriminator that can classify (state, instruction) pairs as either being a correct goal state or not, and the "how" is encoded by a policy. They assume they have some human-annotated goal states for instructions. The discriminator is then trained with supervised learning, where the positive examples are the human-annotated goal states, and the negative examples are states that the policy achieves during training (which are usually failures). The policy is trained using A3C with a reward function that is 1 if the discriminator says the state is more likely than not to be a goal state, and 0 otherwise. Of course, if the policy actually achieves the goal state, there is no way of knowing this apart from the discriminator -- so by default *all* of the states that the policy achieves (including goal states) are treated as negative examples for the dsicriminator. This leads to the discriminator getting slightly worse over time as the policy becomes better, since it is incorrectly told that certain states are not goal states. To fix this issue, the authors drop the top 25% of states achieved by the policy that have the highest probability of being a goal state (according to the discriminator).

The authors compare AGILE against A3C with the true reward function (i.e. the reward function implied by a perfect discriminator) and found that AGILE actually performed *better*, implying that the inaccuracy of the discriminator actually *helped* with learning. The authors hypothesize that this is because when the discriminator incorrectly rewards non-goal states, it is actually providing useful reward shaping that rewards progress towards the goal, leading to faster learning. Note though that A3C with an auxiliary reward prediction objective performed best. They have several other experiments that look at individual parts of the system.

My opinion: I like the idea of separating "what to do" from "how to do it", since the "what to do" is more likely to generalize to new circumstances. Of course, this can also be achieved by learning a reward function, which is one way to encode "what to

do". I'm also happy to see progress on the front of learning what humans want where we can take advantage of adversarial training that leads to a natural curriculum -- this has been key in many systems, most notably AlphaZero.
I'm somewhat surprised that dropping the top 25% of states ranked highly by the discriminator works. I would have guessed that states that are "near" the goal states might be misclassified by the discriminator, and the mistake will never be fixed because those states will always be in the top 25% and so will never show up as negative examples. I don't know whether I should expect this problem to show up in other environments, or whether there's a good reason to expect it won't happen.

I'm also surprised at the results from one of their experiments. In this experiment, they trained the agent in the normal environment, but then made red squares immovable in the test environment. This only changes the dynamics, and so the discriminator should work just as well (about 99.5% accuracy). The policy performance tanks (from 98% to 52%), as you'd expect when changing dynamics, but if you then finetune the policy, it only gets back to 69% success. Given that the discriminator should be just as accurate, you'd expect the policy to get back to 98% accuracy. Partly the discrepancy is that some tasks become unsolvable when red squares are immovable, but they say that this is a small effect. My hypothesis is before finetuning, the policy is very certain of what to do, and so doesn't explore enough during finetuning, and can't learn new behaviors effectively. This would mean that if they instead retrained the policy starting from a random initialization, they'd achieve better performance (though likely requiring many more samples).

A general model of safety-oriented AI development *(Wei Dai)*: A general model for developing safe powerful AI systems is to have a team of humans and AIs, which continually develops and adds more AIs to the team, while inductively preserving alignment.

My opinion: I'm glad this was finally written down -- I've been calling this the "induction hammer" and have used it a lot in my own thinking. Thinking about this sort of a model, and in particular what kinds of properties we could best preserve inductively, has been quite helpful for me.

AGI Strategy - List of Resources: Exactly what it sounds like.

Technical AI alignment

Agent foundations

Counterfactual Mugging Poker Game *(Scott Garrabrant)*: This is a variant of counterfactual mugging, in which an agent doesn't take the action that is locally optimal, because that would provide information in the counterfactual world where one aspect of the environment was different that would lead to a large loss in that setting.

My opinion: This example is very understandable and very short -- I haven't summarized it because I don't think I can make it any shorter.

Weak arguments against the universal prior being malign *(X4vier)*: In an earlier post, Paul Christiano has argued that if you run Solomonoff induction and use its predictions for important decisions, most of your probability mass will be placed on universes with

intelligent agents that make the right predictions so that their predictions will influence your decisions, and then use that influence to manipulate you into doing things that they value. This post makes a few arguments that this wouldn't actually happen, and Paul responds to the arguments in the comments.

My opinion: I still have only a fuzzy understanding of what's going on here, so I'm going to abstain from an opinion on this one.

Prerequisities: What does the universal prior actually look like?

Learning human intent

Learning to Follow Language Instructions with Adversarial Reward Induction *(Dzmitry Bahdanau et al)*: Summarized in the highlights!

An Efficient, Generalized Bellman Update For Cooperative Inverse Reinforcement Learning *(Dhruv Malik, Malayandi Palaniappan et al)*: Previously, Cooperative Inverse Reinforcement Learning (CIRL) games were solved by reducing them to a POMDP with an exponentially-sized action space, and then solving with POMDP algorithms that are exponential in the size of the action space, leading to a doubly-exponential algorithm. This paper leverages the fact that the human has perfect information to create a modified Bellman update that still computes the optimal policy, but no longer requires an exponential action space. The modified Bellman update works with the human's policy, and so we can now swap in more accurate models of the human, including eg. noisy rationality (whereas previously the human had to be exactly optimal). They show huge speedups in experiments, and discuss some interesting qualitative behavior that arises out of CIRL games -- for example, sometimes the human *waits* instead of making progress on the task, because it is a good signal to the robot of what the human wants.

My opinion: I'm excited by this improvement, since now we can actually solve non-trivial CIRL games -- one of the games they solve has around 10 billion states. With this we can run experiments with real humans, which seems really important, and the paper does mention a very preliminary pilot study run with real humans.

Prerequisities: Cooperative Inverse Reinforcement Learning

Learning a Prior over Intent via Meta-Inverse Reinforcement Learning *(Kelvin Xu et al)*: For complex rewards, such as reward functions defined on pixels, standard IRL methods require a large number of demonstrations. However, many tasks are very related, and so we should be able to leverage demonstrations from one task to learn rewards for other tasks. This naturally suggests that we use meta learning. The authors adapt MAML to work with maximum entropy IRL (which requires differentiating through the MaxEnt IRL gradient). They evaluate their approach, called MandRIL, on a navigation task whose underlying structure is a gridworld, but the state is represented as an image so that the reward function is nonlinear and requires a convnet.

My opinion: In one of the experiments, the baseline of running IRL from scratch performed second best, beating out two other methods of meta-learning. I'd guess that this is because both MandRIL and standard IRL benefit from assuming the maxent IRL distribution over trajectories (which I believe is how the demonstrations were synthetically generated), whereas the other two meta learning baselines do not have any such assumption, and must learn this relationship.

Imitating Latent Policies from Observation *(Ashley D. Edwards et al)*: Typically in imitation learning, we assume that we have access to demonstrations that include the actions that the expert took. However, in many realistic settings we only have access to state observations (eg. driving videos). In this setting, we could still infer a reward function and then use reinforcement learning (RL) to imitate the behavior, but this would require a lot of interaction with the environment to learn the dynamics of the environment. Intuitively, even demonstrations with only states and no actions should give us a lot of information about the dynamics -- if we can extract this information, then we would need much less environment interaction during RL. (For example, if you watch a friend play a video game, you only see states, not actions; yet you can infer a lot about the game rules and gameplay.) The key idea is that each action probably causes similar effects on different states. So, they create a model with hidden action nodes z, and use the state observations to learn a policy P(z | s) and dynamics s' = g(s, z) (they assume deterministic dynamics). This is done end-to-end with neural nets, but essentially the net is looking at the sequence of states and figuring out how to assign actions z to each s (this is P(z | s)), such that we can learn a function g(s, z) that outputs the next observed state s'. Once this is trained, intuitively g(s, z) will already have captured most of the dynamics, and so now we only require a small number of environment actions to figure out how the true actions a correspond to the hidden actions z -- concretely, we train a model P(a | s, z). Then, in any state s, we first choose the most likely hidden action z *according to P(z | s), and then the most likely action a* according to P(a | s, z*).

My opinion: The intuition behind this method makes a lot of sense to me, but I wish the experiments were clearer in showing how the method compares to other methods. They show that, on Cartpole and Acrobat, they can match the results of behavioral cloning with 50,000 state-action pairs using 50,000 state observations and 200 environment interactions, but I don't know if behavioral cloning actually needed that many state-action pairs. Similarly, I'm not sure how much environment interaction would be needed if you inferred a reward function but not the dynamics, since they don't compare against such a method. I'm also unclear on how hard it is to assign transitions to latent actions -- they only test on MDPs with at most 3 actions, it's plausible to me that with more actions it becomes much harder to figure out which hidden action a state transition should correspond to.

Preventing bad behavior

Worrying about the Vase: Whitelisting *(TurnTrout)*: It's really hard to avoid negative side effects because explicitly listing out all possible side effects the agent should avoid would be far too expensive. The issue is that we're trying to build a blacklist of things that can't be done, and that list will never be complete, and so some bad things will still happen. Instead, we should use whitelists, because if we forget to add something to the whitelist, that only limits the agent, it doesn't lead to catastrophe. In this proposal, we assume that we have access to the agent's ontology (in current systems, this might be the output of an object detection system), and we operationalize an "effect" as the transformation of one object into another (i.e. previously the AI believed an object was most likely an A, and now it believes it is most likely a B). We then whitelist allowed transformations -- for example, it is allowed to transform a carrot into carrot slices. If the agent causes any transformations not on the whitelist (such as "transforming" a vase into a broken vase), it incurs a negative reward. We also don't have to explicitly write down the whitelist -- we can provide demonstrations of acceptable behavior, and any transitions in these demonstrations

can be added to the whitelist. The post and paper have a long list of considerations on how this would play out in a superintelligent AI system.

My opinion: Whitelisting seems like a good thing to do, since it is safe by default. (Computer security has a similar principle of preferring to whitelist instead of blacklist.) I was initially worried that we'd have the problems of symbolic approaches to AI, where we'd have to enumerate far too many transitions for the whitelist in order to be able to do anything realistic, but since whitelisting could work on learned embedding spaces, and the whitelist itself can be learned from demonstrations, this could be a scalable method. I'm worried that it presents generalization challenges -- if you are distinguishing between different colors of tiles, to encode "you can paint any tile" you'd have to whitelist transitions (redTile -> blueTile), (blueTile -> redTile), (redTile -> yellowTile) etc. Those won't all be in the demonstrations. If you are going to generalize there, how do you *not* generalize (redLight -> greenLight) to (greenLight -> redLight) for an AI that controls traffic lights? On another note, I personally don't want to assume that we can point to a part of the architecture as the AI's ontology. I hope to see future work address these challenges!

Handling groups of agents

Adaptive Mechanism Design: Learning to Promote Cooperation *(Tobias Baumann et al)*

Multi-Agent Deep Reinforcement Learning with Human Strategies *(Thanh Nguyen et al)*

Interpretability

Neural Stethoscopes: Unifying Analytic, Auxiliary and Adversarial Network Probing *(Fabian B. Fuchs et al)*

Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning *(Leilani H. Gilpin et al)*

Miscellaneous (Alignment)

A general model of safety-oriented AI development *(Wei Dai)*: Summarized in the highlights!

To Trust Or Not To Trust A Classifier *(Heinrich Jiang, Been Kim et al)*: The confidence scores given by a classifier (be it logistic regression, SVMs, or neural nets) are typically badly calibrated, and so it is hard to tell whether or not we should trust our classifier's prediction. The authors propose that we compute a *trust score* to tell us how much to trust the classifier's prediction, computed from a training set of labeled datapoints. For every class, they filter out some proportion of the data points, which removes outliers. Then, the trust score for a particular test point is the ratio of (distance to nearest non-predicted class) to (distance to predicted class). They have theoretical results showing that a high trust score means that the classifier likely agrees with the Bayes-optimal classifier, as well as empirical results showing that this method does better than several baselines for determining when to trust a classifier. One cool thing about this method is that it can be done with any representation of the input data points -- they find that working with the activations of deeper layers of a neural net improves the results.

My opinion: I'm a big fan of trying to understand when our AI systems work well, and when they don't. However, I'm a little confused by this -- ultimately the trust score is just comparing the given classifier with a nearest neighbor classifier. Why not just use the nearest neighbor classifier in that case? This paper is a bit further out of my expertise than I'd like to admit, so perhaps there's an obvious answer I'm not seeing.

Podcast: Astronomical Future Suffering and Superintelligence with Kaj Sotala *(Lucas Perry)*

Near-term concerns

Adversarial examples

Defense Against the Dark Arts: An overview of adversarial example security research and future research directions *(Ian Goodfellow)*

AI strategy and policy

AGI Strategy - List of Resources: Summarized in the highlights!

Accounting for the Neglected Dimensions of AI Progress *(Fernando Martinez-Plumed et al)*

Artificial Intelligence and International Affairs: Disruption Anticipated *(Chatham House)*

India's National Strategy for Artificial Intelligence

AI capabilities

Reinforcement learning

Self-Imitation Learning *(Junhyuk Oh et al)*

Deep learning

Neural scene representation and rendering *(S. M. Ali Eslami, Danilo J. Rezende et al)*

Improving Language Understanding with Unsupervised Learning *(Alec Radford et al)*

Meta learning

Unsupervised Meta-Learning for Reinforcement Learning *(Abhishek Gupta et al)*

Bayesian Model-Agnostic Meta-Learning *(Taesup Kim et al)*

News

Research Scholars Programme: From the website: "The Future of Humanity Institute is launching a Research Scholars Programme, likely to start in October 2018. It is a selective, two-year research programme, with lots of latitude for exploration as well as significant training and support elements. We will offer around six salaried positions to early-career researchers who aim to answer questions that shed light on the big-picture questions critical to humanity's wellbeing. We are collecting formal applications to the programme from now until 11 July, 2018."

Announcing the second AI Safety Camp *(Anne Wissemann)*: I forgot to mention last week that the second AI safety camp will be held Oct 4-14 in Prague.

Human-aligned AI Summer School: The first Human-aligned AI Summer School will be held in Prague from 2nd to 5th August, with a focus on "learning from humans" (in particular, IRL and models of bounded rationality). Applications are open till July 14, but may close sooner if spots are filled up.

spaced repetition & Darwin's golden rule

This is a linkpost for https://mindhacks.com/2018/02/26/spaced-repetition-darwins-golden-rule/

We Agree: Speeches All Around!

In the Catalan autobiography of James I *Llibre dels Fets,* King James often describes the advice given to him by different nobles and princes of the Church (read bishops). Oftentimes they disagree; sometimes he turns out right, and sometimes they turn out the wiser counsellors. Scholars often regard this frequent decision-making dialogue as evidence that James wanted not only to give an account of the great accomplishments of his life, but also provide insight for future kings and ministers of Aragon-Catalonia. There is much to say about the nature of this advice, the strategic and tactical reasoning, the difficulty of passing down rational statesmanship, and interrogation into just how "rational" this statesmanship actually was.

I am not going to focus on those issues. Instead, I want to bring to light a common knowledge dynamic I noticed in this book that resonated in my daily life.

My day job requires a lot of meetings. Oftentimes in these meetings my colleagues and I will hit on an agreed course of action, but then instead of saying, "We are agreed. Let's go!" We will continue talking ourselves into the decision. Once a decision has been reached, each person inexplicably waxes poetic about their own reason for why they believe this is a good or right decision. This happens quite frequently, I do not think anyone recognizes it as weird. To be clear, this is not part of some in-house "Guideline For Decision-Making"; it is a spontaneous event of human interaction.

Up until this week, I thought this exercise was either an attempt to cover up uncertainty or a waste of time. But perhaps there is some utility here. Is this practice a way creating more agreeance? Congratulating ourselves on being in charge? What's the deal? Is it a way of rebuilding bonds that may have been strained over the course of discussion? Or is it just a 'Midwestern USA' thing?

James I helped me see the light. Before the invasion of the island of Mallorca, the *Corts* and councils convened to decide whether to invade. Into the mouths of a noble merchant, a general, a landed aristocrat, and a bishop additional words of approval came *after* they had already decided to launch this campaign. Since the campaign had already been approved in prior discussion by leading parties, why do they need more words of approval again after the decision has been made?

I think the answer is that although these speeches might be boring to read or a seeming waste of a Wednesday afternoon at work, they also provide an additional fact for everyone present. We know that everyone approves the course of action. Now in addition, we also know why everyone approves the course of action, what their slant, and what their motivations. From these, we can adjust our beliefs about to what extent and under what circumstances the other actors will support the course of action – how far are the others willing to go to support this? This additional knowledge should facilitate future coalitions, strategy, and decision-making. The more we understand each other's motivations, the more we can communicate effectively, find shared goals, and create a dynamic organization, one which can conquer western Mediterranean islands.

Next time, you are impatient hearing the reasons for a course of action you already agree with, it's not the course of action which you can learn about, but common knowledge about the motivations and interests of other actors. Common knowledge

about intentions within the coalition is the first step to sustained conquests... err success.

Anthropics and Fermi

tl;dr *There is no well-defined "probability" of intelligent life in the universe. Instead, we can use proper scoring functions to penalise bad probability estimates. If we average scores across all existent observers, we get SSA-style probability estimates; if we sum them, we get SIA-style ones.*

When presenting "anthropic decision theory" (the anthropic variant of UDT/FDT), I often get the response "well, that's all interesting, but when we look out to the stars with telescopes, probes, what do we *really* expect to see?" And it doesn't quite answer the question to say that "*really expect to see* is incoherent".

So instead of evading the question, let's try and answer it.

Proper scoring rules

Giving the best guess about the probability of X, is the same as maximising a proper scoring rule conditional on X. For example, someone can be asked to name a $0 \leq p \leq 1$, and they will get a reward of $-(I_X - p)^2$, where I_X is the indicator variable that which is 1 if X happens and 0 if it doesn't.

Using a logarithmic proper scoring rule, Wei Dai demonstrated that an updatless agent can behave like an updating one.

So let's apply the proper scoring rule to the probability that there is an alien civilization in our galaxy. As above, you guess p, and are given $-(1-p)^2$ if there is an alien civilization in our galaxy, and $-p^2$ if there isn't.

Summing over different agents

But how do we combine estimates from different agents? If you're merely talking about probability - there are several futures you could experience, and you don't know which yet - then you simply take an expectation over these.

But what about duplication, which is not the same as probability? What if there are two identical versions of you in the universe, but you expect them to diverge soon, and maybe one will find aliens in their galaxy while the other will not?

One solution is to treat duplication as probability. If your two copies diverge, that's exactly the same as if there was a 50-50 split into possible futures. In this case, the total score is the *average* of all scores in this one universe. In that case, one should use SSA-style probability, and update one's estimates using that.

Or we could treat duplication as separate entities, and ensure that as many as possible are as correct as possible. This involves totalling up the scores in the

universe, and so we use SIA-style probability.

In short:

- SSA: in every universe, the average score is as good as can be.
- SIA: for every observer, the score is as good as can be.

Thus the decision between SSA-style and SIA-style probabilities, is the decision as to which summed proper scoring function one tries to maximise.

So, which of these approaches is correct? Well, you can't say from intrinsic factors. How do you know that any probability you utter is correct? Frequentists talk about long-run empirical frequencies, while Bayesians allow themselves to chunk a lot of this empirical data into the same category (your experience of people in the construction industry is partially applicable to academia). But, all in all, both are correcting their estimates according to observations. And the two scoring totals are just two ways of correcting this estimate - neither is better than the other.

Reference classes and linked decisions

I haven't yet touched upon the reference class issue. Independently of what we choose to sum over - the scores of all human estimates, all conscious entities, all humans subjectively indistinguishable from me - by choosing our own estimates, we are affecting the estimates of those whose estimates are 'linked' with ours (in the same way that our decisions are linked with those of identical copies in the Prisoner's Dilemma). If we total up the scores, then as long as the summing includes all 'linked' scores, then it doesn't matter how many other scores are included in the total: that's just an added constant, fixed in any given universe, that we cannot change. This is the decision theory version of "SIA doesn't care about reference classes".

If we are averaging, however, then it's very important which scores we use. If we have large reference classes, then the large amount of other observers will dilute the effect of linked decisions. Thus universes will get downgraded in probability if they contain a higher proportion of non-linked estimates to linked ones. This is the decision theory version of "SSA is dependent on your choice of reference classes".

However, unlike standard SSA, decision theory has a natural reference class: just use the class of all linked estimates.

Boltzmann brains and simulations

Because the probability is defined in terms of a score in the agent's "galaxy", it makes sense to exclude Boltzmann brains from the equation, as their entire beliefs are wrong - they don't inhabit the galaxy they believe they are in, and their believed reality is entirely wrong. So from a decision theoretic perspective, so that the scoring rule makes sense, we should exclude them.

Simulations are more tricky, because they may discovered simulated aliens within their simulated galaxies. If we have a well defined notion of simulation - I'd argue that in general, that term is ill-defined - then we can choose to include or not include that in the calculation, and both estimates would makes perfect sense.

Loss aversion is not what you think it is

This is a linkpost for https://www.basilhalperin.com/blog/2015/12/loss-aversion-is-not-what-you-think-it-is/

Three types of "should"

Note: This post is about "should" in how people think, in human psychology -- not about "should" in some deeper/broader philosophical sense that might apply to general agents.

I think when people talk about how they "should" do something, there's basically three different types of motivation/shouldness that are meant, and I think noticing this helps make sense of the idea of "getting rid of 'should'" that some people talk about. I'll call the three "external", "internal", and "interalized". (I don't really like this terminology but it's what I've come up with. If you have a suggestion for something better, let me know.)

External here refers to things that are motivated entirely externally -- I don't particularly want to do X, but things will not go well for me if I don't. External "should" has no real moral component to it -- here by "moral" I mean that not in the broad consequentialist sense of what one should do (here by "should" I *do* mean that in a broader sense rather than a human sense!) but rather that thing that people think of as morality (like for example things that deal with other people's welfare and not one's own, or things being forbidden-or-allowed-or-mandatory).

Internal is the opposite, purely internal -- I feel that X is something that needs to be done and so I am internally motivated to do it. Well, OK -- that description has the problem that it doesn't do enough to distinguish it from the third one, "internalized". I'm hoping the distinction will become clear in a moment when I discuss the third.

Internalized is the nasty one that you want to avoid, the source of scrupulosity, the reason that people talk about "getting rid of 'should'". It's when you take someone *else's* morality, that you are not allowed to question, and internalize it as binding on yourself. As I said -- this is where scrupulosity comes from; it's not a good thing.

The thing is that if you're not already aware of the distinction it can be hard to describe internal or internalized in a way that couldn't also describe the other. Like, internalized masks itself as internal. Above I said that internal is things that *you* want to do, that *you* feel need to be done -- but the person in the grip of scrupulosity would just say, yes, *I* want to do these things, *I* feel they need to be done; the obligation is not externally imposed but a result of my own internal desire to do the right thing. Such a person would honestly have trouble recognizing the distinction.

But there is a distinction. The two, well, *feel* different. Internalized "should" feels, well, bad; it feels like something that oppresses you and gets in your way, even as it's notionally your own internal motivation. Whereas what I'm calling "internal should" doesn't. Another detectable difference, I think, is that internalized should has a certain indirectness to it; it's less "I want to do this thing", as it is "I want to do what is right, and I have concluded that this is what is right". But perhaps that's not the best distinguisher since I guess there are circumstances where internal should can have that indirectness as well.

So when people talk of "getting rid of 'should'", it seems to me they mean "internalized should". Taken literally it doesn't make a lot of sense -- you want to get rid of your motivation? You want to get rid of your notion of right and wrong? But fortunately a person can't actually get rid of their own internal should; but in the attempt to get rid of "should", they can free themselves from their internalized

shoulds, from *other* people's senses of right and wrong that they've tried to incoporate unquestioningly into their own. (External "should" has, as mentioned, no real moral component to it and so isn't relevant here.)

Anyway I think a number of discussions of "should" (again, in the human psychology sense, not in a deeper philosophical sense) make more sense in light of this distinction and the frequent failure to recognize it.

SIAM Lecture: How Paradoxes Shape Mathematics and Give Us Self-Verifying Computer Programs

This is a linkpost for http://meetings.siam.org/sess/dsp_programsess.cfm?SESSIONCODE=65101

Abstract:

> A paradox is a seeming contradiction. The liar's paradox is one of the best known: "This statement is a false." If the statement is true, then it is false; if it is false, then it is true.
>
> Paradoxes can be so amusing that we might think that paradoxes are nothing more than a game. However, paradoxes triggered a crisis in math a century ago when a paradox similar to the barber paradox was found: a barber named Bertie shaves exactly those who do not shave themselves. Does Bertie shave himself? If he does, then he doesn't; if he doesn't, then he does.
>
> Other clever paradoxes show us the disturbing limits of computation and mathematics. These results are mathematical bombshells.
>
> Today, we design computer programs that check that other computers programs have no bugs. Can computer programs be fed into themselves to check their own correctness? Or does paradox stop us in our tracks? And can we know that beneficial artificial intelligence will not turn evil when it starts to modify its own computer code?

The Alignment Newsletter #9: 06/04/18

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

Highlights

Playing hard exploration games by watching YouTube *(Yusuf Aytar, Tobias Pfaff et al)*: There are many YouTube videos demonstrating how to play levels of eg. Montezuma's Revenge. Can we use these demonstrations to solve the hard exploration tasks in Atari? One challenge is that the videos have slightly different visual properties (like color and resolution). They propose to learn a shared feature space by using an auxiliary loss where the network must predict the number of timesteps between two frames of a video, or to predict the delay between a video and audio clip from the same trajectory. Using this shared feature space, they can define a reward function that encourages the agent to take trajectories whose features match those of the demonstrations. In experiments they exceed human performance on Atari games with hard exploration problems.

My opinion: It seems to me that this is how we'll have to solve exploration in practice if we don't want to have a huge sample complexity, though I know other researchers are optimistic about solving exploration using curiosity or diversity. It's pretty exciting that they could use a source of data that was already present in the real world.

Technical AI alignment

Problems

The simple picture on AI safety *(alexflint)*: Argues that we should distill the problem of AI safety into a simple core. The author proposes it be distilled into two simple (but not easy) problems -- the technical engineering problem of how to build a safe superintelligence, and the coordination problem of how to prevent an unaligned superintelligence from being built first.

Iterated distillation and amplification

Amplification Discussion Notes *(William_S)*

Learning human intent

Learning Safe Policies with Expert Guidance *(Jessie Huang et al)*: Expert demonstrations can be consistent with many possible reward functions. Instead of simply trying to mimic the demonstration, the authors consider all possible rewards that are consistent with the demonstration, and then maximize the worst reward, leading to safe behavior.

My opinion: This is very related to Inverse Reward Design, where instead of maxmin planning we use risk-averse planning, and instead of considering all rewards compatible with an expert demonstration we consider all reward functions that are probable based on which reward function the designer wrote down.

Handling groups of agents

Scalable Centralized Deep Multi-Agent Reinforcement Learning via Policy Gradients *(Arbaaz Khan et al)*

Verification

Training verified learners with learned verifiers *(Krishnamurthy (Dj) Dvijotham, Sven Gowal, Robert Stanforth et al)*

Miscellaneous (Alignment)

How To Solve Moral Conundrums with Computability Theory *(Jongmin Jerome Baek)*

AI strategy and policy

How a Pentagon Contract Became an Identity Crisis for Google *(Scott Shane et al)*: After Google accepted a share of the contract for the Maven program run by the Defense Department, Google has been internally fractured, with many employees strongly opposing the use of AI for military applications.

My opinion: Stories like this make me optimistic that we can actually coordinate AI researchers to take appropriate safety precautions when developing advanced AI systems, even if the economic incentives point in the other direction (and I'm not sure they do).

AI capabilities

Reinforcement learning

Playing hard exploration games by watching YouTube *(Yusuf Aytar, Tobias Pfaff et al)*: Summarized in the highlights!

Meta-Gradient Reinforcement Learning *(Zhongwen Xu et al)*

Deep learning

Do Better ImageNet Models Transfer Better?: See Import AI

Meta learning

Meta-Learning with Hessian Free Approach in Deep Neural Nets Training *(Boyu Chen et al)*

May gwern.net newsletter

This is a linkpost for https://www.gwern.net/newsletter/2018/05

Resolving the Dr Evil Problem

Deep in Dr Evil's impregnable fortress (paraphrased):

> Dr Evil is just about to complete his evil plan of destroying the Earth, when he receives a message from the Philosophy Defence Force on Mars. They have created a clone in the exact same subjective situation Dr Evil now occupies; he believes he is Dr Evil and is currently in a recreation of the fortress. If the clone of Dr Evil tries to destroy the Earth, they will torture him, otherwise they will treat him well. Dr Evil wants to destroy the Earth, but he would prefer to avoid being tortured much, much more and he is now uncertain about whether he should surrender or not. Should Dr Evil surrender?

The paper then concludes:

> I conclude that Dr. Evil ought to surrender. I am not entirely comfortable with that conclusion. For if INDIFFERENCE is right, then Dr. Evil could have protected himself against the PDF's plan by (in advance) installing hundreds of brains in vats in his battlestation—each brain in a subjective state matching his own, and each subject to torture if it should ever surrender

This article will address two areas of this problem:

- Firstly, it argue that Dr Evil should surrender, however focusing on a different path, particularly what it means to "know" and how this is a leaky abstraction
- Secondly, it will argue that hundreds of brains in a vat would indeed secure him against this kind of blackmail

I'll note that this problem is closely related to The AI that Boxes You. Several people noted there that you could avoid blackmail by pre-committing to reboot the AI if it tried to threaten you, although my interest is in the rational behaviour of an agent who has failed to pre-commit.

Why Dr Evil Should Surrender

I think there's a framing effect in the question that is quite misleading. Regardless of whether we say, "You are Dr Evil. What do you do?" or "What should Dr Evil do?" we are assuming that you or a third-party Dr Evil knows that they are Dr Evil? However that's an assumption that needs to be questioned.

If you *knew* that you were Dr Evil, then knowing that a clone has been created and placed in a situation that appears similar wouldn't change what you should do if you don't care about the clone. However, you strictly can't ever actually *know* that you are Dr Evil. All you can know is that you have memories of being Dr Evil and you appear to be in Dr Evil's fortress (technically we could doubt this too. Maybe Dr Evil doesn't actually exist, but we don't have to go that far to prove our point).

Before this event, you would have placed the probability of you being Dr Evil really high as you had no reason to believe that you might have been a clone or in a simulation. After you receive the message, you have to rate the chance of you being a clone much higher and this breaks the leaky abstraction that normally allows you to say that you *know* you are Dr Evil.

If we did say that you knew you were Dr Evil, then on receiving the message, you would somehow have to magically come to un-know something without someone erasing your memories or otherwise interfering with your brain. However, since you only know that you have memories of being Dr Evil, you haven't lost information You've actually gained it, no nothing magical is happening at all.

Why Clones Protect Against the Attack

The idea of creating clones to protect yourself against these kinds of attacks seems weird. However, I will argue that this strategy is actually effective.

I'll first note that the idea of setting up a punishment to prevent you giving in to blackmail isn't unusual at all. It's well known that if you can credibly pre-commit, there's no incentive to blackmail you. If you have a device that will torture you if you surrender, then have no incentive to surrender unless the expected harm from not surrendering exceeds the torture.

Perhaps then this issue is that you want to protect yourself by intentionally limiting messing up your beliefs about what is true? There's no reason why this shouldn't be effective. If you can self-modify yourself to disbelieve any blackmail threat, no-one can blackmail you. One way to do this would be to self-modify yourself to believe you are in a simulation that will end just before you are tortured. Alternatively, you could protect yourself by self-modifying to believe that you would be tortured worse if you accepted the blackmail. Creating the clones is just another way of achieving this, though less effective as you only believe there is a probability that you will be tortured if you surrender.