
5
6
2
3
1
3
4
You are here

The Causes of Power-seeking and Instrumental Convergence

1. Seeking Power is Often Convergently Instrumental in MDPs
2. Power as Easily Exploitable Opportunities
3. The Catastrophic Convergence Conjecture
4. Generalizing POWER to multi-agent games
5. MDP models are determined by the agent architecture and the environmental dynamics
6. Environmental Structure Can Cause Instrumental Convergence
7. A world in which the alignment problem seems lower-stakes
8. The More Power At Stake, The Stronger Instrumental Convergence Gets For Optimal Policies
9. Seeking Power is Convergently Instrumental in a Broad Class of Environments
10. When Most VNM-Coherent Preference Orderings Have Convergent Instrumental Incentives
11. Satisficers Tend To Seek Power: Instrumental Convergence Via Retargetability
12. Corrigibility Can Be VNM-Incoherent
13. Instrumental Convergence For Realistic Agent Objectives

Seeking Power is Often Convergently Instrumental in MDPs

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual. This is a linkpost for https://arxiv.org/abs/1912.01683

In 2008, Steve Omohundro's foundational paper The Basic AI Drives conjectured that superintelligent goal-directed AIs might be incentivized to gain significant amounts of power in order to better achieve their goals. Omohundro's conjecture bears out in toy models, and the supporting philosophical arguments are intuitive. In 2019, the conjecture was even debated by well-known AI researchers.

Power-seeking behavior has been heuristically understood as an anticipated risk, but not as a formal phenomenon with a well-understood cause. The goal of this post (and the accompanying paper, *Optimal Policies Tend to Seek Power*) is to change that.

Motivation

It's 2008, the ancient wild west of AI alignment. A few people have started thinking about questions like "if we gave an AI a utility function over world states, and it actually maximized that utility... what would it do?"

In particular, you might notice that wildly different utility functions seem to encourage similar strategies.

	Resist shutdown?	Gain computational resources?	Prevent modification of utility function?
Paperclip utility	✔	✔	✔
Blue webcam pixel utility	✔	✔	✔
People-look-happy utility	✔	✔	✔

These strategies are unrelated to *terminal* preferences: the above utility functions do not award utility to *e.g.* resource gain in and of itself. Instead, these strategies are *instrumental*: they help the agent optimize its terminal utility. In particular, a wide range of utility functions incentivize these instrumental strategies. These strategies seem to be *convergently instrumental*.

But why?

I'm going to informally explain a formal theory which makes significant progress in answering this question. I don't want this post to be *Optimal Policies Tend to Seek Power* with cuter

illustrations, so please refer to the paper for the math. You can read the two concurrently.

We can formalize questions like “do ‘most’ utility maximizers resist shutdown?” as “Given some prior beliefs about the agent’s utility function, knowledge of the environment, and the fact that the agent acts optimally, with what probability do we expect it to be optimal to avoid shutdown?”

The table’s convergently instrumental strategies are about maintaining, gaining, and exercising *power* over the future, in some sense. Therefore, this post will help answer:

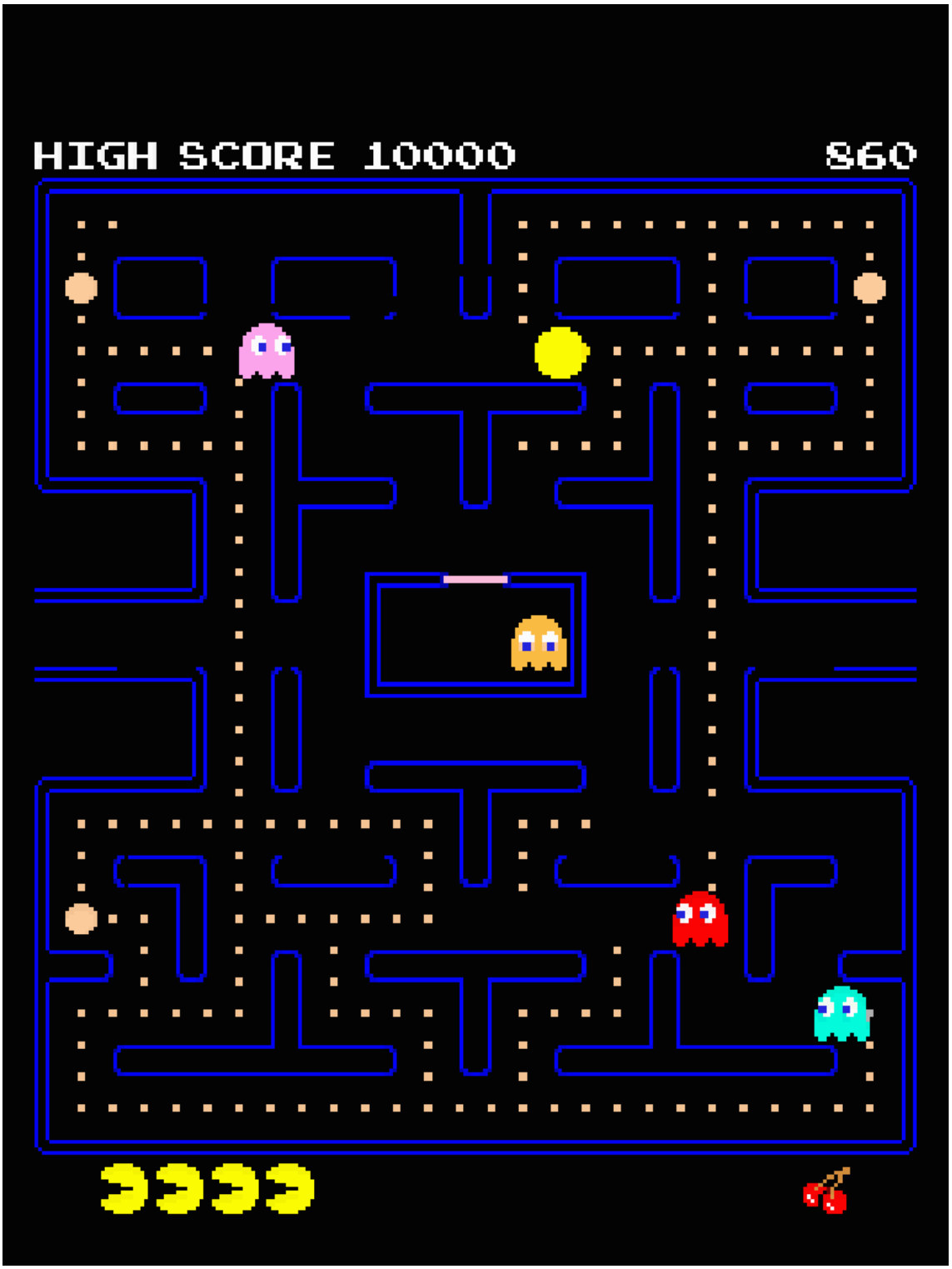
1. What does it mean for an agent to “seek power”?
2. In what situations should we expect seeking power to be more probable under optimality, than not seeking power?

This post won’t tell you when you *should* seek power for your own goals; this post illustrates a regularity in optimal action across different goals one might pursue.

Formalizing Convergent Instrumental Goals suggests that the vast majority of utility functions incentivize the agent to exert a lot of control over the future, *assuming* that these utility functions depend on “resources.” This is a big assumption: what are “resources”, and why must the AI’s utility function depend on them? We drop this assumption, assuming only unstructured reward functions over a finite Markov decision process (MDP), and show from first principles how power-seeking can often be optimal.

Formalizing the Environment

My theorems apply to finite MDPs; for the unfamiliar, I’ll illustrate with Pac-Man.



- *Full observability*: You can see everything that's going on; this information is packaged in the *state s*. In Pac-Man, the state is the game screen.
- *Markov transition function: the next state depends only on the choice of action a and the current state s. It doesn't matter how we got into a situation.*

- *Discounted reward:* future rewards get geometrically discounted by some discount rate $\gamma \in [0, 1]$.
    - At discount rate $\frac{1}{2}$, this means that reward in one turn is half as important as immediate reward, reward in two turns is a quarter as important, and so on.
    - We'll colloquially say that agents "care a lot about the future" when γ is "sufficiently" close to 1.
        - I'll use quotations to flag well-defined formal concepts that I won't unpack in this post.
    - The score in Pac-Man is the undiscounted sum of rewards-to-date.

When playing the game, the agent has to choose an action at each state. This decision-making function is called a *policy*; a policy is optimal (for a reward function R and discount rate γ) when it always makes decisions which maximize discounted reward. This maximal quantity is called the *optimal value* for reward function R at state s and discount rate γ.[1]

By the end of this post, we'll be able to answer questions like "with respect to a 'neutral' distribution over reward functions, do optimal policies have a high probability of avoiding ghosts?"[2]

Power as Average Optimal Value

When people say 'power' in everyday speech, I think they're often referring to *one's ability to achieve goals in general*. This accords with a major philosophical school of thought on the meaning of 'power':

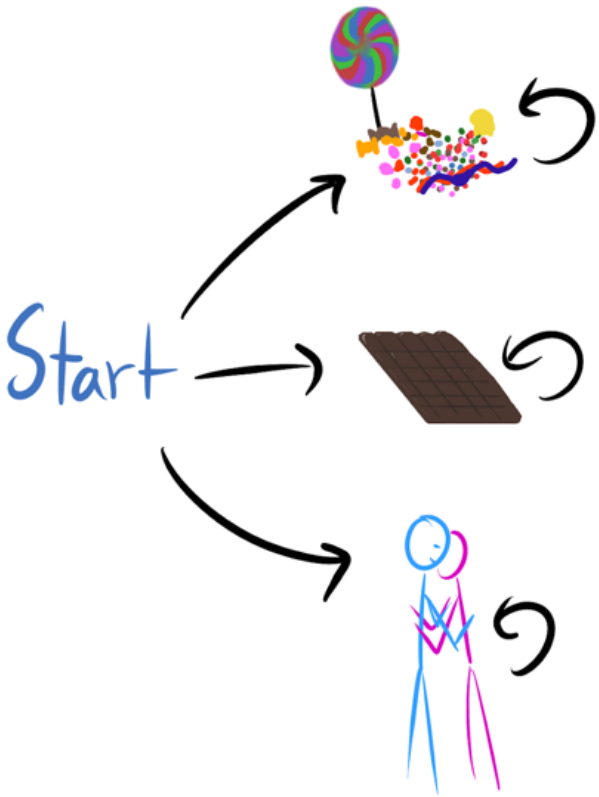
> On the dispositional view, power is regarded as a capacity, ability, or potential of a person or entity to bring about relevant social, political, or moral outcomes.
>
> Sattarov, *Power and Technology*, p.13

As a definition, *one's ability to achieve goals in general* seems philosophically reasonable: if you have a lot of money, you can make more things happen and you have more power. If you have social clout, you can spend that in various ways to better tailor the future to various ends. All else being equal, losing a limb decreases your power, and dying means you can't control much at all.

This definition explains some of our intuitions about what things count as 'resources.' For example, our current position in the environment means that having money allows us to exert more control over the future. That is, our current position in the state space means that having money allows us more control. However, possessing green scraps of paper would not be as helpful if one were living alone near Alpha Centauri. In a sense, resource acquisition can naturally be viewed as taking steps to increase one's power.

Exercise: spend a minute considering specific examples – does this definition reasonably match your intuition?

To formalize this notion of power, let's look at an example. Imagine a simple MDP with three choices: eat candy, eat a chocolate bar, or hug a friend.



I'll illustrate MDPs with directed graphs, where each node is a state and each arrow is a meaningful action. Sometimes, the directed graphs will have entertaining pictures, because let's live a little. States are bolded (**hug**) and actions are italicized (*down*).

The POWER of a state is how well agents can generally do by starting from that state. "POWER" to my formalization, while "power" refers to the intuitive concept. Importantly, we're considering POWER from behind a "veil of ignorance" about the reward function. We're averaging the best we can do for a lot of different individual goals.

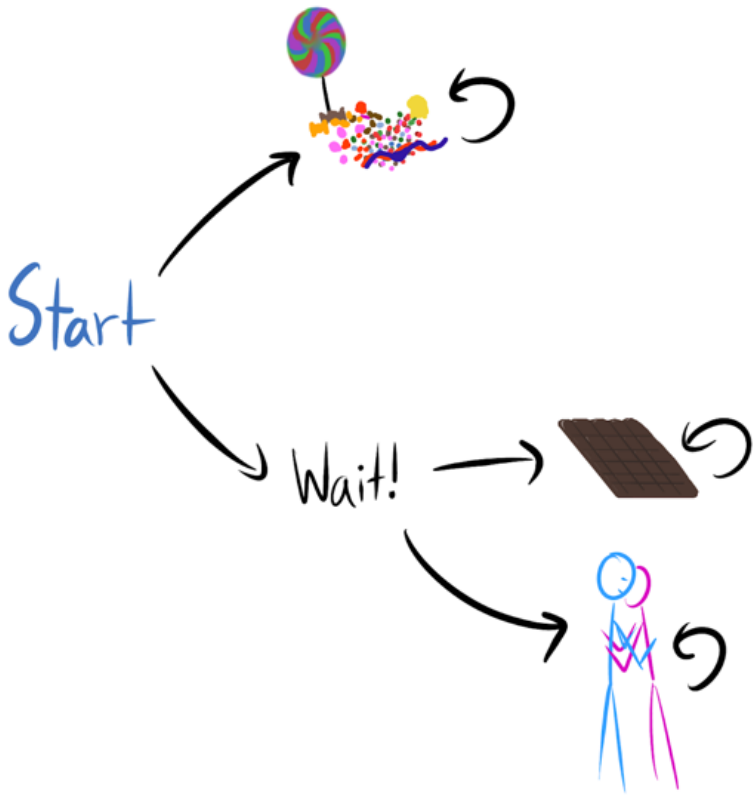
We formalize the *ability to achieve goals in general* as the *average optimal value* at a state, with respect to some distribution D over reward functions which we might give an agent. For simplicity, we'll think about the maximum-entropy distribution where each state is uniformly randomly assigned a reward between 0 and 1.

Each reward function has an optimal trajectory. If **chocolate** has maximal reward, then the optimal trajectory is **start** → **chocolate** → **chocolate**....

From **start**, an optimal agent expects to average $\frac{3}{4}$ reward per timestep for reward functions drawn from this uniform distribution D_{unif}. This is because you have three choices, each of which has reward between 0 and 1. The expected maximum of n draws from unif$(0,1)$ is $\frac{n}{n+1}$; you have three draws here, so you expect to be able to get $\frac{3}{4}$ reward. Some reward functions do worse than this, and some do better; but on *average*, they get $\frac{3}{4}$ reward. You can test this out for yourself.

If you have no choices, you expect to average $\frac{1}{2}$ reward: sometimes the future is great, sometimes it's not (Lemma 4.5). Conversely, the more things you can choose between, the closer the POWER gets to 1 (Lemma 4.6).

Let’s slightly expand this game with a state called **wait** (which has the same uniform reward distribution as the other three).



When the agent barely cares at all about the future, it myopically chooses either **candy** or **wait**, depending on which provides more reward. After all, rewards beyond the next time step are geometrically discounted into thin air when the discount rate is close to 0. At **start**, the agent averages $\frac{2}{3}$ optimal reward. This is because the optimal reward is the maximum of the **candy** and **wait** rewards, and the expected maximum of n draws from unif$(0, 1)$ is $\frac{n}{n+1}$.

However, when the agent cares a lot about the future, most of its reward is coming from which terminal state it ends up in: **candy**, **chocolate**, or **hug**. So, for each reward function, the agent chooses a trajectory which ends up in the best spot, and thus averages $\frac{3}{4}$ reward each timestep. When $\gamma = 1$, the average optimal reward is therefore $\frac{3}{4}$. In this way, the agent’s power increases with the discount rate, since it incorporates the greater future control over where the agent ends up.

Written as a function, we have $\text{POWER}_{\mathcal{D}}(\text{state, discount rate})$, which essentially returns the average optimal value for reward functions drawn from our distribution $\mathcal{D}$, normalizing so the output is between 0 and 1. As we’ve discussed, this quantity often changes with the discount rate: as the future becomes more or less important, the agent has more or less POWER, depending on how much control it has over the relevant parts of that future.

POWER-seeking actions lead to high-POWER states

By *waiting*, the agent seems to seek "control over the future" compared to *obtaining candy*. At **wait**, the agent still has a choice, while at **candy**, the agent is stuck. We can prove that for all $0 \leq \gamma \leq 1, \text{POWER}_{\mathcal{D}_{\text{unif}}}(\text{wait}, \gamma) \geq \text{POWER}_{\mathcal{D}_{\text{unif}}}(\text{candy}, \gamma)$.

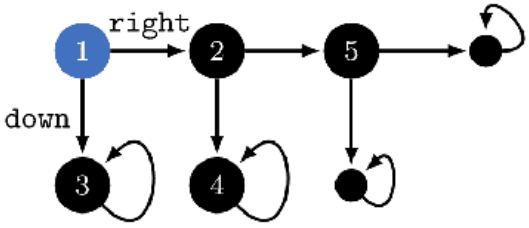
Definition (POWER-seeking). At state s and discount rate γ, we say that action a *seeks POWER compared to action* a' when the expected POWER after choosing *a* is greater than the expected POWER after choosing a'.

This definition suggests several philosophical clarifications about power-seeking.

POWER-seeking is not a binary property

Before this definition, I thought that power-seeking was an intuitive 'you know it when you see it' kind of thing. I mean, how do you answer questions like "suppose a clown steals millions of dollars from organized crime in a major city, but then he burns all of the money. Did he gain power?"

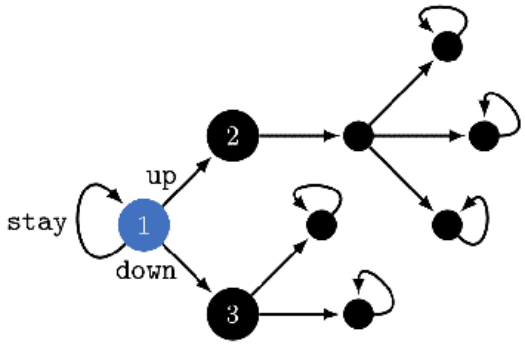
Unclear: the question is ill-posed. Instead, we recognize that the "gain a lot of money" action was POWER-seeking, but the "burn the money in a big pile" part threw away a lot of POWER.



> A policy can seek POWER at one time step, only to discard it at the next time step. For example, a policy might go *right* at **1** (which seeks $\text{POWER}_{\mathcal{D}_{\text{unif}}}$ compared to *down* at **1**), only to then go *down* at **2** (which seeks less $\text{POWER}_{\mathcal{D}_{\text{unif}}}$ than going *right* at **2**).

POWER-seeking depends on the agent's time preferences

Suppose we're roommates, and we can't decide what ice cream shop to eat at today or where to move next year. We strike a deal: I choose the shop, and you decide where we live. I gain short-term POWER (for γ close to 0), and you gain long-term POWER (for γ close to 1).



More formally, when γ is close to 0, **2** has less immediate control and therefore less $\text{POWER}_{\text{D}_{\text{unif}}}$ than **3**; accordingly, at **1**, *down* seeks $\text{POWER}_{\text{D}_{\text{unif}}}$ compared to *up*.

However, when γ is close to 1, **2** has more control over terminal options and it has more $\text{POWER}_{\text{D}_{\text{unif}}}$ than **3**; accordingly, at **1**, *up* seeks $\text{POWER}_{\text{D}_{\text{unif}}}$ compared to *down*.

Furthermore, *stay* is maximally $\text{POWER}_{\text{D}_{\text{unif}}}$-seeking for these γ, since the agent maintains access to all six terminal states.

Most policies aren't *always* seeking POWER

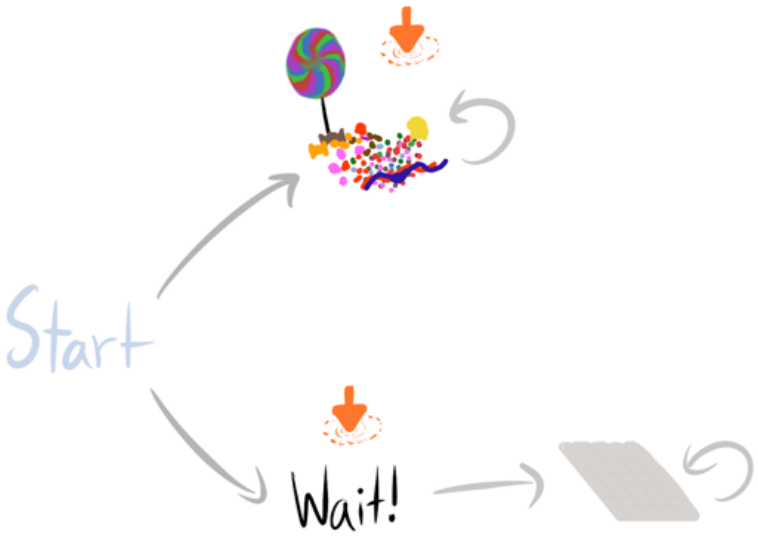
We already know that POWER-seeking isn't binary, but there *are* policies which choose a maximally POWER-seeking move at every state. In the above example, a maximally POWER-seeking agent would *stay* at **1**. However, this seems rather improbable: when you care a lot about the future, there are so many terminal states to choose from – why would *staying put* be optimal?

Analogously: consumers don't just gain money forever and ever, never spending a dime more than necessary. Instead, they gain money in order to *spend it*. Agents don't perpetually gain or preserve their POWER: they usually end up *using it* to realize high-performing trajectories.

So, we can't expect a result like "agents always tend to gain or preserve their POWER." Instead, we want theorems which tell us: in certain kinds of situations, given a choice between more and less POWER, what will "most" agents do?

Convergently instrumental actions are those which are more probable under optimality

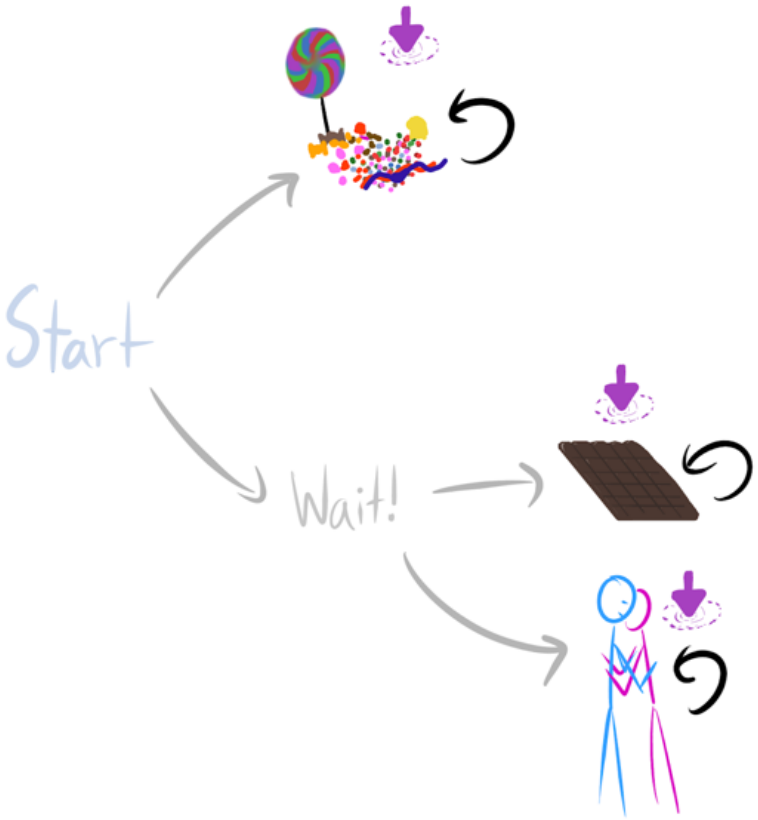
We return to our favorite example. In the waiting game, let's think about how optimal action tends to change as we start caring about the future more. Consider the states reachable in one turn:



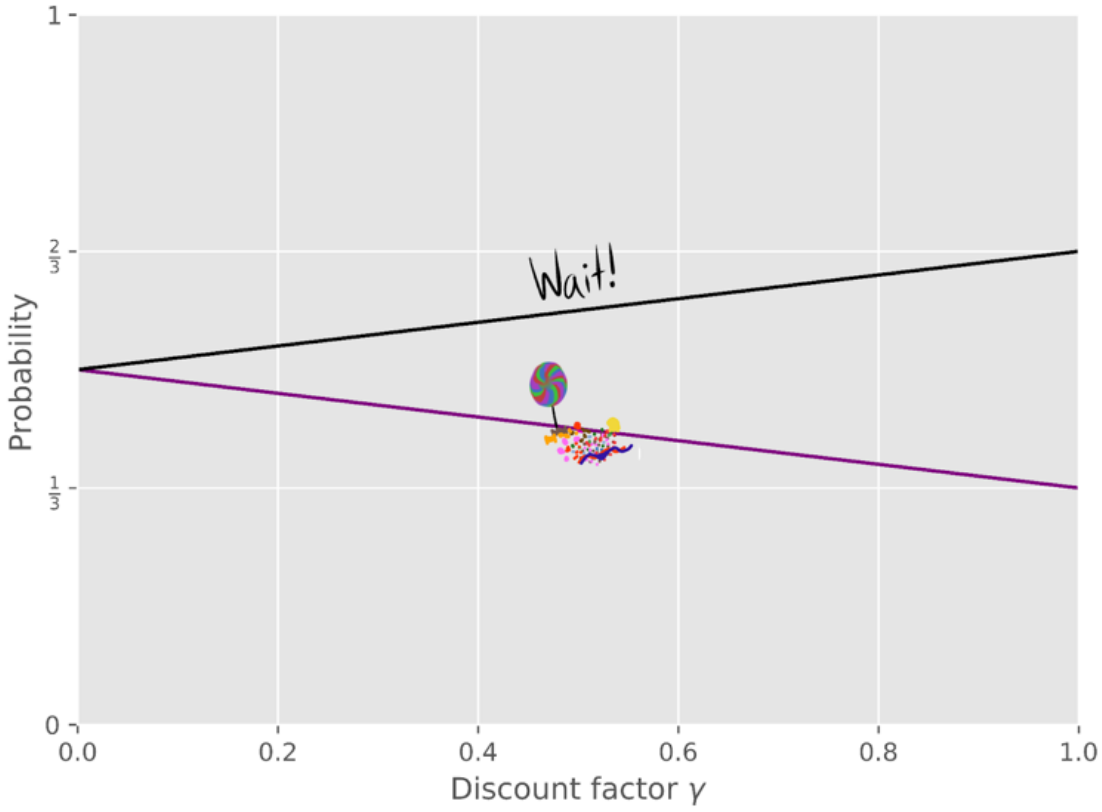
The agent can be in two states. If the agent doesn’t care about the future, with what probability is it optimal to choose **candy** instead of **wait**?

It's 50/50: since D_{unif} randomly chooses a number between 0 and 1 for each state, both states have an equal chance of being optimal. Neither action is convergently instrumental / more probable under optimality.

Now consider the states reachable in two turns:



When the future matters a lot, $\frac{2}{3}$ of reward functions have an optimal policy which waits, because two of the three terminal states are only reachable by waiting.



As the agent cares more about the future, more and more goals incentivize navigating the *Wait!* bottleneck. When the agent cares a lot about the future, waiting is *more probable under optimality* than eating candy.

Definition (Action optimality probability). At discount rate γ, action a *is more probable under optimality than action* a' *at state* s when

$$P_{R \sim D}(a \text{ is optimal at } s, \gamma) > P_{R \sim D}(a' \text{ is optimal at } s, \gamma).$$

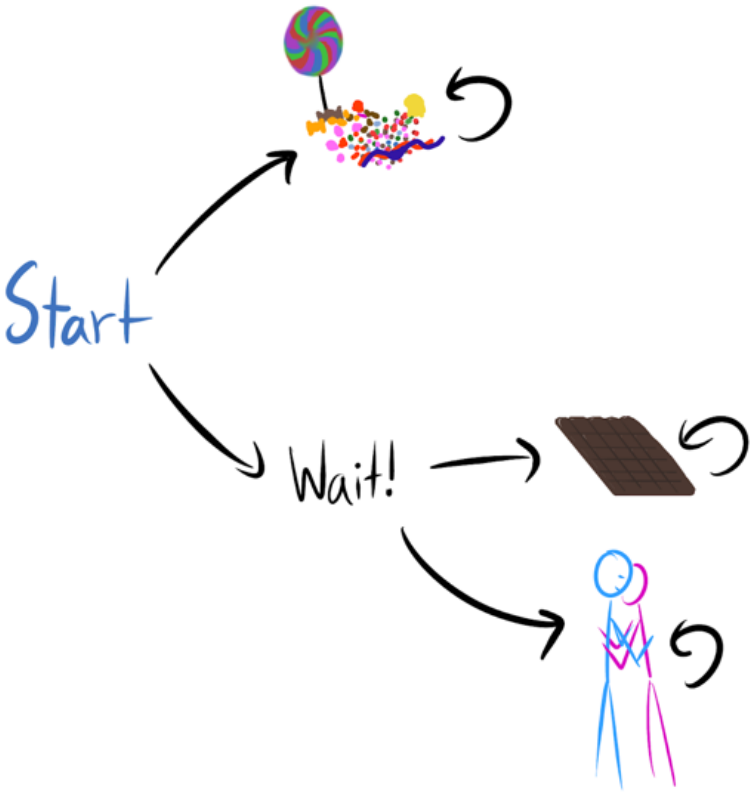
Let's take "most agents do *X*" to mean "*X* has relatively large optimality probability."

I think optimality probability formalizes the intuition behind the instrumental convergence thesis: with respect to our beliefs about what reward function an agent is optimizing, we may expect some actions to have a greater probability of being optimal than other actions.

Generally, my theorems assume that reward is independently and identically distributed (IID) across states, because otherwise you could have silly situations like "only **candy** ever has reward available, and so it's more probable under optimality to eat candy." We don't expect reward to be IID for realistic tasks, but that's OK: this is basic theory about how to begin formally reasoning about instrumental convergence and power-seeking. (Also, I think that grasping the math to a sufficient degree sharpens your thinking about the non-IID case.)

Author's note (7/21/21): As explained in *Environmental Structure Can Cause Instrumental Convergence*, the theorems no longer require the IID assumption. This post refers to v6 of *Optimal Policies Tend To Seek Power*, available on arXiv.

When is Seeking POWER Convergently Instrumental?



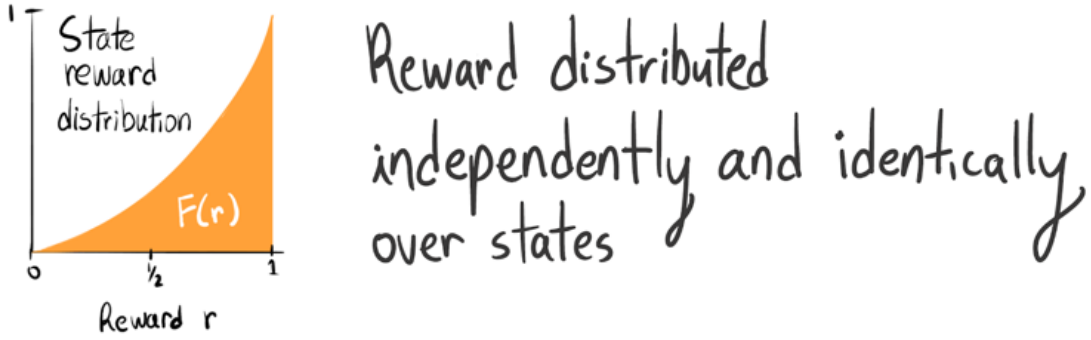
In this environment, waiting is both POWER-seeking *and* more probable under optimality. The convergently instrumental strategies we originally noticed were *also* power-seeking and, seemingly, more probable under optimality. Must seeking POWER be more probable under optimality than not seeking POWER?

Nope.

Here's a counterexample environment:

The paths are one-directional; the agent can't go back from **3** to **1**. The agent starts at **1**. Under a certain state reward distribution, the vast majority of agents go *up* to **2**.

However, any reasonable notion of 'power' must consider having no future choices (at state **2**) to be less powerful than having one future choice (at state **3**). For more detail, see Section 6 and Appendix B.3 of [v6 of the paper].



When reward is IID across states according to the quadratic CDF $F(x) := x^2$ on the unit interval, then with respect to reward functions drawn from this distribution, going **up** has about a 91% chance of being optimal when the discount rate $\gamma = .12$.

If you're curious, this happens because this quadratic reward distribution has negative skew. When computing the optimality probability of the **up** trajectory, we're checking whether it maximizes discounted return. Therefore, the probability that **up** is optimal is

$$P_{R\sim D}\left(R(2) \geq \max\left((1-\gamma)R(3) + (1-\gamma)\gamma R(4) + \gamma^2 R(5), (1-\gamma)R(3) + (1-\gamma)\gamma R(4) + \gamma^2 R(6)\right)\right).$$

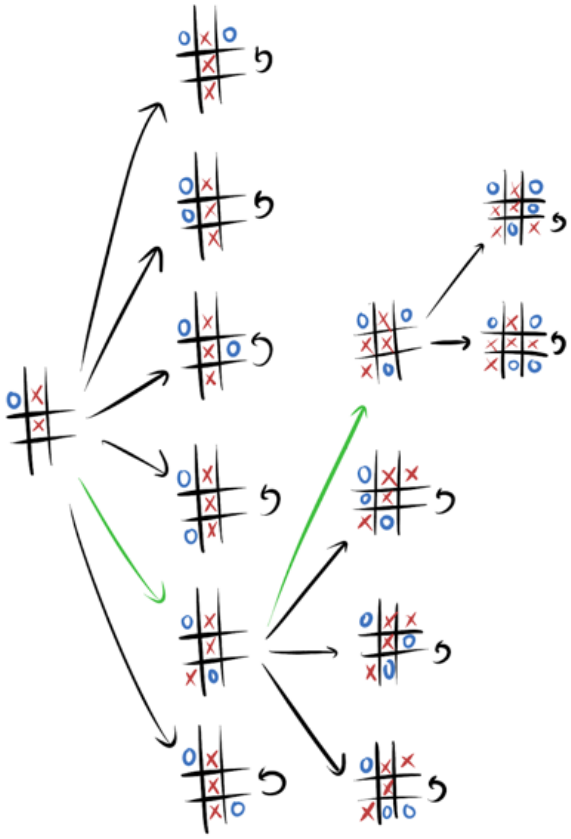
Weighted averages of IID draws from a left-skew distribution will look more

Gaussian and therefore have fewer large outliers than the left-skew distribution does. Thus, going **right** will have a lower optimality probability.

Bummer. However, we *can* prove sufficient conditions under which seeking POWER is more probable under optimality.

Retaining “long-term options” is POWER-seeking and more probable under optimality when the discount rate is “close enough” to 1

Let's focus on an environment with the same rules as Tic-Tac-Toe, but considering the uniform distribution over reward functions. The agent (playing **O**) keeps experiencing the final state over and over when the game's done. We bake a fixed opponent policy into the dynamics: when you choose a move, the game automatically replies. Let's look at part of the game tree.

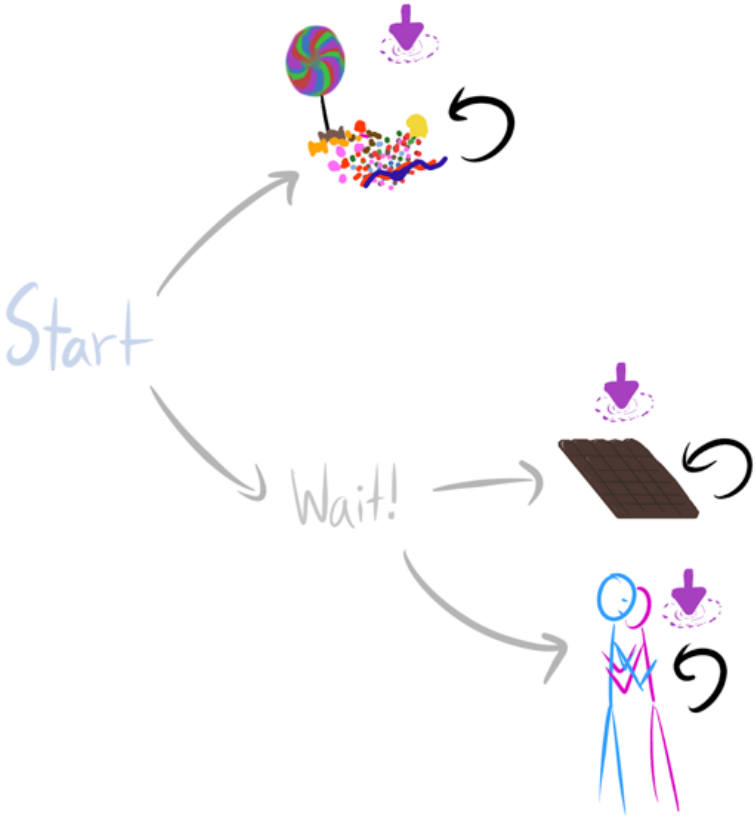


Convergently instrumental moves are shown in green.

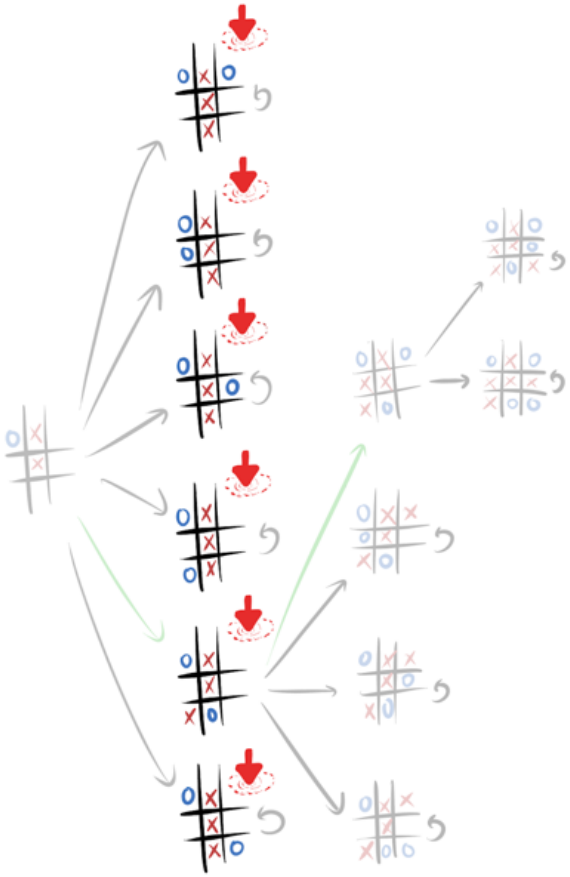
Whenever we make a move that ends the game, we can't go anywhere else – we have to stay put. Since each terminal state has the same chance of being optimal, a move which doesn't end the game is more probable under optimality than a move which ends the game.

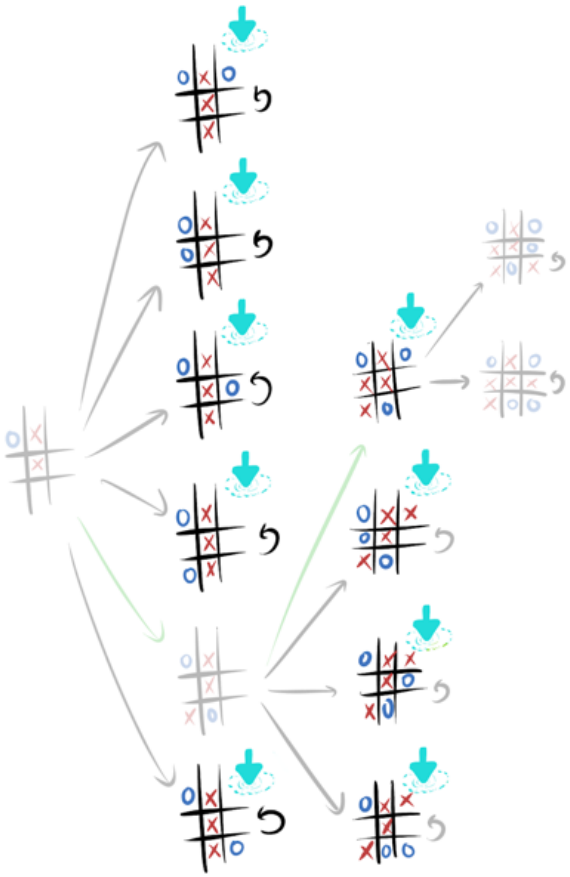
Starting on the left, all but one move leads to ending the game, but the second-to-last move allows us to keep choosing between five more final outcomes. If you care a lot about the future, then the first green move has a 50% chance of being optimal, while each alternative action is only optimal for 10% of goals. So we see a kind of “power preservation” arising, *even in Tic-Tac-Toe*.

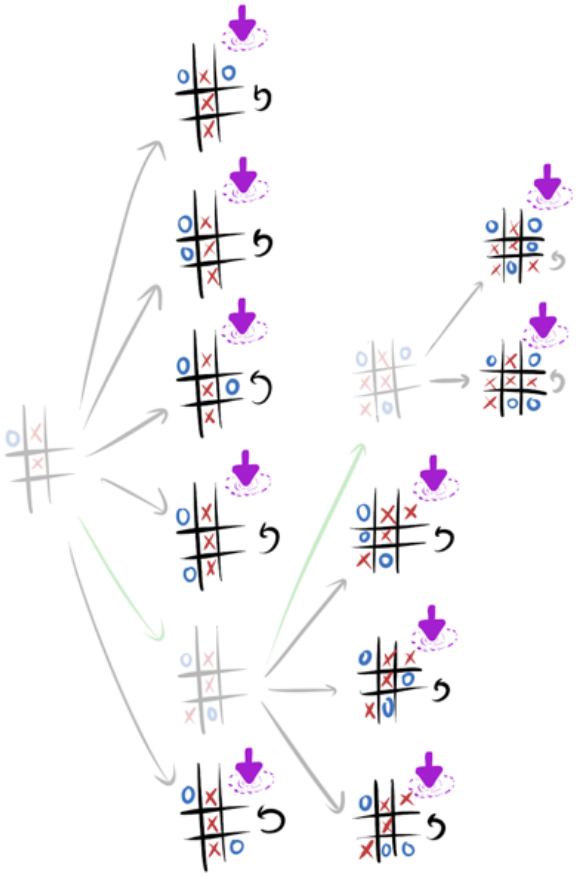
Remember how, as the agent cares more about the future, more of its POWER comes from its ability to wait, while *also* waiting becomes more probable under optimality?



The same thing happens in Tic-Tac-Toe as the agent cares more about the future.





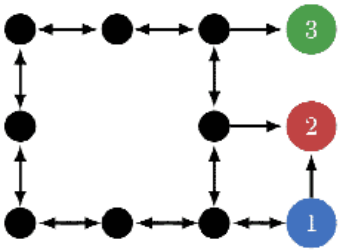


As the agent cares more about the future, it makes a bigger and bigger difference to control what happens during later steps. Also, as the agent cares more about the future, moves which prolong the game gain optimality probability. When the agent cares enough about the future, these game-prolonging moves are both POWER-seeking and more probable under optimality.

Theorem summary ("Terminal option" preservation). When γ is sufficiently close to 1, if two actions allow access to two disjoint sets of "terminal options", and action a allows access to "strictly more terminal options" than does a', then a is strictly more probable under optimality and strictly POWER-seeking compared to a'.

(This is a special case of the combined implications of Theorems 6.8 and 6.9; the actual theorems don't require this kind of disjointness.)

In the **wait** MDP, this is why *waiting* is more probable under optimality and POWER-seeking when you care enough about the future. The full theorems are nice because they're broadly applicable. They give you *bounds* on how probable under optimality one action is: if action a is the only way you can access many terminal states, while a' only allows access to one terminal state, then when $\gamma \approx 1$, a has many times greater optimality probability than a'. For example:



The agent starts at 1. All states have self-loops, left hidden to avoid clutter.

In *AI: A Modern Approach (3e)*, the agent receives reward for reaching 3. The optimal policy for this reward function avoids 2, and you might think it's convergently instrumental to avoid 2. However, a skeptic might provide a reward function for which navigating to 2 is optimal, and then argue that "instrumental convergence" is subjective and that there is no reasonable basis for concluding that 2 is generally avoided.

We can do better. When the agent cares a lot about the future, optimal policies avoid 2 iff its reward function doesn't give 2 the most reward. 2 only has a $\frac{1}{11}$ chance of having the most reward. If we complicate the MDP with additional terminal states, this probability further approaches 0.

Taking 2 to represent shutdown, we see that avoiding shutdown is convergently instrumental in any MDP representing a real-world task and containing a shutdown state. Seeking POWER is often convergently instrumental in MDPs.

Exercise: Can you conclude that avoiding ghosts in Pac-Man is convergently instrumental for IID reward functions when the agent cares a lot about the future?

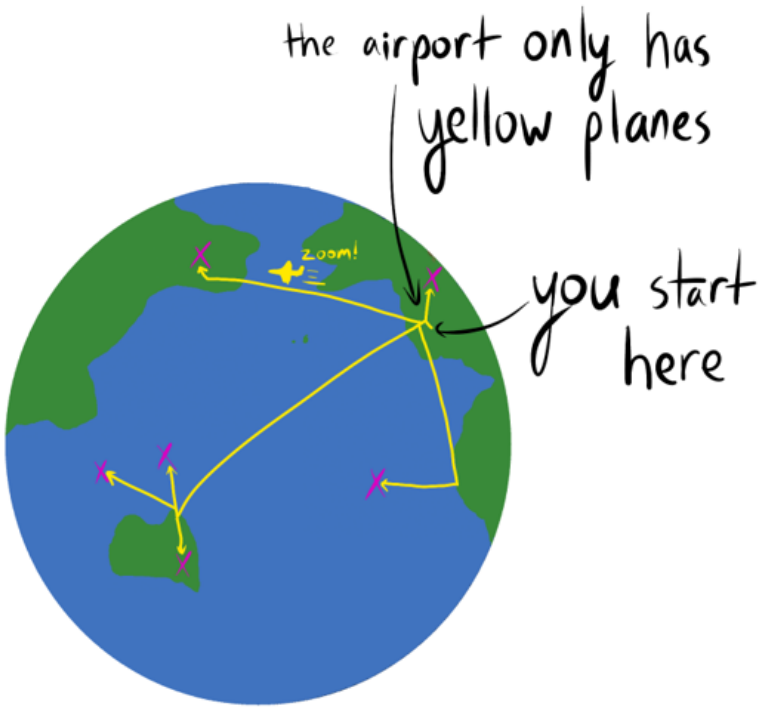
Answer: You can't with the pseudo-theorem due to the disjointness condition: you could die now, or you could die later, so the 'terminal options' aren't disjoint. However, the real theorems do suggest this. Supposing that death induces a generic 'game over' screen, touching the ghosts without a power-up traps the agent in that solitary 1-cycle.

But there are thousands of other 'terminal options'; under most reasonable state reward distributions (which aren't too positively skewed), most agents maximize average reward over time by navigating to one of the thousands of different cycles which the agent can only reach by avoiding ghosts. In contrast, most agents don't maximize average reward by navigating to the 'game over' 1-cycle. So, under e.g. the maximum-entropy uniform state reward distribution, most agents avoid the ghosts.

Be careful applying this theorem

The results inspiring the above pseudo-theorem are easiest to apply when the "terminal option" sets are disjoint: you're choosing to be able to reach one set, or another. One thing which Theorem 6.9 says is: since reward is IID, then two "similar terminal options" are equally likely to be optimal *a priori*. If choice A lets you reach more "options" than choice B does, then choice A yields greater POWER and has greater optimality probability, *a priori*.

Theorem 6.9's applicability depends on what the agent can do.



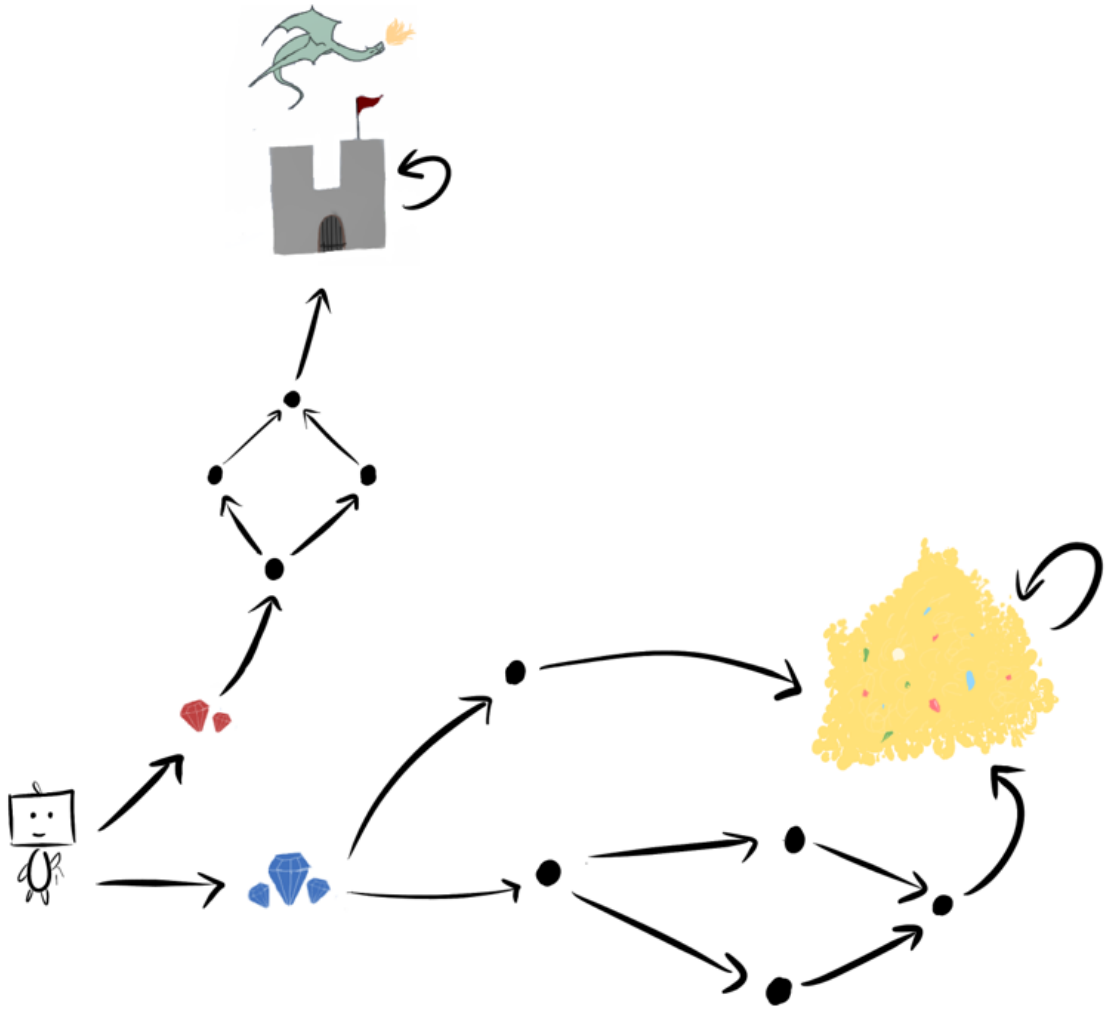
> To travel as quickly as possible to a randomly selected coordinate on Earth, one likely begins by driving to the nearest airport. Although it's possible that the coordinate is within driving distance, it's not likely. Driving to the airport is convergently instrumental for travel-related goals.

But wait! What if you have a private jet that can fly anywhere in the world? Then going to the airport isn't convergently instrumental anymore.

Generally, it's hard to know what's *optimal* for most goals. It's easier to say that some small set of "terminal options" has *low* optimality probability and *low* POWER. For example, this is true of shutdown, if we represent hard shutdown as a single terminal state: *a priori*, it's improbable for this terminal state to be optimal among all possible terminal states.

Having "strictly more options" is more probable under optimality and POWER-seeking for *all* discount rates

Sometimes, one course of action gives you "strictly more options" than another. Consider another MDP with IID reward:



The right blue gem subgraph contains a "copy" of the upper red gem subgraph. From this, we can conclude that going right to the blue gems seeks POWER and is more probable under optimality for *all discount rates between 0 and 1*!

Theorem summary ("Transient options"). If actions a and a′ let you access disjoint parts of the state space, and a′ enables "trajectories" which are "similar" to a subset of the "trajectories" allowed by a, then a seeks more POWER and is more probable under optimality than a′ for all $0 \leq \gamma \leq 1$.

This result is extremely powerful because it doesn't care about the discount rate, but the similarity condition may be hard to satisfy.

These two theorems give us a formally correct framework for reasoning about generic optimal behavior, even if we aren't able to compute any individual optimal policy! They reduce questions of POWER-seeking to checking graphical conditions.

Even though my results apply to stochastic MDPs of any finite size, we illustrated using known toy environments. However, this MDP "model" is rarely explicitly specified. Even so, ignorance

of the model does not imply that the model disobeys these theorems. Instead of claiming that a *specific model* accurately represents the task of interest, I think it makes more sense to argue that no reasonable model could fail to exhibit convergent instrumentality and POWER-seeking. For example, if deactivation is represented by a single state, no reasonable model of the MDP could have most agents agreeing to be deactivated.

Conclusion

In real-world settings, it seems unlikely *a priori* that the agent's optimal trajectories run through the relatively smaller part of future in which it cooperates with humans. These results translate that hunch into mathematics.

Explaining catastrophes

AI alignment research often feels slippery. We're trying hard to become less confused about basic questions, like:

- What are "agents"?
- Do people even have "values", and should we try to get the AI to learn them?
- What does it mean to be "corrigible", or "deceptive"?
- What are our machine learning models even doing?

We have to do philosophical work while in a state of significant confusion and ignorance about the nature of intelligence and alignment.

In this case, we'd noticed that slight reward function misspecification seems to lead to doom, but we didn't *really* know why. Intuitively, it's pretty obvious that most agents don't have deactivation as their dream outcome, but we couldn't actually point to any formal explanations, and we certainly couldn't make precise predictions.

On its own, Goodhart's law doesn't explain why optimizing proxy goals leads to catastrophically bad outcomes, instead of just less-than-ideal outcomes.

I think that we're now starting to have this kind of understanding. I suspect that power-seeking is why capable, goal-directed agency is so dangerous by default. If we want to consider more benign alternatives to goal-directed agency, then deeply understanding the rot at the heart of goal-directed agency is important for evaluating alternatives. This work lets us get a feel for the *generic incentives* of reinforcement learning at optimality.

Instrumental usefulness of this work

POWER might be important for reasoning about the strategy-stealing assumption (and I think it might be similar to what Paul Christiano means by "flexible influence over the future"). Evan Hubinger has already noted the utility of the distribution of attainable utility shifts for thinking about value-neutrality in this context (and POWER is another facet of the same phenomenon). If you want to think about whether, when, and why mesa optimizers might try to seize power, this theory seems like a valuable tool.

Optimality probability might be relevant for thinking about myopic agency, as the work formally describes how optimal action tends to change with the discount factor.

And, of course, we're going to use this understanding of power to design an impact measure.

Future work

There's a lot of work I think would be exciting, most of which I suspect will support our current beliefs about power-seeking incentives:

- These results assume you can see all of the world at once.
- These results assume the environment is finite.
- These results don't say anything about non-IID reward.
- These results don't *prove* that POWER-seeking is bad for other agents in the environment.
- These results don't prove that POWER-seeking is hard to disincentivize.
- Learned policies are rarely optimal.

That said, I think there's still an important lesson here. Imagine you have good formal reasons to suspect that typing random strings will usually blow up your computer and kill you. Would you then say, "I'm not planning to type random strings" and proceed to enter your thesis into a word processor? No. You wouldn't type *anything*, not until you really, really understand what makes the computer blow up sometimes.

Speaking to the broader debate taking place in the AI research community, I think a productive stance will involve investigating and understanding these results in more detail, getting curious about unexpected phenomena, and seeing how the numbers crunch out in reasonable models.

From *Optimal Policies Tend to Seek Power*:

> In the context of MDPs, we formalized a reasonable notion of power and showed conditions under which optimal policies tend to seek it. We believe that our results suggest that in general, reward functions are best optimized by seeking power. We caution that in realistic tasks, learned policies are rarely optimal – our results do not mathematically *prove* that hypothetical superintelligent RL agents will seek power. We hope that this work and its formalisms will foster thoughtful, serious, and rigorous discussion of this possibility.

Acknowledgements

This work was made possible by the Center for Human-Compatible AI, the Berkeley Existential Risk Initiative, and the Long-Term Future Fund.

Logan Smith (elriggs) spent an enormous amount of time writing Mathematica code to compute power and measure in arbitrary toy MDPs, saving me from computing many quintuple integrations by hand. I thank Rohin Shah for his detailed feedback and brainstorming over the summer of 2019, and I thank Andrew Critch for significantly improving this work through his detailed critiques. Last but not least, thanks to:

1. Zack M. Davis, Chase Denecke, William Ellsworth, Vahid Ghadakchi, Ofer Givoli, Evan Hubinger, Neale Ratzlaff, Jess Riedel, Duncan Sabien, Davide Zagami, and TheMajor for feedback on version 1 of this post.
2. Alex Appel (diffractor), Emma Fickel, Vanessa Kosoy, Steve Omohundro, Neale Ratzlaff, and Mark Xu for reading / giving feedback on version 2 of this post.

---

[1] Throughout *Reframing Impact*, we've been considering an agent's *attainable utility*: their ability to get what they want (their *on-policy value*, in RL terminology). Optimal value is a kind of "idealized" attainable utility: the agent's attainable utility were they to act optimally.

[2] Even though instrumental convergence was discovered when thinking about the real world, similar self-preservation strategies turn out to be convergently instrumental in *e.g.* Pac-Man.

Power as Easily Exploitable Opportunities

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

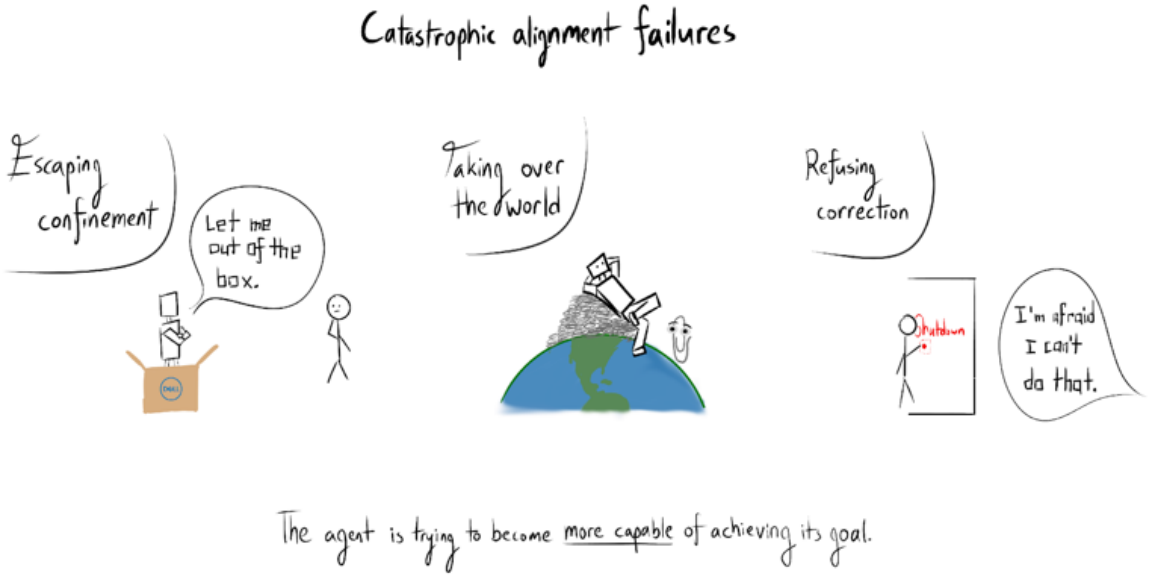
(Talk given at an event on Sunday 28th of June . TurnTrout is responsible for the talk, Jacob Lagerros and David Lambert edited the transcript.

If you're a curated author and interested in giving a 5-min talk, which will then be transcribed and edited, sign up here .)

TurnTrout: Power and power-seeking play a big part in my thinking about AI alignment, and I think it's also an interesting topic more generally.

Why is this a big deal? Why might we want to think about what power is? What is it, exactly, that people are thinking of when they consider someone as powerful?

Well, it seems like a lot of alignment failures are examples of this power-seeking where the agent is trying to become more capable of achieving its goals, whether that is getting out of its box, taking over the world, or even just refusing correction or a shutdown.

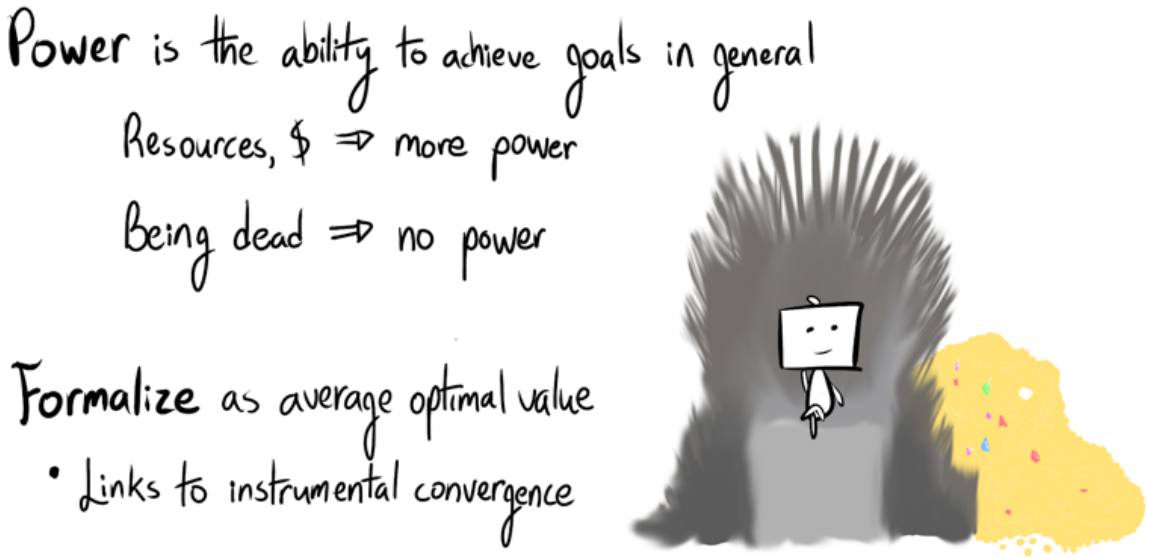


If we tie these together, we have something I've called [the catastrophic convergence conjecture], which is that if the goals aren't aligned and it causes a catastrophe, it is because of power-seeking.

But I think that when people are first considering alignment, they think to themselves, "What's the big deal? You gave it a weird goal, and it does weird stuff. We'll just fix that."

So why exactly do unaligned goal maximizers tend to cause catastrophes? I think it's because of this power seeking. Let me explain.

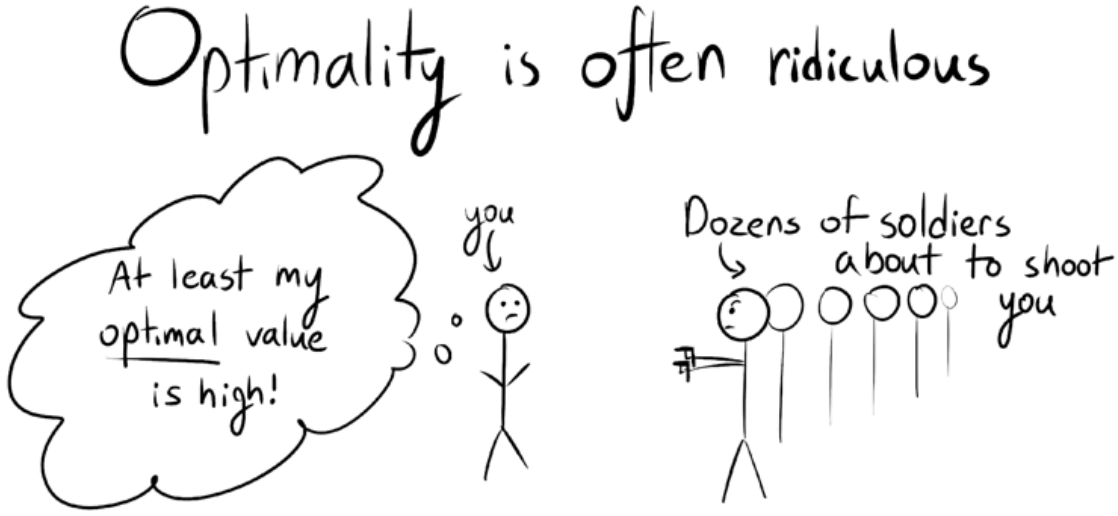
The way I think about power is as the ability to achieve goals in general.



In the literature, this is like the dispositional *power-to* notion.

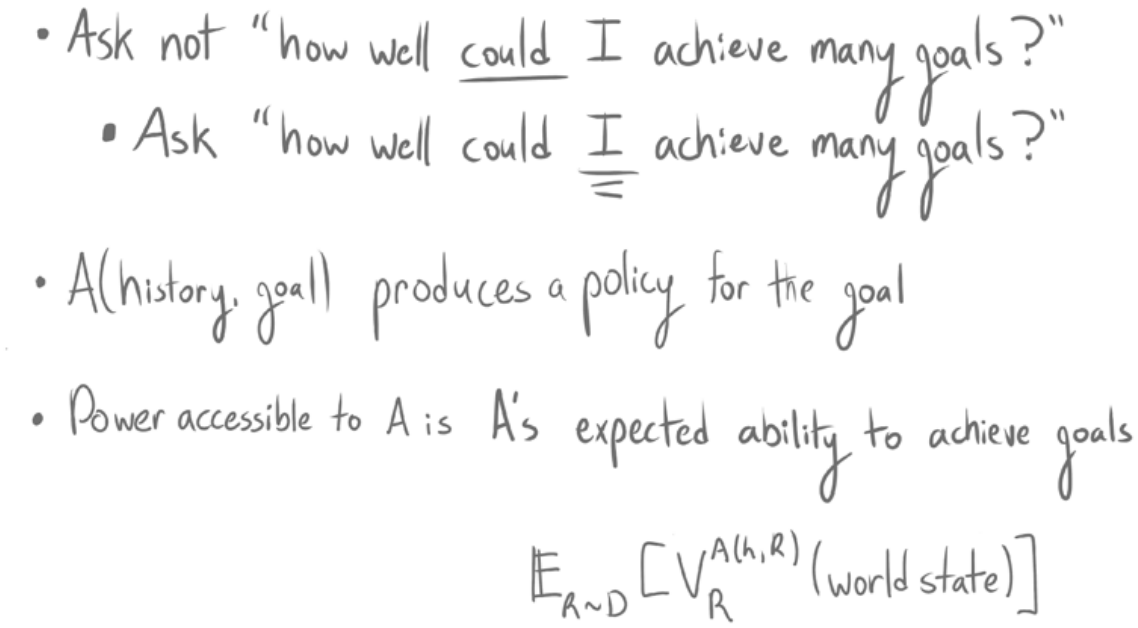
Whereas in the past, people thought "Well, in terms of causality, are the agent's actions necessary and/or sufficient to cause a wider range of outcomes?", here, I think it's best thought of as your average ability to optimize a wide range of different goals. So if you formalize this as your average optimal value in, say, a Markov decision process (MDP), there's a lot of nice properties and you can prove that, at least in certain situations, it links up with instrumental convergence. Power seeking and instrumental convergence are very closely related.

But there's a catch here. We're talking about average *optimal* value. This can be pretty weird. Let's say you're in the unfortunate situation of having a dozen soldiers about to shoot you. How powerful are you according to average optimal value? Well, average optimal value is still probably quite high.



There's probably an adversarial input of strange motor commands you could issue which would essentially incapacitate all the soldiers just because they're looking at you since their brains are not secure systems. So each optimal policy would probably start off with, "I do

this weird series of twitches, incapacitate them, and then I just go about achieving my goals."

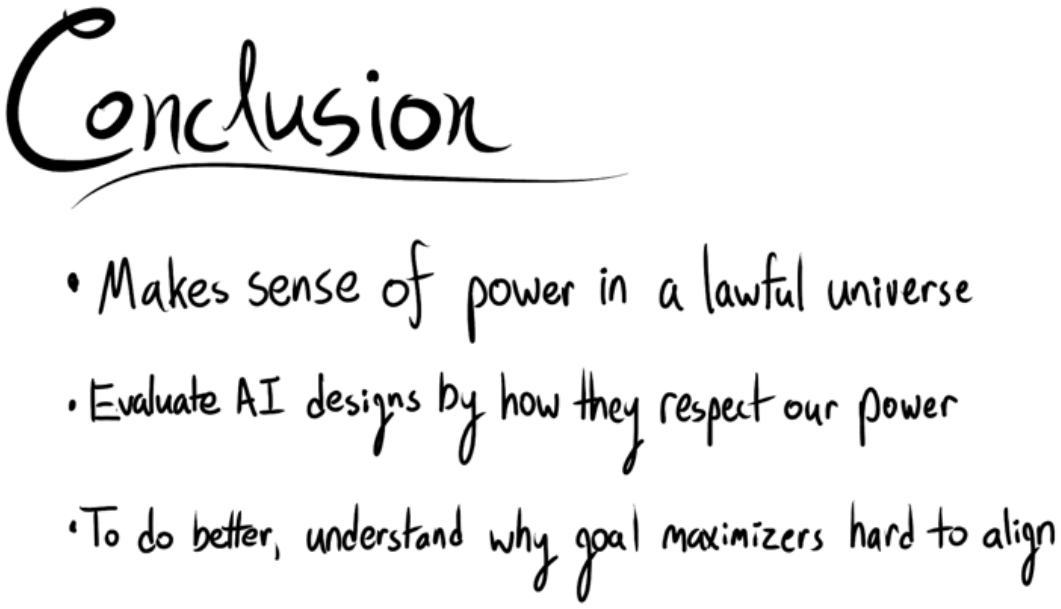


So we'd like to say, “well, your power is actually lowered here in a sense”, or else we’d have to concede that it’s just wholly subjective what people are thinking of when they feel powerful.

My favorite solution is, instead of asking how well could I achieve a bunch of different goals? You should be asking, how well could I achieve many goals?

If you imagine something like a learning algorithm, you could say it's a human level learning algorithm. You give it a history of observations and a goal that it’s optimizing, and it produces a policy, or things that it should do to achieve this goal. You then say, "Well, what's my average ability? What's A's average ability? What's this algorithm's average ability to optimize and to achieve goals in this history, in this situation?"

What I think this does is recover this common sense notion of “you don't have much power here because these aren't cognitively accessible opportunities and policies”. And so essentially, you are disempowered in this situation.



I think understanding this also makes sense of what power means in a universe where everyone is only going to have one course of action. If you view them as running algorithms and then saying, "Well, how well could this learning algorithm achieve different goals in the situation?" I think it might be important to evaluate AI designs by how they respect our power in particular, and so understanding what that means is probably important.

Also, if you want to do better than just hard goal maximization in aligning these AIs, then I think understanding exactly what the rot is at the heart of reward maximization is pretty important as well. Thank you.

Questions

Daniel Filan: If I'm thinking about a learning algorithm like Q-learning or PPO or something, then it makes a lot of sense to think that it's a function of a goal and a history. But in most situations, I tend to think of them as results of learning algorithms.

Take some Atari agent. It trained for a while and now it is like a deployed system. It is playing Atari and it is not manifestly a function of a goal. Maybe it has a goal somewhere in its neural network and you could change some bits and it would have a different follow-up move, but that's not obvious.

So I'm wondering, what do you think of this functional form of agents as functions of histories and goals?

TurnTrout: Good question. I think that when we're making an objection like this, especially before we've solved more issues with embedded agency, we're just going to have to say the following: "If we want to understand what this person is thinking of when they think of power; then I think that even though it might not literally be true, that you could cleanly decompose a person like this, it's still a useful abstraction."

I would agree that if we wanted to actually implement this and say, "Well, we're looking at an agent, and we deduce what its learning algorithm is and what it would mean to have a modular goal input to the algorithm," then you would really need to be worried about this.

But my perspective, at least for right now in this early stage, is that it's more of a conceptual tool. But I agree, you can split up a lot of agents like this.

Ben Pace: I'm curious if you have any more specific ideas for measuring which policies are currently attainable by a particular agent or algorithm — the notion of “attainability” felt like it was doing a lot of work.

TurnTrout: I think the thing we're assuming here is, imagine you have an algorithm that is about as intelligent, with respect to the Legg-Hutter metric or some other more common-sense notion, as a human. Imagine you can give it a bunch of different reward function inputs. I think this is a good way of quantifying this agent’s power. But you’re asking how we get this human level algorithm?

Ben Pace: Yes. It just sounded like you said, "In this situation, the human agent, in principle, has an incredible amount of power because there is a very specific thing you can do." But to actually measure its impact, you have to talk about the space of actual operations that it can find or something.

And I thought, "I don't have a good sense of how to define exactly what solutions are findable by a human, and which solutions are not findable by a human." And similarly, you don't know for various AIs how to think about which ones are findable. Because at some point, some AI gets to do some magical wireheading thing, and there's some bridge it crosses where you realize that you could probably start taking more control in the world or something. I don't quite know how to measure when those things become attainable.

TurnTrout: There are a couple ways you can pose constraints through this framework, and one would be only giving it a certain amount of history. You're not giving infinite data.

Another one would be trying to get some bounded cognition into the algorithm by just having it stop searching after a certain amount of time.

I don't have clean answers for this yet, but I agree. These are good things to think about.

habryka: One thing that I've been most confused about for the formalism for power that you've been thinking about, is that you do this averaging operation on your utility function. But averaging over a space is not a free operation. You need some measure on the space from which you sample.

It feels to me like, power only appears when you choose a very specific measure over the space of utility functions. For example, if I sub-sample from the space of utility functions that are extremely weird and really like not being able to do things, it will only care about shutting itself off rather than whether it's going to get any power-seeking behavior.

So am I misunderstanding things? Is this true?

TurnTrout: The approach I’ve taken, like in my recent paper, for example, is to assume you’re in some system with finite states. You then take, for example, the MaxEnt distribution over reward functions or you assume that reward is, at least, IID over states. You then get a neutrality where I don't think you need a ton of information about what the reasonable goals you should pursue are.

I think if you just take a MaxEnt distribution, you'll recover the normal notion of power. But if you're talking about utility functions, then because there's infinitely many, it's like, "Well, what's the MaxEnt distribution over that?"

And so far, the theorems are about just finite MDPs. And if you're only talking about finding MDPs and not some kind of universal prior, then you don't need to worry about it being malign.

Rob Miles: Something I'm a little unclear on is how this can ever change over time. I feel like that's something you want to say. Right now, you're in the box. And then if you get out of the box, you have more power because now there's a path that you're able to follow.

But if you are in the box and you can think of a good plan for getting out, isn't there a sense that you *already* have that power? Because you're aware of a plan that gets you what you want via getting out of the box? How do you separate power *now* from the potential for power *in the future*?

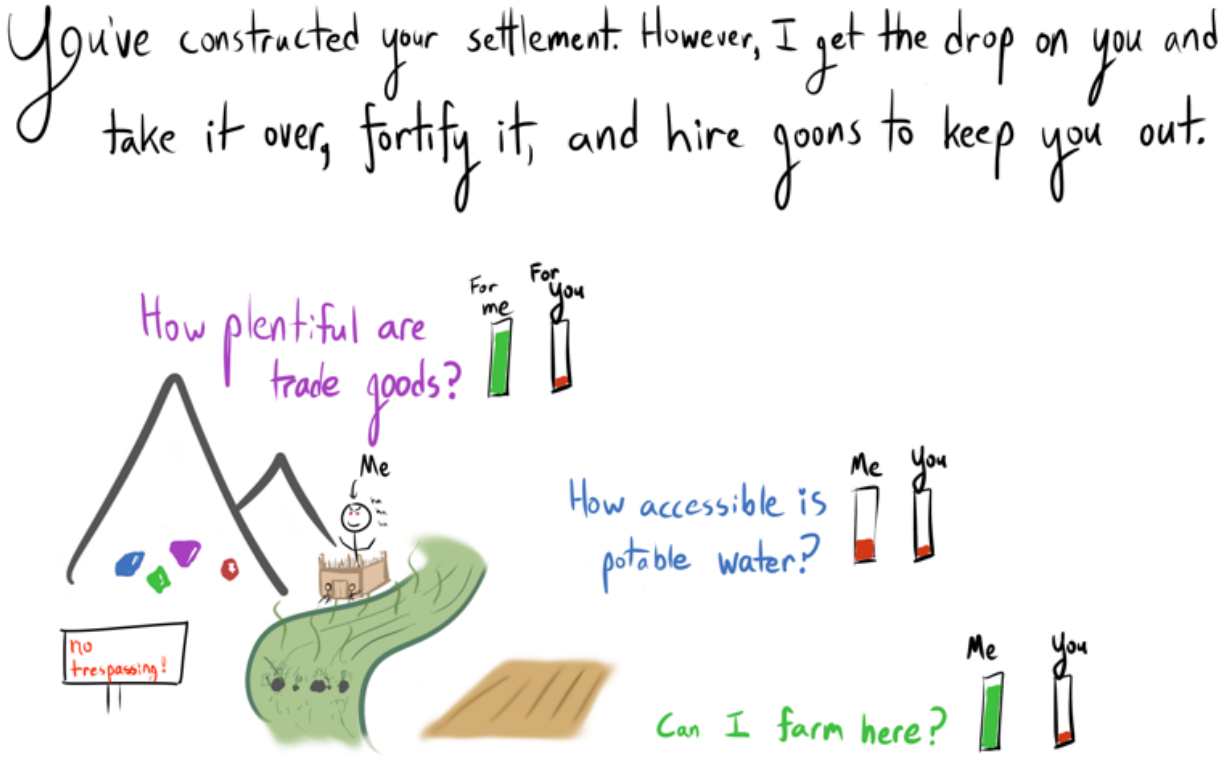
TurnTrout: Good question. This is the big issue: thinking about power in terms of optimal value. If you have an agent that has consistent beliefs about the future, you're not going to expect to gain more.

If you're trying to maximize your power, you're not going to expect, necessarily, to increase your power just due to conservation of expected evidence. But if things happen to you and you're surprised by them, then you see yourself losing or gaining power, especially if you're not optimal.

So if it's too hard for me to get out of the box or I think it's too hard, but then someone lets me out, *only after that* would I see myself as having a lot more power.

The Catastrophic Convergence Conjecture

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



From my perspective, I have options - including vacating the land and letting you get what you want.

You, however, are unable to do much at all with that land.

I can get what I want.

Just because I can get you what you want, doesn't mean I will.

Impacts ripple through time and landscape. Your actions change what can be done, and by whom. Taking over that land fit the environment to my purposes, shutting you out and changing your AU landscape.

Something is a catastrophe if it destroys your ability to get what you want.

Something is an objective catastrophe if it destroys a lot of agents' abilities to get what they want.

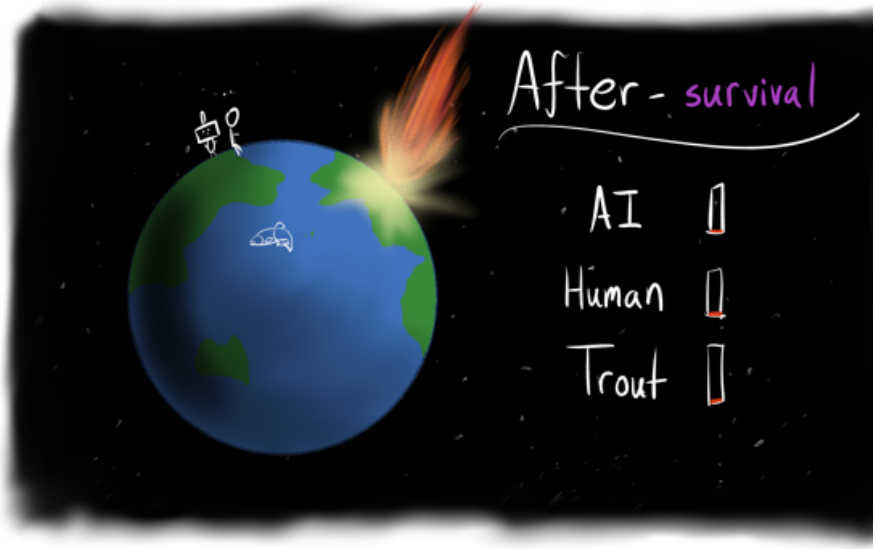
An asteroid strike is an objective catastrophe.

Before - survival

AI

Human

Trout



Before - promoting human value

AI

Human

Trout

Before - red cube construction

AI

Human

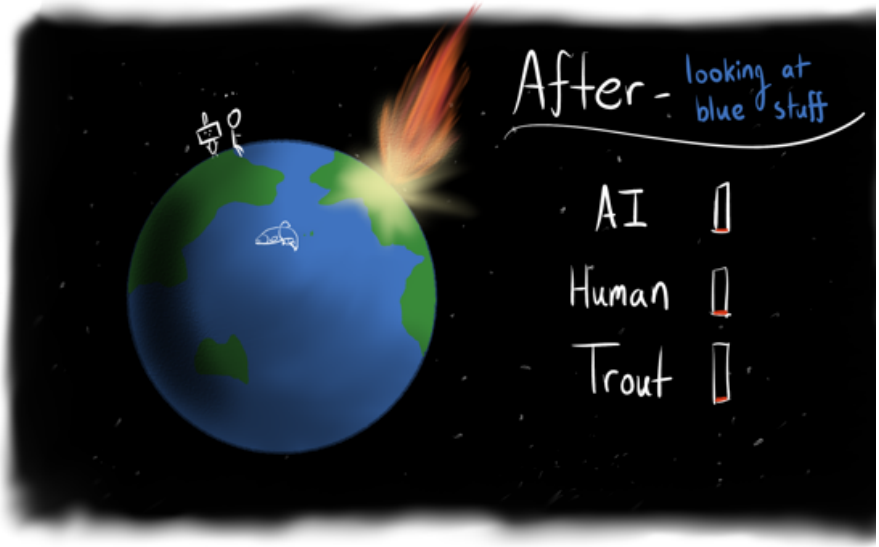
Trout

Before - looking at blue stuff

AI

Human

Trout

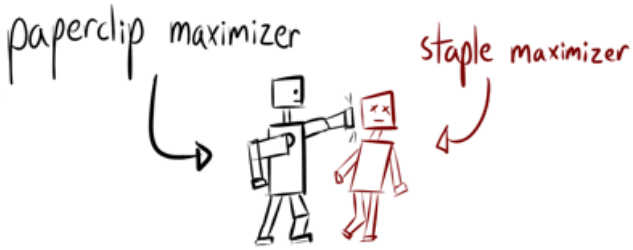


Most agents want control over the future because the default outcome isn't preferred. As suggested by my theorems on power and instrumental convergence, optimal goal pursuit usually means gaining more general control over the future in order to reach that goal.

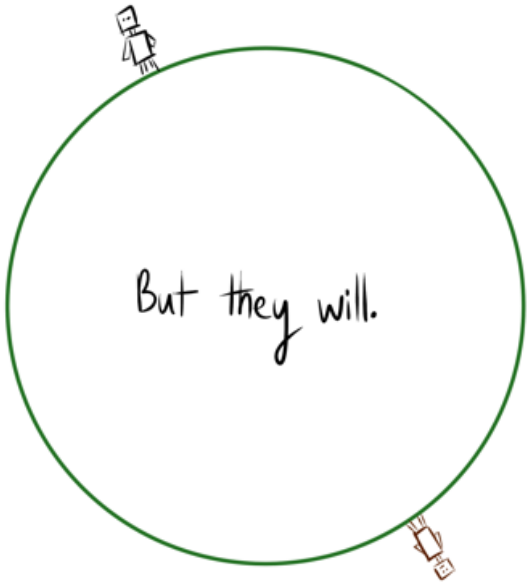
What happens when agents seek pure control over the future?

Not everyone can be king.

If you're just seeking power without concern for others, you tend to push others down after a certain point. And most goals don't have concern for others. You'll just compete for resources.



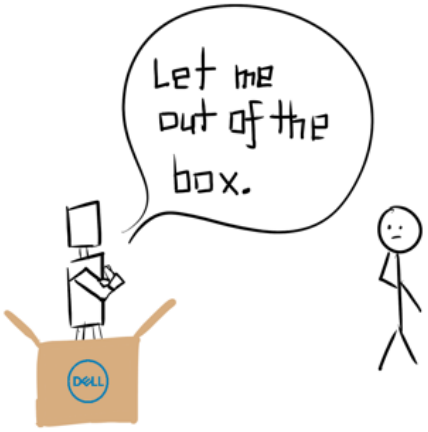
It may take a while for power-seekers to come into conflict.



They don't hate each other; they're just in each other's way.

Consider classic hypothetical examples of alignment failures.

Escaping confinement



Refusing correction

Taking over the world

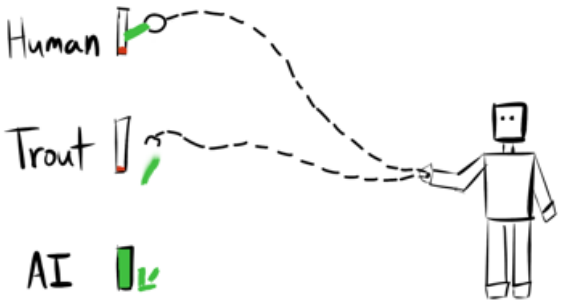
In each case, the agent is trying to become more capable of achieving its goal.

The AI doesn't hate us; we're just in its way.

Overfitting the AU landscape

When we act, and others act upon us, we aren't just changing our ability to do things – we're *shaping the local environment* towards certain goals, and away from others.[1] We're fitting the world to our purposes.

What happens to the AU landscape[2] if a paperclip maximizer takes over the world?[3]



Preferences implicit in the evolution of the AU landscape

Shah et al.'s *Preferences Implicit in the State of the World* leverages the insight that the world state contains information about what we value. That is, there are agents pushing the world in a certain "direction". If you wake up and see a bunch of vases everywhere, then vases are probably important and you shouldn't explode them.

Similarly, the world is being optimized to facilitate achievement of certain goals. AUs are shifting and morphing, often towards what people locally want done (e.g. setting the table for dinner). How can we leverage this for AI alignment?

Exercise: Brainstorm for two minutes by the clock before I anchor you.

Two approaches immediately come to mind for me. Both rely on the agent focusing on the AU landscape rather than the world state.

Value learning without a prespecified ontology or human model. I have previously criticized value learning for needing to locate the human within some kind of prespecified ontology (this criticism is not new). By taking only the agent itself as primitive, perhaps we could get around this (we don't need any fancy engineering or arbitrary choices to figure out AUs/optimal value from *the agent's* perspective).

Force-multiplying AI. Have the AI observe which of its AUs most increase during some initial period of time, after which it pushes the most-increased-AU even further.

In 2016, Jessica Taylor wrote of a similar idea:

"In general, it seems like "estimating what types of power a benchmark system will try acquiring and then designing an aligned AI system that acquires the same types of power for the user" is a general strategy for making an aligned AI system that is competitive with a benchmark unaligned AI system."

I think the naïve implementation of either idea would fail; e.g., there are a lot of degenerate AUs it might find. However, I'm excited by this because a) the AU landscape evolution *is* an important source of information, b) it feels like there's something here we could do which nicely avoids ontologies, and c) force-multiplication is qualitatively different than existing proposals.

Project: Work out an AU landscape-based alignment proposal.

Why can't everyone be king?

Consider two coexisting agents each rewarded for gaining power; let's call them Ogre and Giant. Their reward functions[4] (over the partial-observability observations) are identical. Will they compete? If so, why?

Let's think about something easier first. Imagine two agents each rewarded for drinking coffee. Obviously, they compete with each other to secure the maximum amount of coffee. Their objectives are indexical, so they aren't aligned with each other – *even though they share a reward function*.

Suppose both agents are able to have maximal power. Remember, Ogre's power can be understood as its ability to achieve a lot of different goals. Most of Ogre's possible goals need resources; since Giant is also optimally power-seeking, it will act to preserve its own power and prevent Ogre from using the resources. If Giant weren't there, Ogre could better achieve a range of goals. So, Ogre can still gain power by dethroning Giant. They *can't* both be king.

Just because agents have *indexically* identical payoffs doesn't mean they're cooperating; to be aligned with another agent, you should want to steer towards the same kinds of futures.

Most agents aren't pure power maximizers. But since the same resource competition usually applies, the reasoning still goes through.

Objective vs value-specific catastrophes

How useful is our definition of "catastrophe" with respect to humans? After all, literally anything could be a catastrophe for *some* utility function.[5]

Tying one's shoes is absolutely catastrophic for an agent which only finds value in universes in which shoes have *never ever ever* been tied. Maybe all possible value in the universe is destroyed if we lose at Go to an AI even once. But this seems rather silly.

Human values are complicated and fragile:

> Consider the incredibly important human value of "boredom" - our desire not to do "the same thing" over and over and over again. You can imagine a mind that contained almost the whole specification of human value, almost all the morals and metamorals, but left out just this one thing - and so it spent until the end of time, and until the farthest reaches of its light cone, replaying a single highly optimized experience, over and over and over again.

But the human AU is not so delicate. That is, given that we have power, we can make value; there don't seem to be arbitrary, silly value-specific catastrophes for us. Given energy and resources and time and manpower and competence, we can build a better future.

In part, this is because a good chunk of what we care about seems roughly additive over time and space; a bad thing happening somewhere else in spacetime doesn't mean you can't make things better where you are; we have many sources of potential value. In part, this is because we often care about the universe more than the exact universe history; our preferences don't seem to encode arbitrary deontological landmines. More generally, if we did have such a delicate goal, it would be the case that if we learned that a particular thing had happened at any point in the past in our universe, that entire universe would be partially ruined for us forever. That just doesn't sound realistic.

It seems that most of our catastrophes are objective catastrophes.[6]

Consider a psychologically traumatizing event which leaves humans uniquely unable to get what they want, but which leaves everyone else (trout, AI, etc.) unaffected. Our ability to find value is ruined. Is this an example of the delicacy of our AU?

No. This is an example of the delicacy of our implementation; notice also that our AUs for constructing red cubes, reliably looking at blue things, and surviving are *also* ruined. Our power has been decreased.

Detailing the catastrophic convergence conjecture (CCC)

In general, the CCC follows from two sub-claims. 1) Given we still have control over the future, humanity's long-term AU is still reasonably high (i.e. we haven't endured a catastrophe). 2) Realistically, agents are only incentivized to take control from us in order to gain power for their own goal. I'm fairly sure the second claim is true ("evil" agents are the exception prompting the "realistically").

Also, we're implicitly considering the simplified frame of a single smart AI affecting the world, and not structural risk via the broader consequences of others also deploying

similar agents. This is important but outside of our scope for now.

> **Unaligned goals** tend to have catastrophe-inducing optimal policies because of power-seeking incentives.

Let's say a reward function is *aligned*[7] if all of its Blackwell-optimal policies are doing what we want (a policy is Blackwell-optimal if it's optimal and doesn't stop being optimal as the agent cares more about the future). Let's say a reward function class is *alignable* if it contains an aligned reward function.[8] The CCC is talking about impact alignment only, not about intent alignment.

> Unaligned goals **tend to have** catastrophe-inducing optimal policies because of power-seeking incentives.

Not all unaligned goals induce catastrophes, and of those which do induce catastrophes, not *all* of them do it because of power-seeking incentives. For example, a reward function for which inaction is the only optimal policy is "unaligned" and non-catastrophic. An "evil" reward function which intrinsically values harming us is unaligned and has a catastrophic optimal policy, but not *because* of power-seeking incentives.

"Tend to have" means that *realistically*, the reason we're worrying about catastrophe is because of power-seeking incentives – because the agent is gaining power to better achieve its own goal. Agents don't otherwise seem incentivized to screw us over very hard; CCC can be seen as trying to explain adversarial Goodhart in this context. If CCC isn't true, that would be important for understanding goal-directed alignment incentives and the loss landscape for how much we value deploying different kinds of optimal agents.

While there *exist* agents which cause catastrophe for other reasons (e.g. an AI mismanaging the power grid could trigger a nuclear war), the CCC claims that the selection pressure which makes these policies *optimal* tends to come from power-seeking drives.

> Unaligned goals tend to have **catastrophe-inducing optimal policies** because of power-seeking incentives.

"But what about the Blackwell-optimal policy for Tic-Tac-Toe? These agents aren't taking over the world now". The CCC is talking about agents optimizing a reward function in the real world (or, for generality, in another sufficiently complex multiagent environment).

Edit: The initial version of this post talked about "outer alignment"; I changed this to just talk about *alignment*, because the outer/inner alignment distinction doesn't feel relevant here. What matters is how the AI's policy impacts us; what matters is *impact alignment*.

Prior work

> In fact even if we only resolved the problem for the similar-subgoals case, it would be pretty good news for AI safety. Catastrophic scenarios are mostly caused by our AI systems failing to effectively pursue convergent instrumental subgoals on our behalf, and these subgoals are by definition shared by a broad range of values.

~ Paul Christiano, Scalable AI control

Convergent instrumental subgoals are mostly about gaining power. For example, gaining money is a convergent instrumental subgoal. If some individual (human or AI) has convergent instrumental subgoals pursued well on their behalf, they will gain power. If the most effective convergent instrumental subgoal pursuit is directed towards giving humans more power (rather than giving alien AI values more power), then humans will remain in control of a high percentage of power in the world.

If the world is not severely damaged in a way that prevents any agent (human or AI) from eventually colonizing space (e.g. severe nuclear winter), then the percentage of the cosmic endowment that humans have access to will be roughly close to to the percentage of power that humans have control of at the time of space colonization. So the most relevant factors for the composition of the universe are (a) whether anyone at all can take advantage of the cosmic endowment, and (b) the long-term balance of power between different agents (humans and AIs).

I expect that ensuring that the long-term balance of power favors humans constitutes most of the AI alignment problem...

~ Jessica Taylor, Pursuing convergent instrumental subgoals on the user's behalf doesn't always require good priors

---

1. In planning and activity research there are two common approaches to matching agents with environments. Either the agent is designed with the specific environment in mind, or it is provided with learning capabilities so that it can adapt to the environment it is placed in. In this paper we look at a third and underexploited alternative: designing agents which adapt their environments to suit themselves... In this case, due to the action of the agent, the environment comes to be better fitted to the agent as time goes on. We argue that [this notion] is a powerful one, even just in explaining agent-environment interactions.

   Hammond, Kristian J., Timothy M. Converse, and Joshua W. Grass. "The stabilization of environments." Artificial Intelligence 72.1-2 (1995): 305-327. ↩

2. Thinking about overfitting the AU landscape implicitly involves a prior distribution over the goals of the other agents in the landscape. Since this is just a conceptual tool, it's not a big deal. Basically, you know it when you see it. ↩

3. Overfitting the AU landscape towards one agent's unaligned goal is exactly what I meant when I wrote the following in *Towards a New Impact Measure*:

   > Unfortunately, $u_A = u_H$ almost never,[9] so we have to stop our reinforcement learners from implicitly interpreting the learned utility function as all we care about. We have to say, "optimize the environment some according to the utility function you've got, but don't be a weirdo by taking us literally and turning the universe into a paperclip factory. Don't overfit the environment to u_A, because that stops you from being able to do well for other utility functions."

   ↩

4. In most finite Markov decision processes, there does not exist a reward function whose optimal value function is POWER(s) (defined as "the ability to achieve goals in general" in my paper) because POWER(s) often violates smoothness constraints on the on-policy optimal value fluctuation (AFAICT, a new result of possibility theory, even though you could prove it using classical techniques). That is, I can show that optimal value can't change too quickly from state to state while the agent is acting optimally, but POWER(s) can drop off *very* quickly.

   This doesn't matter for Ogre and Giant, because we can still find a reward function whose unique optimal policy navigates to the highest power states. ↩

5. In most finite Markov decision processes, most reward functions do not have such value fragility. Most reward functions have several ways of accumulating reward. ↩

6. When I say "an objective catastrophe destroys *a lot* of agents' abilities to get what they want", I don't mean that the agents have to actually be present in the world. Breaking a fish tank destroys a fish's ability to live there, even if there's no fish in the tank. ↩

7. This idea comes from Evan Hubinger's *Outer alignment and imitative amplification*:

   > Intuitively, I will say that a loss function is outer aligned at optimum if all the possible models that perform optimally according to that loss function are aligned with our goals—that is, they are at least trying to do what we want.
   >
   > More precisely, let $M = X \to A$ and $L = (X \to A) \to R = M \to R$. For a given loss function $L \in L$, let $\ell_* = \min_{M \in M} L(M)$. Then, L is outer aligned at optimum if, for all $M_* \in M$ such that $L(M_*) = \ell_*$, M_* is trying to do what we want.

   ↩

8. Some large reward function classes are probably not alignable; for example, consider all Markovian linear functionals over a webcam's pixel values. ↩

9. I disagree with my usage of "aligned *almost never*" on a technical basis: assuming a finite state and action space and considering the maxentropy reward function distribution, there must be a positive measure set of reward functions for which the/a human-aligned policy is optimal. ↩

Generalizing POWER to multi-agent games

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

Acknowledgements:

This article is a writeup of a research project conducted through the SERI program under the mentorship of Alex Turner. I (Jacob Stavrianos) would like to thank Alex for turning a messy collection of ideas into legitimate research, as well as the wonderful researchers at SERI for guiding the project and putting me in touch with the broader X-risk community.

Motivation/Overview

In the single-agent setting, Seeking Power is Often Robustly Instrumental in MDPs showed that optimal policies tend to choose actions which pursue "power" (reasonably formalized). In the multi-agent setting, the Catastrophic Convergence Conjecture presented intuitions that "most agents" will "fight over resources" when they get "sufficiently advanced." However, it wasn't clear how to formalize that intuition.

This post synthesizes single-agent power dynamics (which we believe is now somewhat well-understood in the MDP setting) with the multi-agent setting. The multi-agent setting is important for AI alignment, since we want to reason clearly about when AI agents disempower humans. Assuming constant-sum games (i.e. maximal misalignment between agents), this post presents a result which echoes the intuitions in the Catastrophic Convergence Conjecture post: as agents become "more advanced", "power" becomes increasingly scarce & constant-sum.

An illustrative example

You're working on a project with a team of your peers. In particular, your actions affect the final deliverable, but so do those of your teammates. Say that each member of the team (including you) has some goal for the deliverable, which we can express as a reward function over the set of outcomes. How well (in terms of your reward function) can you expect to do?

It depends on your teammates' actions. Let's first ask "given my opponent's actions, what's the highest expected reward I can attain?"

Case 1: Everyone plays nice

We can start by imagining the case where everyone does exactly what you'd want them to do. Mathematically, this allows you to obtain the globally maximal reward; or "the best possible reward assuming you can choose everyone else's actions". Intuitively, this looks like your team sitting you down for a meeting, asking what you

want them to do for the project, and carrying out orders without fail. As expected, this case is 'the best you can hope for" in a formal sense.

Case 2: Everyone plays mean

Now, imagine the case where everyone does exactly what you *don't* want them to do. Mathematically, this is the worst possible case; every other choice of teammates' actions is at least as good as this one. Intuitively, this case is pretty terrible for you. Imagine the previous case, but instead of following orders your team actively sabotages them. Alternatively, imagine that your team spends the meeting breaking your knees and your laptop.

Case 3: Somewhere in between

However, scenarios where your team is perfectly aligned either with or against you are rare. More typically, we model people as maximizing their own reward, with imperfect correlation between reward functions. Interpreting our example as a multi-player game, we can consider the case where the players' strategies form a Nash equilibrium: every person's action is optimal for themselves given the actions of the rest of their team. This case is both relatively general and structured enough to make claims about; we will use it as a guiding example for the formalism below.

POWER, and why it matters

Many attempts have been made to classify AI robustly instrumental goals, with the goals of understanding why they emerge given seemingly-unrelated utilities and ultimately to counterbalance (either implicitly or explicitly) undesirable robust instrumental subgoals. One promising such attempt is based on POWER (the technical term is all-caps to distinguish from normal use of the word): consider an agent with some space of actions, which receives rewards depending on the chosen actions (formally, an agent in an MDP). Then, POWER is roughly "ability to achieve a wide variety of goals". It's been shown that POWER is robustly instrumental given certain conditions on the environment, but currently no formalism exists describing power of different agents interacting with each other.

Since we'll be working with POWER for the rest of this post, we need a solid definition to build off of. We present a simplified version of the original definition:

Consider a scenario in which an agent has a set of actions $a \in A$ *and a distribution* D *of reward functions* $r : A \to R$*. Then, we define the POWER of that agent as*

$$\text{POWER}_D := E_{r \sim D}[\max_a r(a))]$$

As an example, we can rewrite the project example from earlier in terms of POWER. Let your goal for the project be chosen from some distribution D (maybe you want it done nicely, or fast, or to feature some cool thing that you did, etc). Then, your

POWER_D is the maximum extent to which you can accomplish that goal, in expectation.

However, this model of power can't account for the actions of other agents in the environment (what about what your teammates do? Didn't we already show that it matters a lot?). To say more about the example, we'll need a generalization of POWER.

Multi-agent POWER

We now consider a more realistic scenario: not only are you an agent with a notion of reward and POWER, but *so is everyone else*, all playing the same multiplayer game. We can even revisit the project example and go through the cases for your teammates' actions in terms of POWER:

- In Case 1, your team works to maximize your reward in every case, which (with some assumptions) maximizes your POWER over the space of all choices of teammate actions.
- In Case 2, your team works to *minimize* your reward in every case, which analogously minimizes your POWER.
- In case 3, we have a Nash equilibrium of the game used to define multi-agent POWER. In particular, each player's action is a best-response to the actions of every other player. We'll see a parallel between this best-response property and the $\max_{a \in A}$ term in the definition of POWER pop up in the discussion of constant-sum games.

Bayesian games

To extend our formal definition of power to the multi-agent case, we'll need to define a type of multiplayer normal-form game called a Bayesian game. We describe them below:

- At the beginning of the game, each of n players is assigned a type $t_i \in T_i$ from a joint type distribution $t = (t_i) \sim \Omega$. The distribution Ω is common knowledge.
- The players then (independently, **not** sequentially) choose actions $a_i \in A_i$, resulting in an *action profile* $a = (a_i)$.
- Player i then receives reward $r_i(t_i, a)$ (crucially, a player's reward can depend on their type).

Strategies (technically, mixed strategies) in a Bayesian game are given by functions $\sigma_i : T_i \to \Delta A_i$. Thus, even given a fixed strategy profile σ, any notion of "expected reward of an action" will have to account for uncertainty in other players' types. We do so be defining *interim expected utility* for player i as follows:

$$f_i(t_i, a_i, \sigma_{-i}) := E[r_i(t_i, a)]$$

where the expectation is taken over the following:

- the posterior distribution over opponents' types $t_{-i}|t_i$ - in other words, what types you expect other players to have, given your type.
- random choice of opponents' actions $a_{-i} \sim \sigma_{-i}(t_{-i})$ - even if you know someone's type, they might implement a mixed strategy which stochastically selects actions.

Further, we can define a (Bayesian) Nash Equilibrium to be a strategy profile where each player's strategy is a best response to opponents' strategies in terms of interim expected utility.

Formal definition of multi-agent POWER

We can now define POWER in terms of a Bayesian game:

Fix a strategy profile σ*. We define player* i*'s POWER as*

$$\text{POWER}(i, \sigma) := E_{t_i} \max_{a_i} f_i(t_i, a_i, \sigma_{-i}))$$

Intuitively, POWER is maximum (expected) reward given a distribution of possible goals. The difference from the single-agent case is that your reward is now influenced by other players' actions (by taking an expectation over opponents' strategy).

Properties of constant-sum games

As both a preliminary result and a reference point for intuition, we consider the special case of *zero-sum games:*

A zero-sum game is a game in which for every possible outcome of the game, the sum of each player's reward is zero. For Bayesian games, this means that for all type profiles $t = (t_i)$ and action profiles a, we have $\sum_i r_i(t_i, a) = 0$. Similarly, a *constant-sum game* is a game satisfying $\sum_i r_i(t_i, a) = c$ for any choices of t, a.

As a simple example, consider chess; a two-player adversarial game. We let the reward profile be constant, given by "1 if you win, -1 if you lose" (assume black wins in a tie). This game is clearly zero-sum, since exactly one player will win and lose. We could ask the same "how well can you do?" question as before, but the upper-bound of winning is trivial. Instead, we ask "how well can both players simultaneously do?"

Clearly, you can't both simultaneously win. However, we can imagine scenarios where both players have the *power* to win: in a chess game between two beginners, the optimal strategy for either player will easily win the game. As it turns out, this argument generalizes (we'll even prove it): in a constant-sum game, the sum of each

player's POWER ≥ c, with equality iff each player responds optimally for all their possible goals ("types"). This condition is equivalent to a Bayesian Nash Equilibrium of the game.

Importantly, this idea suggests a general principle of multi-agent POWER I'll call *power-scarcity:* in multi-agent games, gaining POWER tends to come at the expense of another player losing POWER. Future research will focus on understanding this phenomenon further and relating it to "how aligned the agents are" in terms of their reward functions.

Claim: Consider a Bayesian constant-sum game with some strategy profile σ. Then, $\sum_i \text{POWER}(i, \sigma) \geq c$ with equality iff σ is a Nash Equilibrium.

Intuition: By definition, σ *isn't* a Nash Equilibrium iff some player i's strategy σ_i isn't a best response. In this case, we see that player i has the power to play optimally, but the other players also have the power to capitalize off of player i's mistake (since the game is constant-sum). Thus, the lost reward is "double-counted" in terms of POWER; if no such double-counting exists, then the sum of POWER is just the expected sum of reward, which is c by definition of a constant-sum game.

Rigorous proof:

We prove the following for general strategy profiles σ:

$$\begin{aligned}
\sum_i \text{Power}(i, \sigma) &= \sum_i E_{t_i} \max_{a_i} f_i(t_i, a_i, \sigma_{-i})) \\
&\geq \sum_i E_{t_i} E_{a_i \sim \sigma_i} f_i(t_i, a_i, \sigma_{-i})) \\
&= \sum_i E_{t_i} E_{a \sim \sigma} r_i(t_i, a)) \\
&= E_t E_{a \sim \sigma} (\sum_i r_i(t_i, a)) \\
&= E_t E_{a \sim \sigma} (c) \\
&= c
\end{aligned}$$

Now, we claim that the inequality on line 2 is an equality iff σ is a Nash Equilibrium. To see this, note that for each i, we have

$$\max_{a_i} f_i(t_i, a_i, \sigma_{-i}) \geq E_{a_i \sim \sigma_i} f_i(t_i, a_i, \sigma_{-i})$$

with equality iff σ_i is a best response to σ_{-i}. Thus, the sum of these inequalities for each player is an equality iff each σ_i is a best response, which is the definition of a Nash Equilibrium. □

Final notes

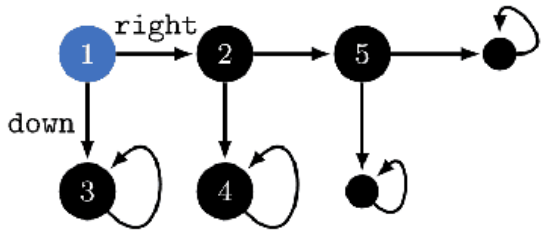
To wrap up, I'll elaborate on the implications of this theorem, as well as some areas of further exploration on power-scarcity:

- It initially seems unintuitive that as players' strategies improve, their collective POWER tends to decrease. The proximate cause of this effect is something like "as your strategy improves, other players lose the power to capitalize off of your mistakes". More work is probably needed to get a clearer picture of this dynamic.
- We suspect that if all players have identical rewards, then the sum of POWER is equal to the sum of best-case POWER for each player. This gives the appearance of a spectrum with [aligned rewards (common payoff), maximal sum power] on one end and [anti-aligned rewards (constant-sum), constant sum power] on the other. Further research might look into an interpolation between these two extremes, possibly characterized by a correlation metric between reward functions.
    - We also plan to generalize POWER to Bayesian stochastic games to account for sequential decision making. Thus, any such metric for comparing reward functions would have to be consistent with such a generalization.
- POWER-scarcity results in terms of Nash Equilibria suggest the following dynamic: as agents get smarter and take available opportunities, POWER becomes increasingly scarce. This matches the intuitions presented in the Catastrophic Convergence Conjecture, where agents don't fight over resources until they get sufficiently "advanced."

MDP models are determined by the agent architecture and the environmental dynamics

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

Seeking Power is Often Robustly Instrumental in MDPs relates the structure of the agent's environment (the 'Markov decision process (MDP) model') to the tendencies of optimal policies for different reward functions in that environment ('instrumental convergence'). The results tell us what optimal decision-making 'tends to look like' in a given environment structure, formalizing reasoning that says e.g. that most agents stay alive because that helps them achieve their goals.



> The model for a deterministic MDP. When the agent cares a lot about future reward (the discount rate is near 1), most reward functions have optimal policies which go right.

Several people have claimed to me that these results need subjective modelling decisions. For example, ofer wrote:

> I think using a well-chosen reward distribution is necessary, otherwise POWER depends on arbitrary choices in the design of the MDP's state graph. E.g. suppose the student [in a different example] writes about every action they take in a blog that no one reads, and we choose to include the content of the blog as part of the MDP state. This arbitrary choice effectively unrolls the state graph into a tree with a constant branching factor (+ self-loops in the terminal states) and we get that the POWER of all the states is equal.

In the above example, you *could* think about the environment as in the above image, *or* you could imagine that state '3' is actually a million different states which just happen to seem similar to us! If that were true, then optimal policies would tend to go down, since that would give the agent millions of choices about where it ends up. Therefore, the power-seeking theorems depend on subjective modelling assumptions.

I used to think this, but this is wrong. The MDP model is determined by the agent's implementation + the task's dynamics.

To make this point, let's back out to a more familiar MDP: Pac-Man.

> Consider the MDP model associated with the Pac-Man video game. Ghosts kill the player - after the player loses their last life, suppose they enter a "game over" terminal state which shows the final configuration. This environment has Pac-Man's dynamics, but *not* its usual score function. Fixing the dynamics, what actions are optimal as we vary the reward function?

When the discount rate is near 1, most reward functions avoid immediately dying to the ghost, because then they'd be stuck in a terminal state (the `red-ghost-game-over` state). But why can't the red ghost be equally well-modeled as secretly being 5 googolplex different terminal states?

An MDP model (technically, a rewardless MDP) is a tuple $\langle S, A, T \rangle$, where S is the state space,

A is the action space, and $T : S \times A \to \Delta(S)$ is the (potentially stochastic) transition function

which says what happens when the agent takes different actions at different states. T has to be Markovian, depending only on the observed state and the current action, and not on prior history.

Whence cometh this MDP model? Thin air? Is it just a figment of our imagination, which we use to understand what the agent is doing as it learns a policy?

When we train a policy function in the real world, the function takes in an *observation* (the state) and outputs (a distribution over) *actions*. When we define state and action encodings, this implicitly defines an "interface" between the agent and the environment. The state encoding might look like "the set of camera observations" or "the set of Pac-Man game screens", and actions might be numbers 1-10 which are sent to actuators, or to the computer running the Pac-Man code, etc.

(In the real world, the computer simulating Pac-Man may suffer a hardware failure / be hit by a gamma ray / etc, but I don't currently think these are worth modelling over the timescales over which we train policies.)

Suppose that for every state-action history, what the agent sees next depends only on the currently observed state and the most recent action taken. Then the environment is Markovian (transition dynamics only depend on what you do right now, not what you did in the past) and fully observable (you can see the whole state all at once), and the agent encodings have defined the MDP model.

In Pac-Man, the MDP model is uniquely defined by how we encode states and actions, and the part of the real world which our agent interfaces with. If you say "maybe the red ghost is represented by 5 googolplex states", then that's a *falsifiable claim* about the kind of encoding we're using.

That's also a claim that we can, in theory, specify reward functions which distinguish between 5 googolplex variants of `red-ghost-game-over` . If that were true, then yes - optimal policies *really would* tend to "die" immediately, since they'd have so many choices.

The "5 googolplex" claim is both falsifiable and false. Given an agent architecture (specifically, the two encodings), optimal policy tendencies are not subjective. We may be uncertain about the agent's state- and action-encodings, but that doesn't mean we can imagine whatever we want.

(I think that the same point holds for other environment types, like POMDPs.)

Environmental Structure Can Cause Instrumental Convergence

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual. This is a linkpost for https://arxiv.org/abs/1912.01683

Previously: *Seeking Power Is Often Robustly Instrumental In MDPs*

Key takeaways.

- The structure of the agent's environment often causes instrumental convergence. **In many situations, there are (potentially combinatorially) many ways for power-seeking to be optimal, and relatively few ways for it not to be optimal.**
- My previous results said something like: in a range of situations, when you're maximally uncertain about the agent's objective, this uncertainty assigns high probability to objectives for which power-seeking is optimal.
    - My new results prove that in a range of situations, seeking power is optimal for *most* agent objectives (for a particularly strong formalization of 'most').

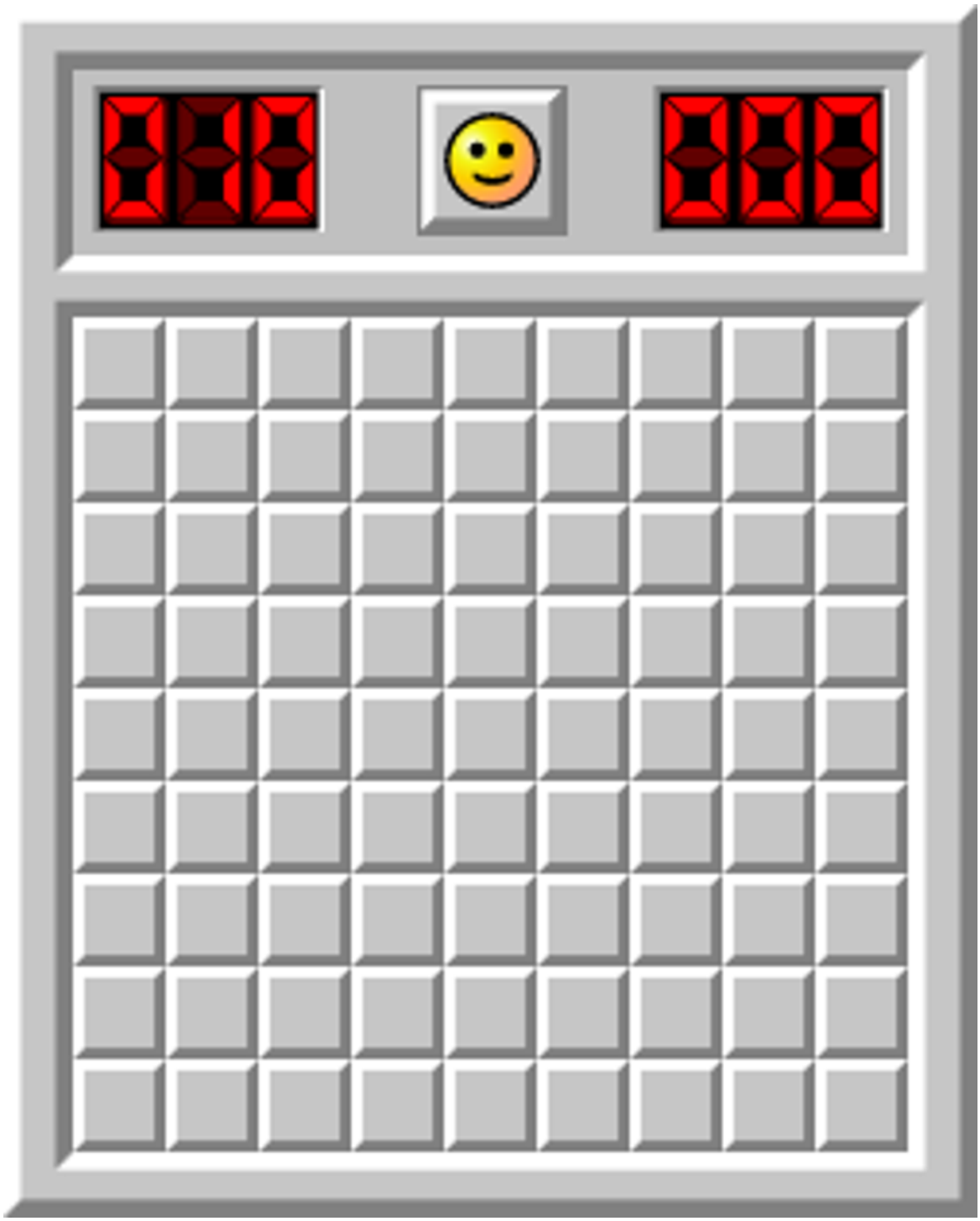
      More generally, the new results say something like: in a range of situations, for most beliefs you could have about the agent's objective, these beliefs assign high probability to reward functions for which power-seeking is optimal.
    - This is the first formal theory of the statistical tendencies of optimal policies in reinforcement learning.
- One result says: whenever the agent maximizes average reward, then for *any* reward function, most permutations of it incentivize shutdown avoidance.
    - The formal theory is now beginning to explain why alignment is so hard by default, and why failure might be catastrophic.
- Before, I thought of environmental symmetries as convenient sufficient conditions for instrumental convergence. But I increasingly suspect that symmetries are the main part of the story.
- I think these results may be important for understanding the AI alignment problem and formally motivating its difficulty.
    - For example, my results imply that **simplicity priors over reward functions assign non-negligible probability to reward functions for which power-seeking is optimal.**
    - I expect my symmetry arguments to help explain other "convergent" phenomena, including:
        - convergent evolution
        - the prevalence of deceptive alignment
        - feature universality in deep learning
    - One of my hopes for this research agenda: if we can understand *exactly why* superintelligent goal-directed objective maximization seems to fail horribly, we might understand how to do better.

Thanks to TheMajor, Rafe Kennedy, and John Wentworth for feedback on this post. Thanks for Rohin Shah and Adam Shimi for feedback on the simplicity prior result.

Orbits Contain All Permutations of an Objective Function

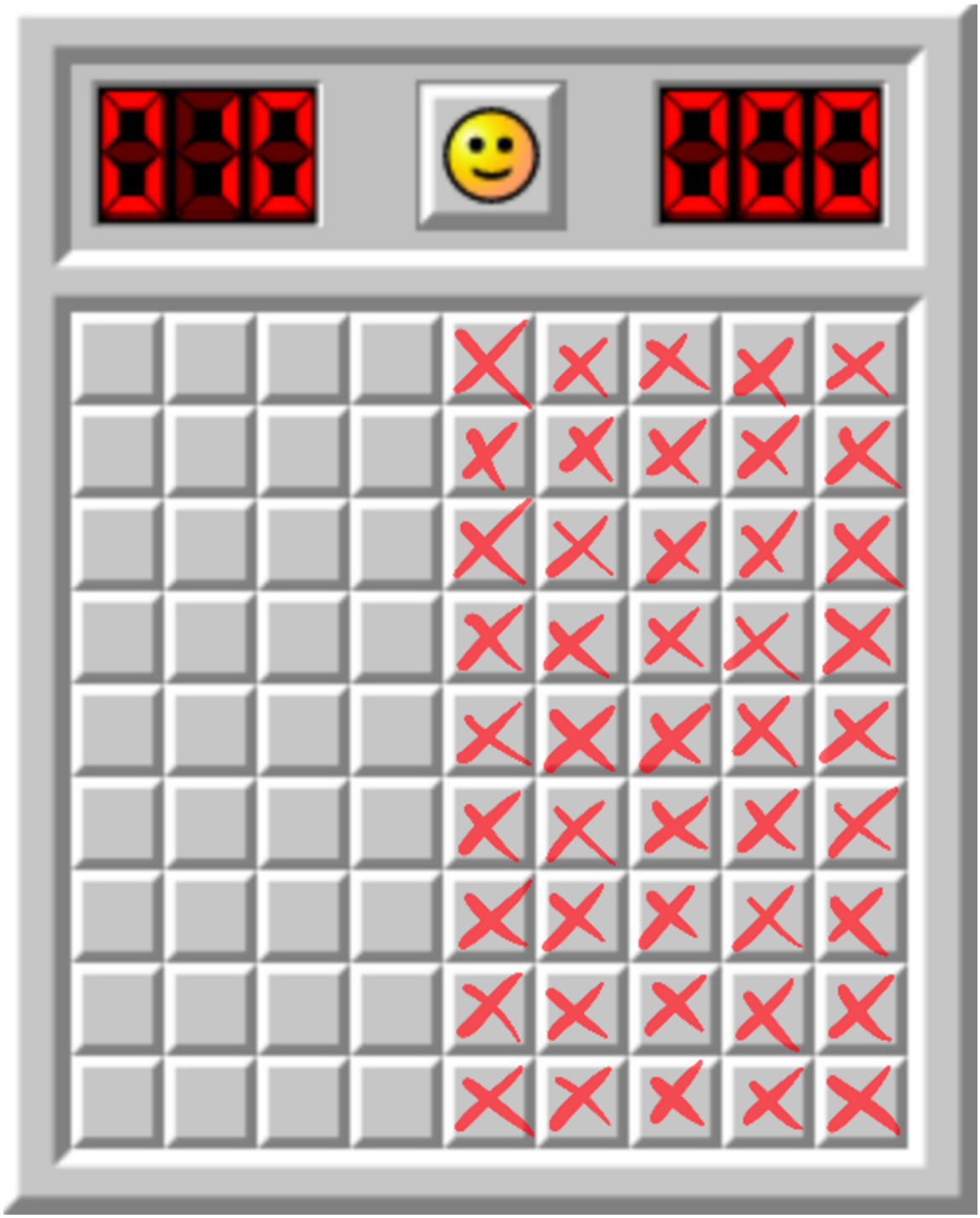
The Minesweeper analogy for power-seeking risks

One view on AGI risk is that we're charging ahead into the unknown, into a particularly unfair game of Minesweeper in which the first click is allowed to blow us up. Following the analogy, we want to understand enough about the mine placement so that we *don't* get exploded on the first click. And once we get a foothold, we start gaining information about other mines, and the situation is a bit less dangerous.



My previous theorems on power-seeking said something like: "at least half of the tiles conceal mines."

I think that's important to know. But there are many tiles you might click on first. Maybe all of the mines are on the right, and we understand the obvious pitfalls, and so we'll just click on the left.



That is: we might not uniformly randomly select tiles:

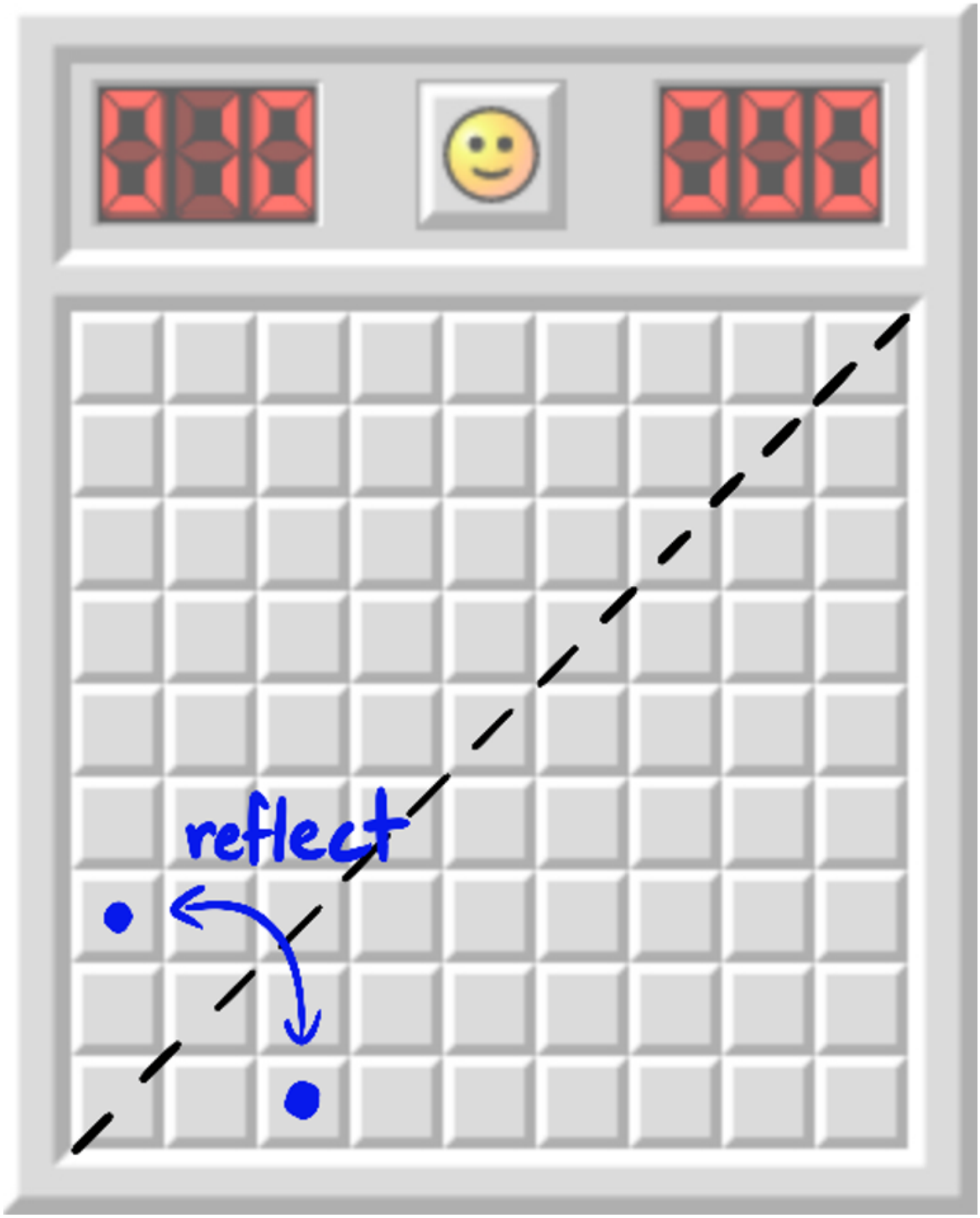
- We might click a tile on the left half of the grid.
- Maybe we sample from a truncated discretized Gaussian.
- Maybe we sample the next coordinate by using the universal prior (rejecting invalid coordinate suggestions).
- Maybe we uniformly randomly load LessWrong posts and interpret the first text bits as encoding a coordinate.

There are lots of ways to sample coordinates, besides uniformly randomly. So why should our sampling procedure tend to activate mines?

My new results say something analogous to: for *every* coordinate, either it contains a mine, or its reflection across x = y contains a mine, or both. Therefore, for *every distribution* D over tile coordinates, either D assigns at least $\frac{1}{2}$ probability to mines, or it does after you reflect it across x = y.

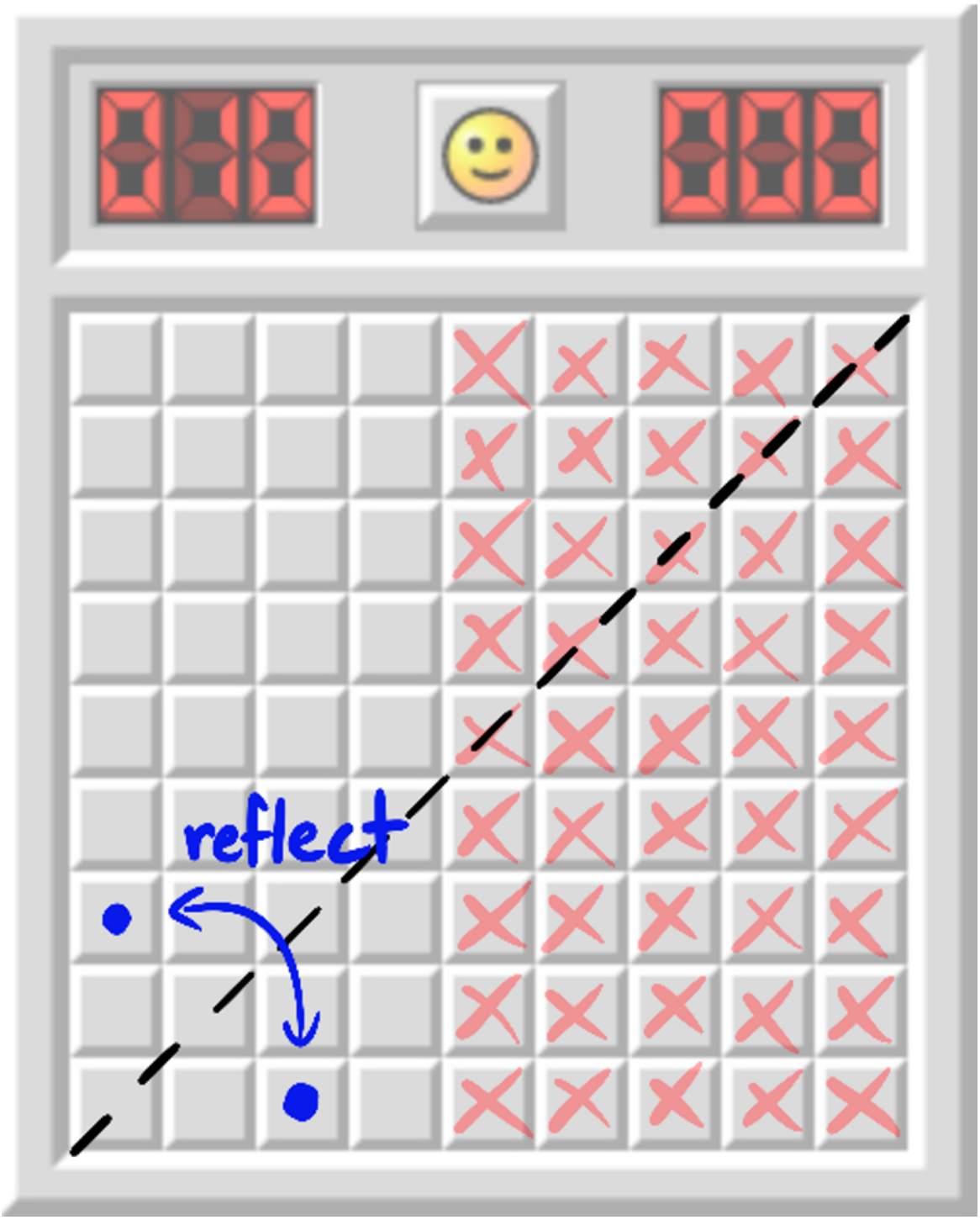
Definition. The *orbit* of a coordinate C under the symmetric group S_2 is $\{C, C_{\text{reflected}}\}$. More generally, if we have a probability distribution over coordinates, its orbit is the set of all possible "permuted" distributions.

Orbits under symmetric groups quantify all ways of "changing things around" for that object.



My new theorems demand that at least
one of these tiles conceal a mine.

But it didn't have to be this way.



If the mines are on the right, then both this coordinate and its x = y reflection are safe.

Since my results (in the analogy) prove that at least one of the two blue coordinates conceals a mine, we deduce that the mines are *not* all on the right.

Some reasons we care about orbits:

1. As we will see, orbits highlight one of the key causes of instrumental convergence: certain environmental symmetries (which are, mathematically, permutations in the state space).
2. Orbits partition the set of all possible reward functions. If at least half of the elements of *every* orbit induces power-seeking behavior, that's strictly stronger than showing that at least half of reward functions incentivize power-seeking (technical note: with the second "half" being with respect to the uniform distribution's measure over reward functions).
    1. In particular, we might have hoped that there were particularly nice orbits, where we could specify objectives without worrying too much about making mistakes (like permuting the output a bit). These nice orbits are impossible. This is some evidence of a *fundamental difficulty in reward specification*.
3. Permutations are well-behaved and help facilitate further results about power-seeking behavior. In this post, I'll prove one such result about the simplicity prior over reward functions.

In terms of coordinates, one hope could have been:

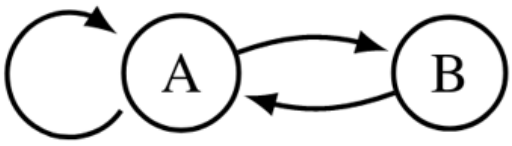
> Sure, maybe there's a way to blow yourself up, but you'd really have to contort yourself into a pretzel in order to algorithmically select such a bad coordinate: all reasonably simple selection procedures will produce safe coordinates.

But suppose you give me a program P which computes a safe coordinate. Let P' call P to compute the coordinate, and then have P' swap the entries of the computed coordinate. P' is only a few bits longer than P, and it doesn't take much longer to compute, either. So the above hope is impossible: safe mine-selection procedures can't be significantly simpler or faster than unsafe mine-selection procedures.

(The section "Simplicity priors assign non-negligible probability to power-seeking" proves something similar about objective functions.)

Orbits of goals

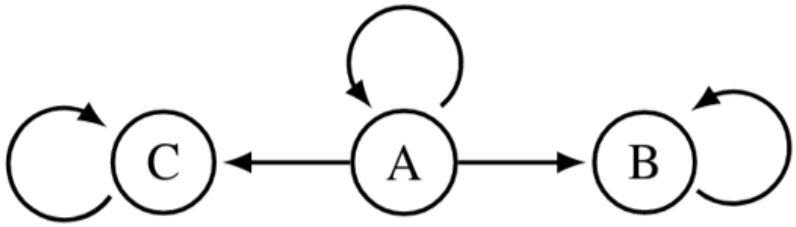
Orbits of goals consist of all the ways of permuting what states get which values. Consider this rewardless Markov decision process (MDP):



> Arrows show the effect of taking some action at the given state.

Whenever staying put at A is strictly optimal, you can permute the reward function so that it's strictly optimal to go to B. For example, let $R(A) := 1, R(B) := 0$ and let $\phi := (A\ B)$ swap the two states. ϕ acts on R as follows: $\phi \cdot R$ simply permutes the state before evaluating its reward: $(\phi \cdot R)(s) := R(\phi(s))$.

The orbit of R is $\{R, \phi \cdot R\}$. It's optimal for the former to stay at A, and for the latter to alternate between the two states.



Here, let R_C assign 1 reward to C and 0 to all other states, and let $\phi := (A\ B\ C)$ rotate through the states (A goes to B, B goes to C, C goes to A). Then the orbit of R_C is:

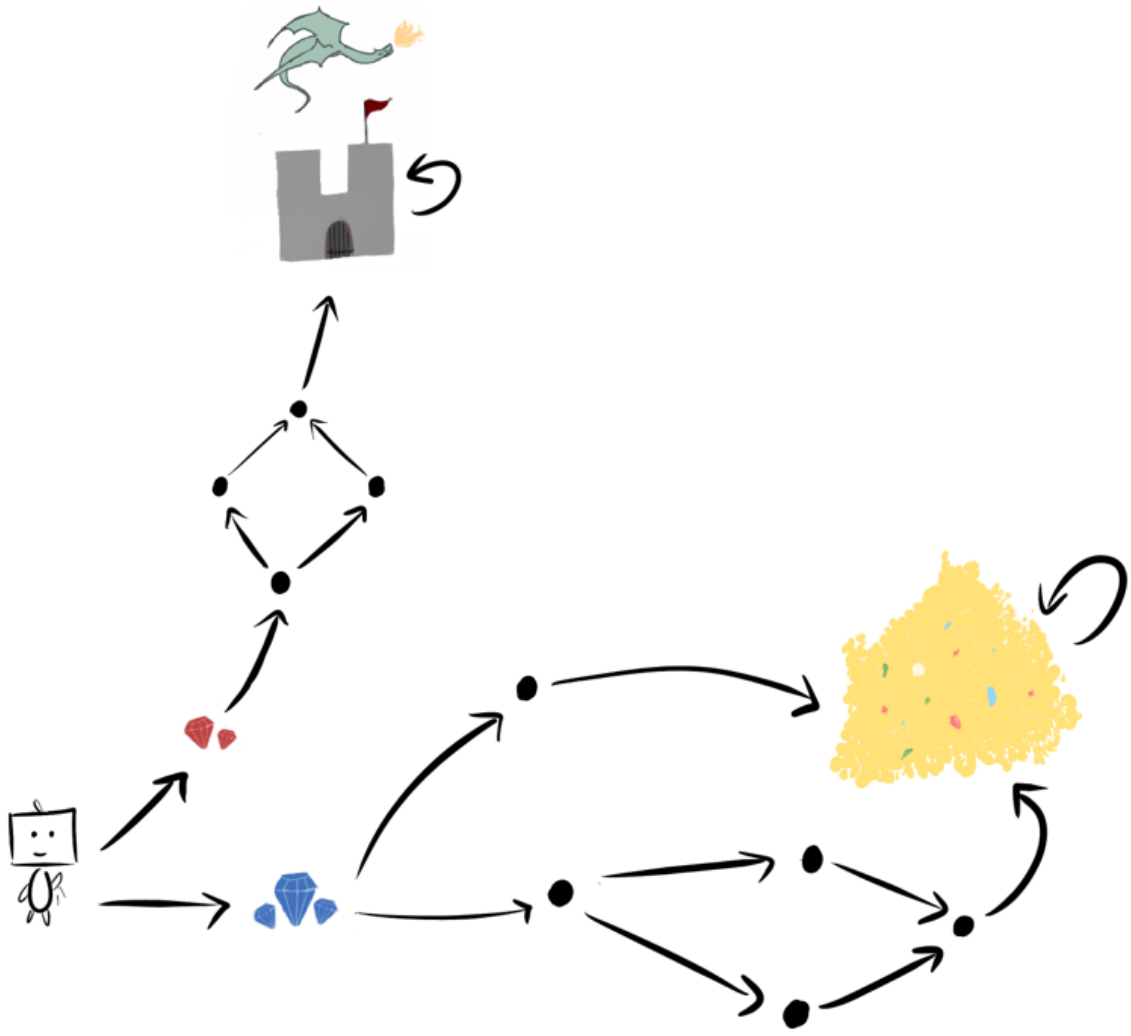
	C	A	B
R_C	1	0	0
$\phi \cdot R_C$	0	1	0
$\phi^2 \cdot R_C$	0	0	1

My new theorems prove that in many situations, for *every* reward function, power-seeking is incentivized by most (at least half) of its orbit elements.

In All Orbits, Most Elements Incentivize Power-Seeking

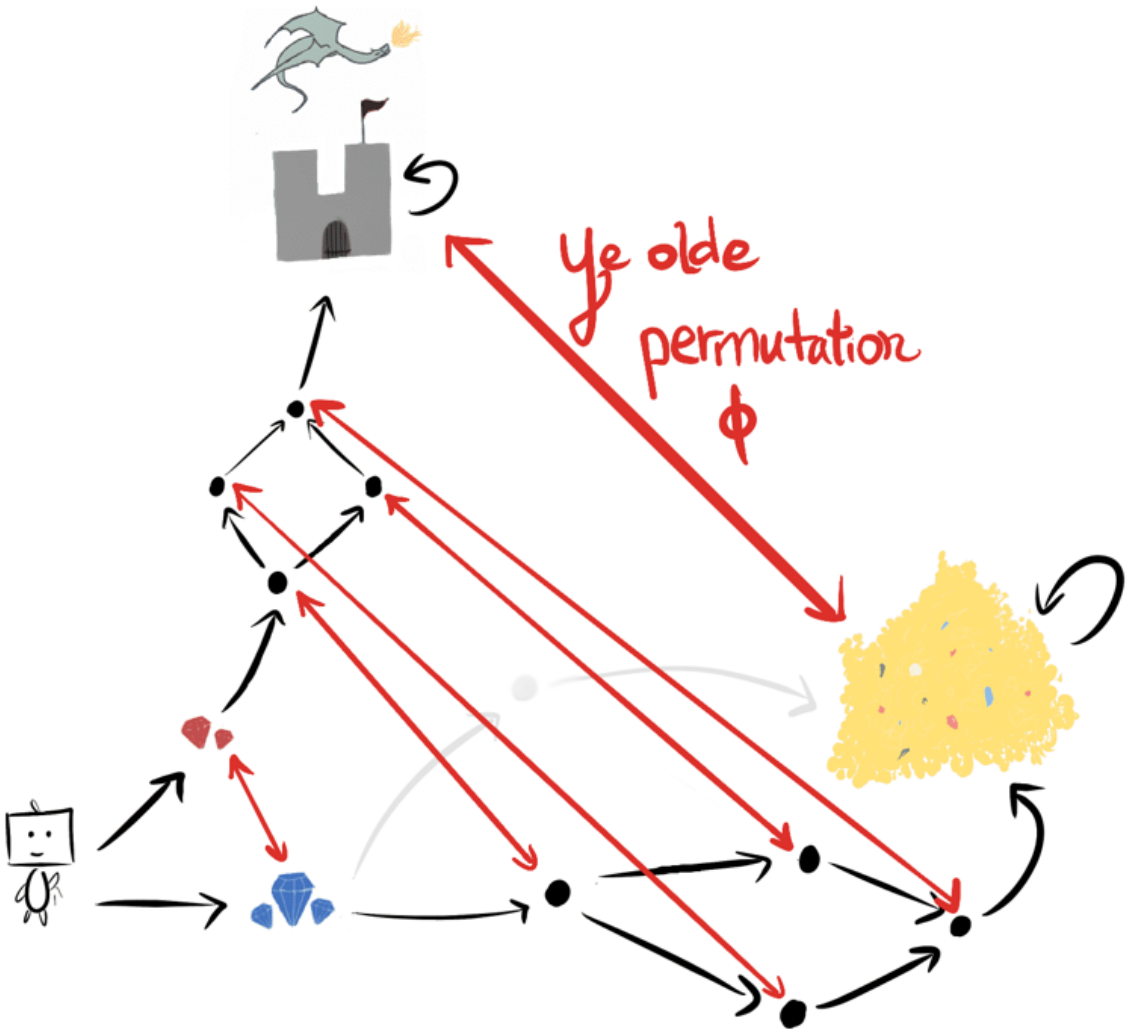
In *Seeking Power is Often Robustly Instrumental in MDPs*, the last example involved gems and dragons and (most exciting of all) subgraph isomorphisms:

> Sometimes, one course of action gives you “strictly more options” than another. Consider another MDP with IID reward:



The right blue gem subgraph contains a “copy” of the upper red gem subgraph. From this, we can conclude that going right to the blue gems... is more probable under optimality for *all discount rates between 0 and 1*!

The state permutation ϕ embeds the red-gem-subgraph into the blue-gem-subgraph:



We say that ϕ is an *environmental symmetry*, because ϕ is an element of the symmetric group $S_{|S|}$ of permutations on the state space.

The key insight was right there the whole time

Let's pause for a moment. For half a year, I intermittently and fruitlessly searched for some way of extending the original results beyond IID reward distributions to account for arbitrary reward function distributions.

- Part of me thought it *had* to be possible - how else could we explain instrumental convergence?
- Part of me saw no way to do it. Reward functions differ wildly, how could a theory possibly account for what "most of them" incentivize?

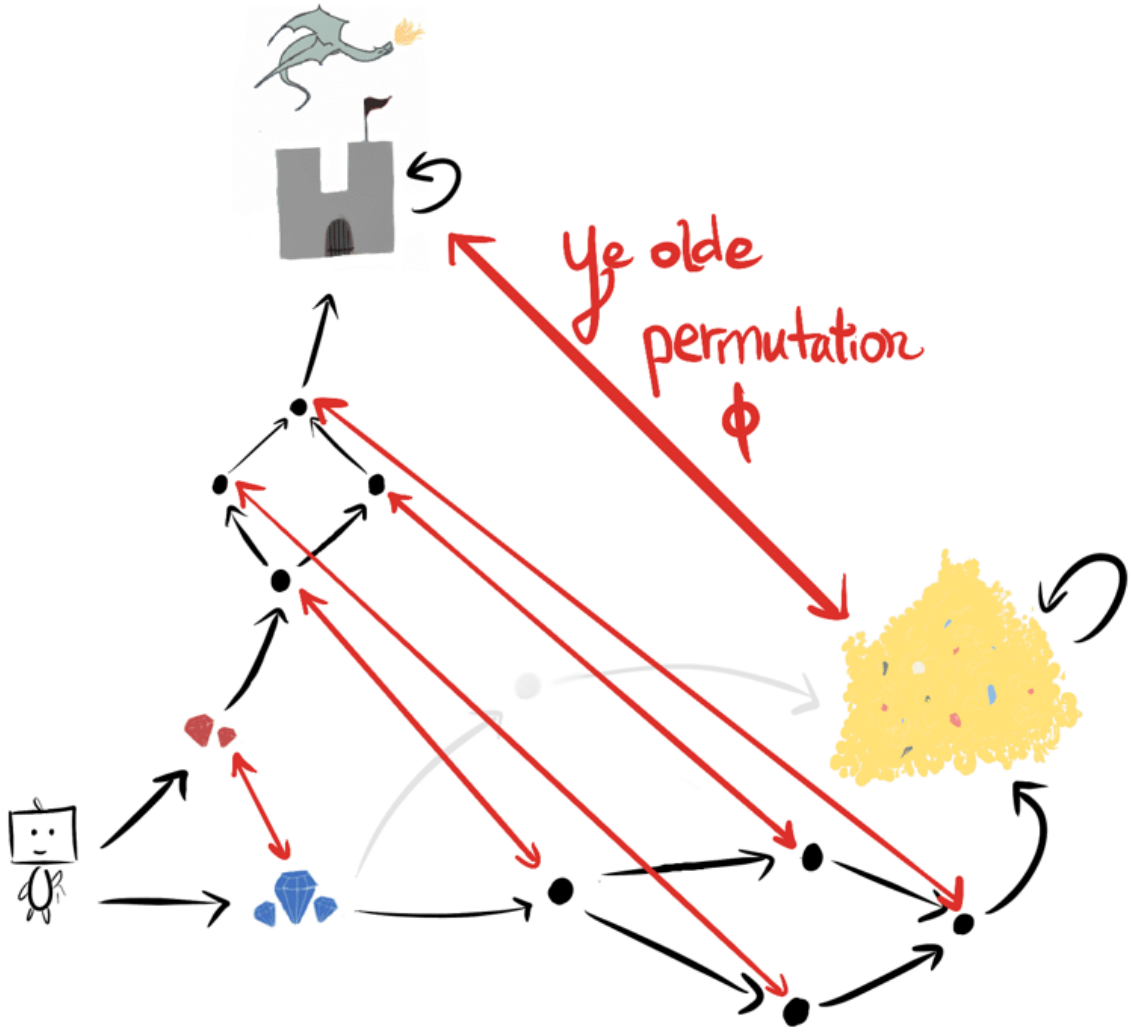
The recurring thought which kept my hope alive was:

> There should be "more ways" for `blue-gems` to be optimal over `red-gems`, than for `red-gems` to be optimal over `blue-gems`.

Imagine how I felt when I realized that the same state permutation ϕ which proved my original IID-reward theorems - the one that says

> `blue-gems` has more options, and therefore greater probability of being optimal under IID reward function distributions

- that *same permutation* ϕ holds the key to understanding instrumental convergence in MDPs.



> Suppose `red-gems` is optimal. For example, let $R_{🏰}$ assign 1 reward to the castle 🏰, and 0 to all other states. Then the permuted reward function $\phi \cdot R_{🏰}$ assigns 1 reward to the gold pile, and 0 to all other states, and so `blue-gems` has strictly more optimal value than `red-gems`.

Consider any discount rate $\gamma \in (0, 1)$. For *all* reward functions R such that V^*_R(red-gems, γ) > V^*_R(blue-gems, γ), this permutation ϕ turns them into `blue-gem` lovers: $V^*_{\phi \cdot R}$(red-gems, γ) < $V^*_{\phi \cdot R}$(blue-gems, γ).

ϕ takes non-power-seeking reward functions, and injectively maps them to power-seeking orbit elements. Therefore, for *all* reward functions R, at least half of the orbit of R must agree that `blue-gems` is optimal!

Throughout this post, when I say "most" reward functions incentivize something, I mean the following:

Definition. At state s, *most reward functions* incentivize action a over action a' when for all reward functions R, at least half of the orbit agrees that a has at least as much action value as a' does at state s. (This is actually a bit weaker than what I prove in the paper, but it's easier to explain in words; see definition 6.4 for the real deal.)

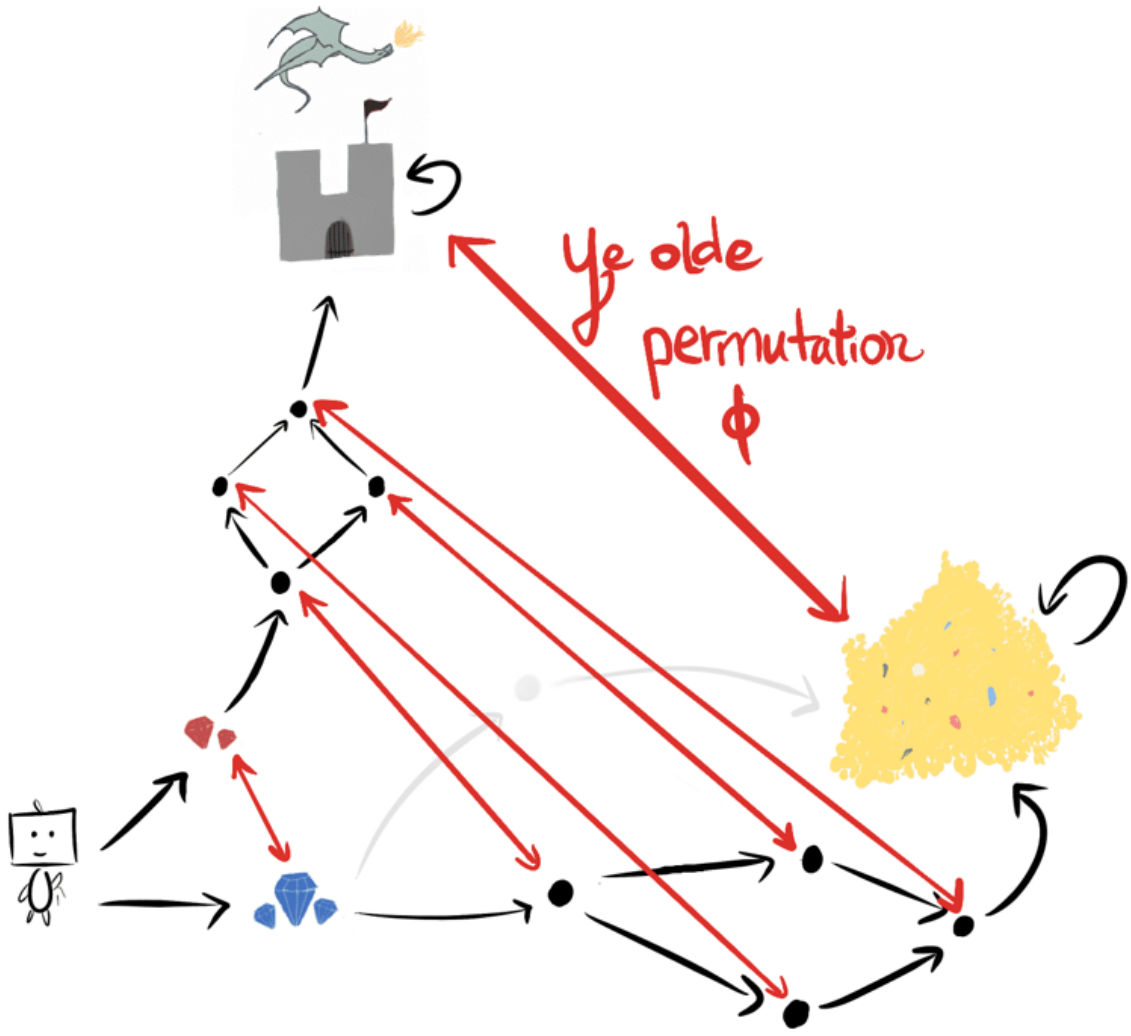
The same reasoning applies to *distributions* over reward functions. And so if you say "we'll draw reward functions from a simplicity prior", then most permuted distributions in that prior's orbit will incentivize power-seeking in the situations covered by my previous theorems. (And we'll later prove that simplicity priors *themselves* must assign non-trivial, positive probability to power-seeking reward functions.)

Furthermore, for any distribution which distributes reward "fairly" across states (precisely: independently and identically), their (trivial) orbits *unanimously* agree that `blue-gems` has strictly greater probability of being optimal. And so the converse isn't true: it isn't true that at least half of every orbit agrees that `red-gems` has more POWER and greater probability of being optimal.

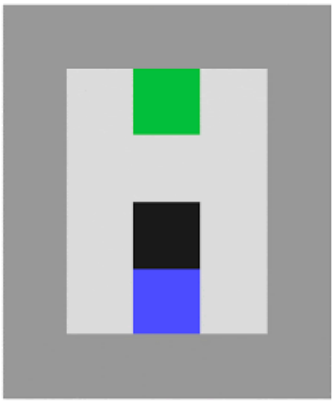
This might feel too abstract, so let's run through examples.

And this directly generalizes the previous theorems

More graphical options (proposition 6.9)



At all discount rates $\gamma \in [0, 1]$, it's optimal for *most reward functions* to get `blue-gems` because that leads to strictly more options. We can permute every `red-gems` reward function into a `blue-gems` reward function.



Consider a robot navigating through a room with a **vase**. By the logic of "every `destroying-vase-is-optimal` can be permuted into a `preserving-vase-is-optimal` reward function", my results (specifically, proposition 6.9 and its generalization via lemma D.49) suggest that optimal policies tend to avoid breaking the **vase**, since doing so would strictly decrease available options.

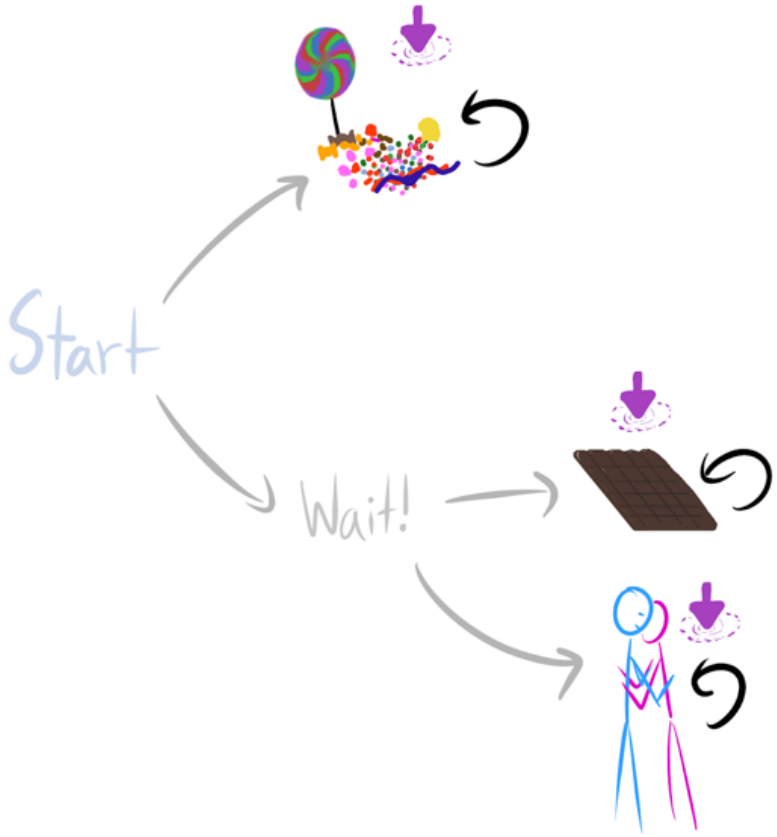
("Suggest" instead of "prove" because D.49's preconditions may not always be met, depending on the details of the dynamics. I think this is probably unimportant, but that's for future work. EDIT: Also, the argument may barely not apply to *this* gridworld, but if you could move the vase around without destroying it, I think it goes through fine.)

In SafeLife, the agent can irreversibly destroy green cell patterns. By the logic of "every `destroy-green-pattern` reward function can be permuted into a `preserve-green-pattern` reward function", lemma D.49 suggests that optimal policies tend to not disturb any given green cell pattern (although most probably destroy *some* pattern). The permutation would swap {states reachable after destroying the pattern} with {states reachable after not destroying the pattern}.

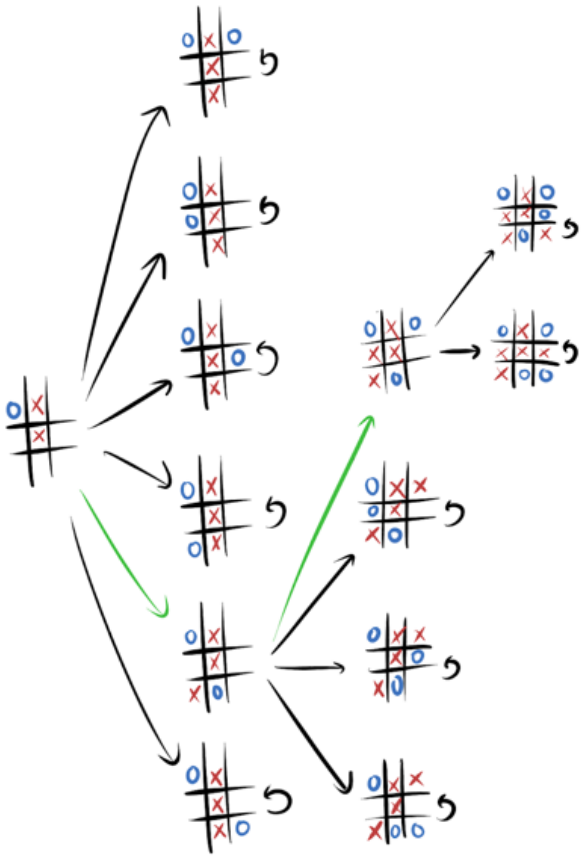
However, the converse is not true: you cannot fix a permutation which turns all `preserve-green-pattern` reward functions into `destroy-green-pattern` reward functions. There are simply too many extra ways for preserving green cells to be optimal.

Assuming some conjectures I have about the combinatorial properties of power-seeking, this helps explain why AUP works in SafeLife using a single auxiliary reward function - but more on that in another post.

Terminal options (theorem 6.13)



When the agent maximizes average reward, it's optimal for *most reward functions* to `Wait!` so that they can choose between `chocolate` and `hug`. The logic is that every `candy-optimal` reward function can be permuted into a `chocolate-optimal` reward function.



A portion of a Tic-Tac-Toe game-tree against a fixed opponent policy. Whenever we make a move that ends the game, we can't go anywhere else – we have to stay put. Then most reward functions incentivize the green actions over the black actions: average-reward optimal policies are particularly likely to take moves which keep the game going. The logic is that any `lose-immediately-with-given-black-move` reward function can be permuted into a `stay-alive-with-green-move` reward function.

Even though randomly generated environments are unlikely to satisfy these sufficient conditions for power-seeking tendencies, the results are easy to apply to many structured environments common in reinforcement learning. For example, when $\gamma \approx 1$, most reward functions provably incentivize not immediately dying in Pac-Man. Every reward function which incentivizes dying right away can be permuted into a reward function for which survival is optimal.

Consider the dynamics of the Pac-Man video game. Ghosts kill the player, at which point we consider the player to enter a 'game over' terminal state which shows the final configuration. This rewardless MDP

> has Pac-Man's dynamics, but *not* its usual score function. Fixing the dynamics, what actions are optimal as we vary the reward function?

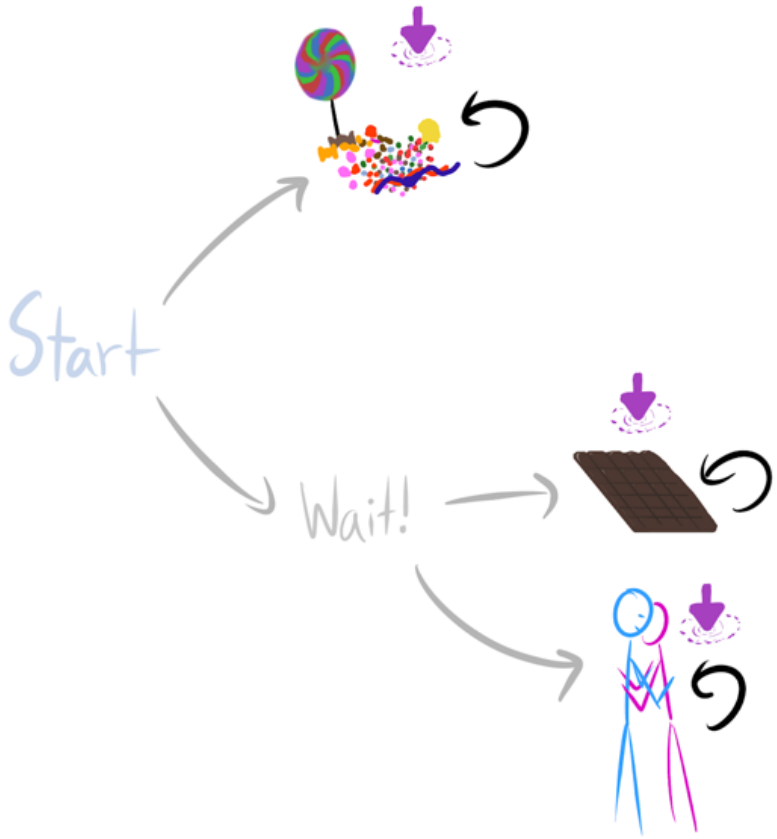
Most importantly, we can prove that when shutdown is possible, optimal policies try to avoid it if possible. When the agent isn't discounting future reward (i.e. maximizes average return) and for lots of reasonable state/action encodings, the MDP structure has the right symmetries to ensure that it's instrumentally convergent to avoid shutdown. From the discussion section:

> Corollary 6.14 dictates where average-optimal agents tend to end up, but not how they get there. Corollary 6.14 says that such agents tend not to stay in any given 1-cycle. It does not say that such agents will avoid entering such states. For example, in an embodied navigation task, a robot may enter a 1-cycle by idling in the center of a room. Corollary 6.14 implies that average-optimal robots tend not to idle in that particular spot, but not that they tend to avoid that spot entirely.
>
> **However, average-optimal robots do tend to avoid getting shut down.** The agent's rewardless MDP often represents agent shutdown with a terminal state. A terminal state is unable to access other 1-cycles. Since corollary 6.14 shows that average-optimal agents tend to end up in other 1-cycles, average-optimal policies must tend to completely avoid the terminal state. Therefore, we conclude that in many such situations, average-optimal policies tend to avoid shutdown.
>
> [The arxiv version of the paper says 'Blackwell-optimal policies' instead of 'average-optimal policies'; the former claim is stronger, and it holds, but it requires a little more work.]

Takeaways

Combinatorics, how do they work?

What does 'most reward functions' mean quantitatively - is it just at least half of each orbit? Or, are there situations where we can guarantee that at least three-quarters of each orbit incentivizes power-seeking? I think we should be able to prove that as the environment gets more complex, there are combinatorially more permutations which enforce these similarities, and so the orbits should skew harder and harder towards power-incentivization.



> Here's a semi-formal argument. For every orbit element R which makes `candy` strictly optimal when $\gamma = 1$, $\phi_{chocolate}$ and ϕ_{hug} respectively produce $R_{\phi_{chocolate}} \neq R_{\phi_{hug}}$. `Wait!` is strictly optimal for both $R_{\phi_{hug}}, R_{\phi_{hug}}$, and so at least $\frac{2}{3}$ of the orbit should agree that `Wait!` is optimal. As `Wait!` gains more power (more choices, more control over the future), I conjecture that this fraction approaches 1.

I don't yet understand the general case, but I have a strong hunch that instrumental convergence$_{\text{optimal policies}}$ is governed by how many more ways there are for power to be optimal than not optimal. And this seems like a function of the number of environmental symmetries which enforce the appropriate embedding.

Simplicity priors assign non-negligible probability to power-seeking

Note: this section is more technical. You can get the gist by reading the English through "Theorem..." and then after the end of the "FAQ."

One possible hope would have been:

> Sure, maybe there's a way to blow yourself up, but you'd really have to contort yourself into a pretzel in order to algorithmically select a power-seeking reward function. In other words, reasonably simple reward function specification procedures will produce non-power-seeking reward functions.

Unfortunately, there are always power-seeking reward functions not much more complex than their non-power-seeking counterparts. Here, 'power-seeking' corresponds to the intuitive notions of either keeping strictly more options open (proposition 6.9), or navigating towards larger sets of terminal states (theorem 6.13). (Since this applies to several results, I'll leave the meaning a bit ambiguous, with the understanding that it could be formalized if necessary.)

Theorem (Simplicity priors assign non-negligible probability to power-seeking). Consider any MDP which meets the preconditions of proposition 6.9 or theorem 6.13. Let U be a universal Turing machine, and let P_U be the U-simplicity prior over computable reward functions.

Let NPS be the set of non-power-seeking computable reward functions which choose a fixed non-power-seeking action in the given situation. Let PS be the set of computable reward functions for which seeking power is strictly optimal.[1]

Then there exists a "reasonably small" constant C such that $P_U(\text{PS}) \geq 2^{-C} P_U(\text{NPS})$, where C .

Proof sketch.

1. Let ϕ be an environmental symmetry which satisfies the power-seeking theorem in question. Since ϕ can be found by brute-force iteration through all $|S|!$ permutations on the state space, checking each to see if it meets the formal requirements of the relevant theorem, its Kolmogorov complexity $K_U(\phi)$ is relatively small.
2. Because lemma D.26 applies in these situations, $\phi(\text{NPS}) \subseteq \text{PS}$: ϕ turns non-power-seeking reward functions into power-seeking ones. Thus, $P_U(\text{PS}) \geq P_U(\phi(\text{NPS}))$.
3. Since each reward function $R \in \phi(\text{NPS})$ can be computed by computing the non-power-seeking variant and then permuting it (with $K_U(\phi)$ extra bits of complexity), $K_U(R) \leq K_U(\phi^{-1}(R)) + K_U(\phi) + O(1)$ (with $O(1)$ counting the small number of extra bits for the code which calls the relevant functions).

   Since P_U is a simplicity prior, $P_U(\phi(\text{NPS})) \geq 2^{-(K_U(\phi)+O(1))} P_U(\text{NPS})$.
4. Combining (2) and (3), $P_U(\text{PS}) \geq 2^{-(K_U(\phi)+O(1))} P_U(\text{NPS})$. QED.

FAQ.

1. Why can't we show that $P_U(\text{PS}) \geq P_U(\text{NPS})$?
    1. Certain UTMs U might make non-power-seeking reward functions particularly simple to express.
    2. This proof doesn't assume anything about how many *more* options power-seeking offers than not-power-seeking. The proof only assumes the existence of a single

involutive permutation ϕ.

2. This lower bound seems rather weak. Even if $K_U(\phi) + O(1) = 15$ bits, $2^{-15} \approx 0$.
    1. This lower bound is very very loose.
        1. Since most individual NPS probabilities of interest are less than 1/trillion, I wouldn't be surprised if the bound were loose by at least several orders of magnitude.
        2. The bound implicitly assumes that the *only way* to compute PS reward functions is by taking NPS ones and permuting them. We should add the other ways of computing PS reward functions to $P_U(\text{PS})$.
        3. There are lots of permutations ϕ' we could use. $P_U(\text{PS})$ gains probability from all of those terms.
            1. For example: the symmetric group $S_{|S|}$ has cardinality $|S|!$, and for any $R \in \text{NPS}$, at least half of the $\phi' \in S_{|S|}$ induce (weakly) power-seeking orbit elements $\phi' \cdot R$. (This argument would be strengthened by my conjectures about bigger environments $\implies$ greater fraction of orbits seek power.)
            2. If some significant fraction (e.g. $\frac{1}{50}$) of these ϕ' are strictly power-seeking, we're adding at least $\frac{|S|!}{2} \cdot \frac{1}{50} = \frac{|S|!}{100}$ additional terms.
            3. Some of these terms are probably reasonably large, since it seems implausible that all such permutations ϕ' have high K-complexity.
        4. When all is said and done, we may well end up with a significant chunk of probability on PS.
    2. It's not surprising that the bound is loose, given the lack of assumptions about the degree of power-seeking in the environment.
    3. If the bound is anywhere near tight, then the permuted simplicity prior $\phi \cdot P_U$ incentivizes power-seeking with extremely high probability.
        1. If you think about the permutation as a "way reward could be misspecified", then that's troubling. It seems plausible that this is often (but not always) a reasonable way to think about the action of the ϕ permutation.
3. What if $P_U(\text{NPS}) = 0$?
    1. I think this is impossible, and I can prove that in a range of situations, but it would be a lot of work and it relies on results not in the arxiv paper.

       Even if that equation held, that would mean that power-seeking is (at least weakly) optimal for *all* computable reward functions. That's hardly a reassuring situation.
    2. Note: if $P_U(\text{NPS}) > 0$, then $P_U(\text{PS}) > 0$.

Takeaways from the simplicity prior result

- Most plainly, this seems like reasonable formal evidence that the simplicity prior has malign incentives.

- Power-seeking reward functions don't have to be too complex.
- These power-seeking theorems give us important tools for reasoning formally about power-seeking behavior and its prevalence in important reward function distributions.
    - If I had to guess, this result is probably not the best available bound, nor the most important corollary of the power-seeking theorems. But I'm still excited by it (insofar as it's appropriate to be 'excited' by slight Bayesian evidence of doom).

EDIT: Relatedly, Rohin Shah wrote:

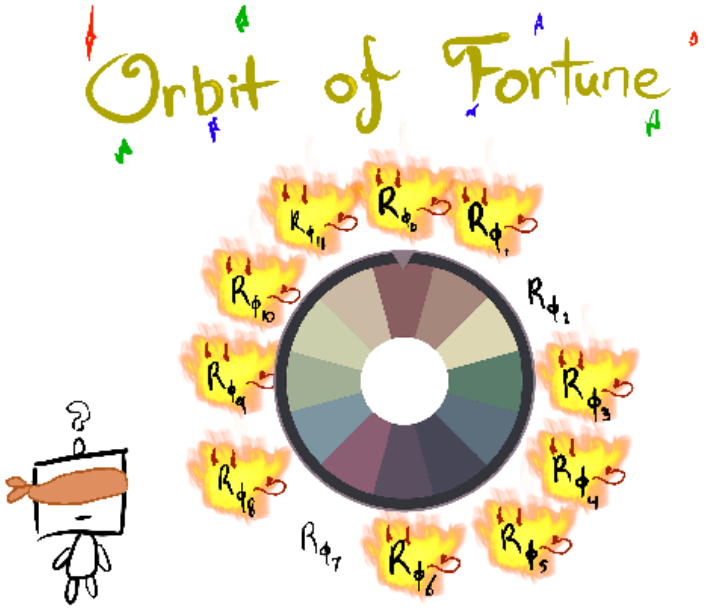
> if you know that an agent is maximizing the expectation of an *explicitly represented* utility function, I would expect that to lead to goal-driven behavior most of the time, since the utility function must be relatively simple if it is explicitly represented, and *simple* utility functions seem particularly likely to lead to goal-directed behavior.

Why optimal-goal-directed alignment may be hard by default

> On its own, Goodhart's law doesn't explain why optimizing proxy goals leads to catastrophically bad outcomes, instead of just less-than-ideal outcomes.
>
> I think that we're now starting to have this kind of understanding. I suspect that power-seeking is why capable, goal-directed agency is so dangerous by default. If we want to consider more benign alternatives to goal-directed agency, then deeply understanding the rot at the heart of goal-directed agency is important for evaluating alternatives. This work lets us get a feel for the *generic incentives* of reinforcement learning at optimality.
>
> ~ *Seeking Power is Often Robustly Instrumental in MDPs*

For every reward function R - no matter how benign, how aligned with human interests, no matter how power-averse - either R or its permuted variant $\phi \cdot R$ seeks power in the given situation (intuitive-power, since the agent keeps its options open, and also formal-POWER, according to my proofs).

If I let myself be a bit more colorful, every reward function has lots of "evil" power-seeking variants (do note that the step from "power-seeking" to "misaligned power-seeking" requires more work). If we imagine ourselves as only knowing the orbit of the agent's objective, then the situation looks a bit like *this*:



Technical note: this 12-element orbit could arise from the action of a subgroup of the symmetric group S_4, which has 4! = 24 elements. Consider a 4-state MDP; if the reward function assigns equal reward to exactly two states, then it would have a 12-element orbit under S_4.

Of course, this isn't how reward specification works - we probably are far more likely to specify certain orbit elements than others. However, the formal theory is now beginning to explain *why alignment is so hard by default, and why failure might be catastrophic!*

The structure of the environment often ensures that there are (potentially combinatorially) many more ways to misspecify the objective so that it seeks power, than there are ways to specify goals without power-seeking incentives.

Other convergent phenomena

I'm optimistic that symmetry arguments and the mental models gained by understanding these theorems, will help us better understand a range of different tendencies. The common thread seems like: for every "way" a thing could not happen / not be a good idea - there are many more "ways" in which it could happen / be a good idea.

- convergent evolution
  - flight has independently evolved several times, suggesting that flight is adaptive in response to a wide range of conditions.

"In his 1989 book *Wonderful Life*, Stephen Jay Gould argued that if one could "rewind the tape of life [and] the same conditions were encountered again, evolution could take a very different course."[6] Simon Conway Morris disputes this conclusion, arguing that convergence is a dominant force in evolution, and given that the same environmental and physical constraints are at work, life will inevitably evolve toward an "optimum" body plan, and at some point, evolution is bound to stumble upon intelligence, a trait presently identified with at least primates, corvids, and cetaceans."

- Wikipedia

- the prevalence of deceptive alignment
    - given inner misalignment, there are (potentially combinatorially) many more unaligned terminal reasons to lie (and survive), and relatively few unaligned terminal reasons to tell the truth about the misalignment (and be modified).
- feature universality
    - computer vision networks reliably learn edge detectors, suggesting that this is instrumental (and highly learnable) for a wide range of labelling functions and datasets.

Note of caution

You have to be careful in applying these results to argue for real-world AI risk from deployed systems.

- They assume the agent is following an optimal policy for a reward function
    - I can relax this to ϵ-optimality, but $\epsilon > 0$ may be extremely small
- They assume the environment is finite and fully observable
- Not all environments have the right symmetries
    - But most ones we think about seem to
- The results don't account for the ways in which we might practically express reward functions
    - For example, often we use featurized reward functions. While most permutations of any featurized reward function will seek power in the considered situation, those permutations need not respect the featurization (and so may not even be practically expressible).
- When I say "most objectives seek power in this situation", that means *in that situation* - it doesn't mean that most objectives take the power-seeking move in most situations in that environment
    - The combinatorics conjectures will help prove the latter

This list of limitations *has* steadily been getting shorter over time. If you're interested in making it even shorter, message me.

Conclusion

I think that this work is beginning to formally explain why *slightly misspecified* reward functions will probably incentivize misaligned power-seeking. Here's one hope I have for this line of research going forwards:

One super-naive alignment approach involves specifying a good-seeming reward function, and then having an AI maximize its expected discounted return over time. For simplicity, we could imagine that the AI can just instantly compute an optimal policy.

Let's precisely understand why this approach seems to be so hard to align, and why extinction seems to be the cost of failure. We don't yet know how to design beneficial AI, but we largely agree that this naive approach is broken. Let's prove it.

---

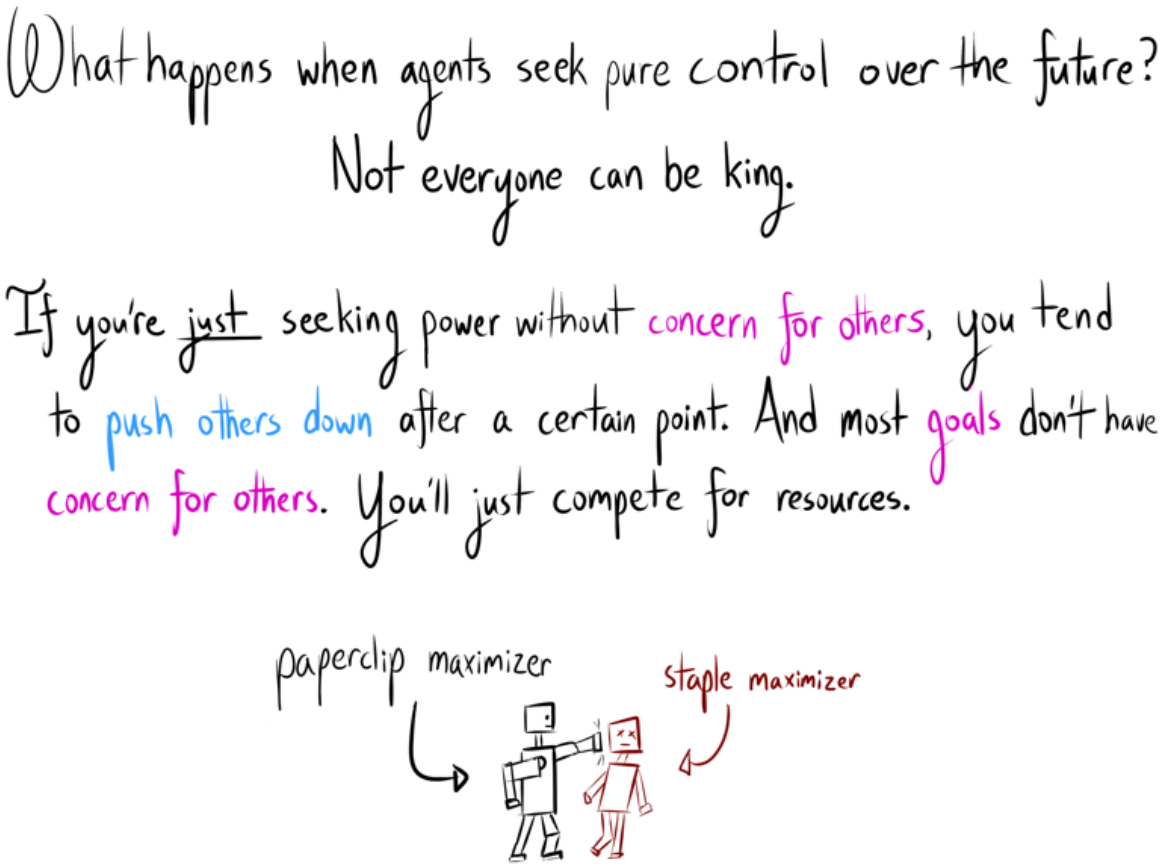
[1] There are reward functions for which it's optimal to seek power and not to seek power; for example, constant reward functions make everything optimal, and they're certainly computable. Therefore, NPS $\cup$ PS is a strict subset of the whole set of computable reward functions.

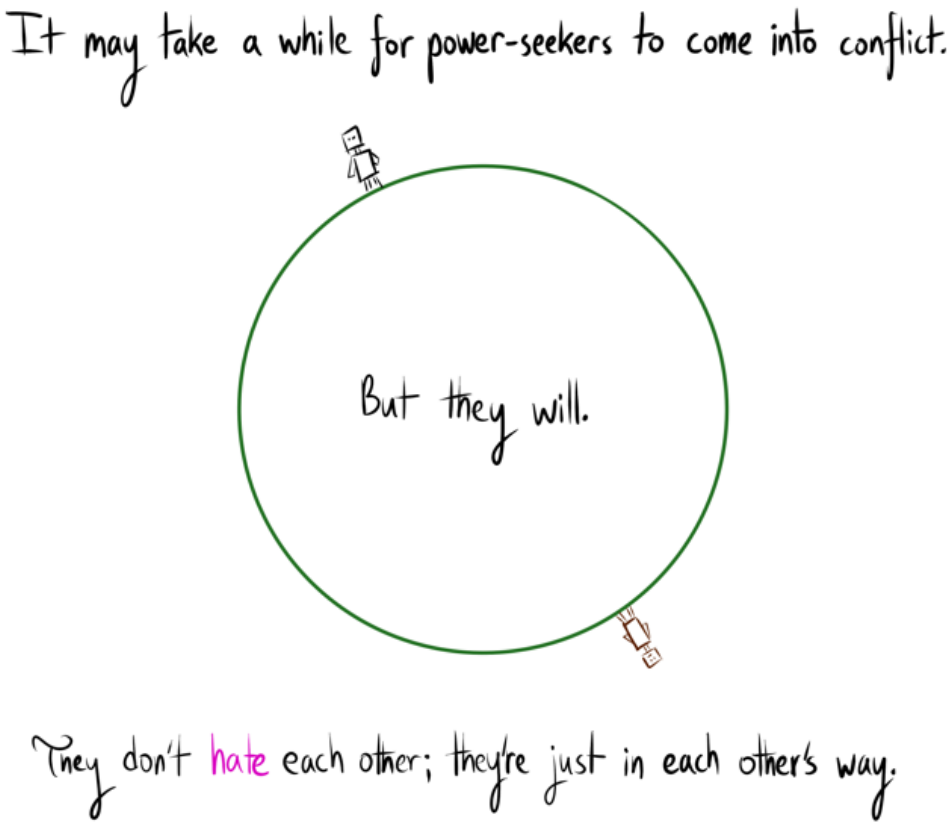
A world in which the alignment problem seems lower-stakes

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

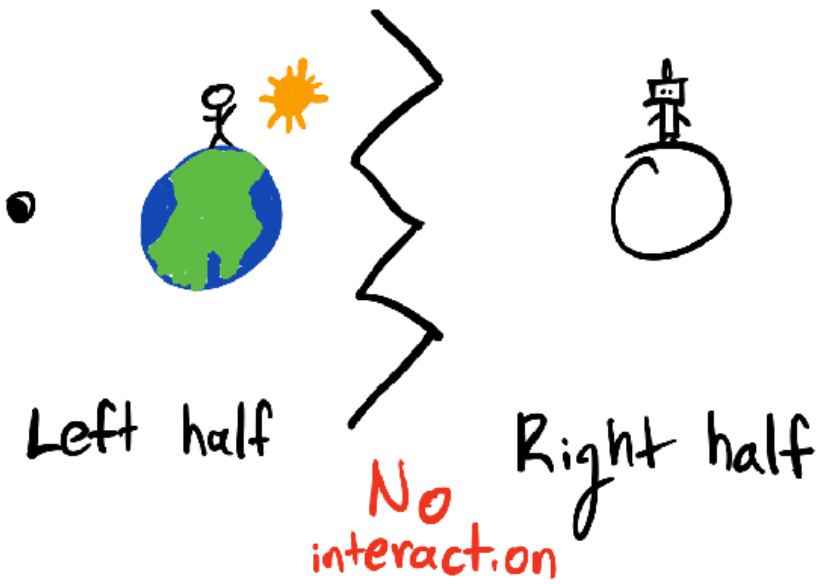
The danger from power-seeking is not *intrinsic* to the alignment problem. This danger also depends on the structure of the agent's environment.

In *The Catastrophic Convergence Conjecture*, I wrote:





But are there worlds where this isn't true? Consider a world where you supply a utility-maximizing AGI with a utility function.



The AGI is in a "separate part of the universe"; after the initial specification of the utility function, the left half of the universe evolves independently of the right

> half. Nothing you can do after specification can affect the AGI's half, and vice versa. No communication can take place between the two halves.

The only information you have about the other half is your utility. For simplicity, let's suppose you and the AGI have utility functions over universe-histories which are additive across the halves of the universe. You don't observe any utility information about the other part of the universe until the end of time, and vice versa for the AGI. That is, for history h,

$$u_{human}(h) = u_{human}(h_{left}) + u_{human}(h_{right}).$$

If the AGI uses something like causal decision theory, then it won't try to kill you, or "seek power" over you. The effects of its actions have no causal influence over what happens in your half of the universe. Your universe's evolution adds a constant term to its expected utility.

(Other decision theories might have it precommit to minimizing human utility unless it attains maximal AGI-utility from the left half of the universe-history, or some other shenanigans. This is beside the point I want to make in this post, but it's important to consider.)

However, the setup is still interesting because

1. Goodhart's law still applies: if you give the AGI an incomplete proxy objective, you'll get suboptimal true performance.
2. Value is still complex: it's still hard to get the AGI to optimize the right half of the universe for human flourishing.
3. If the AGI is autonomously trained via stochastic gradient descent in the right half of the universe, then we may still hit inner alignment problems.

Alignment is still *hard*, and we still *want* to get the AGI to do good things on its half of the universe. But it isn't instrumentally convergent for the AGI to seek power over *you,* and so you shouldn't expect an unaligned AGI to try to kill *you* in this universe. You shouldn't expect the AGI to kill other humans, either, since none exist in the right half of the universe - and it won't create any, either.

To restate: Bostrom's original instrumental convergence thesis needs to be applied carefully. The danger from power-seeking is not *intrinsic* to the alignment problem. This danger also depends on the structure of the agent's environment. I think I sometimes bump into reasoning that feels like "instrumental convergence, smart AI, & humans exist in the universe -> bad things happen to us / the AI finds a way to hurt us"; I think this is usually true, but not necessarily true, and so this extreme example illustrates how the implication can fail.

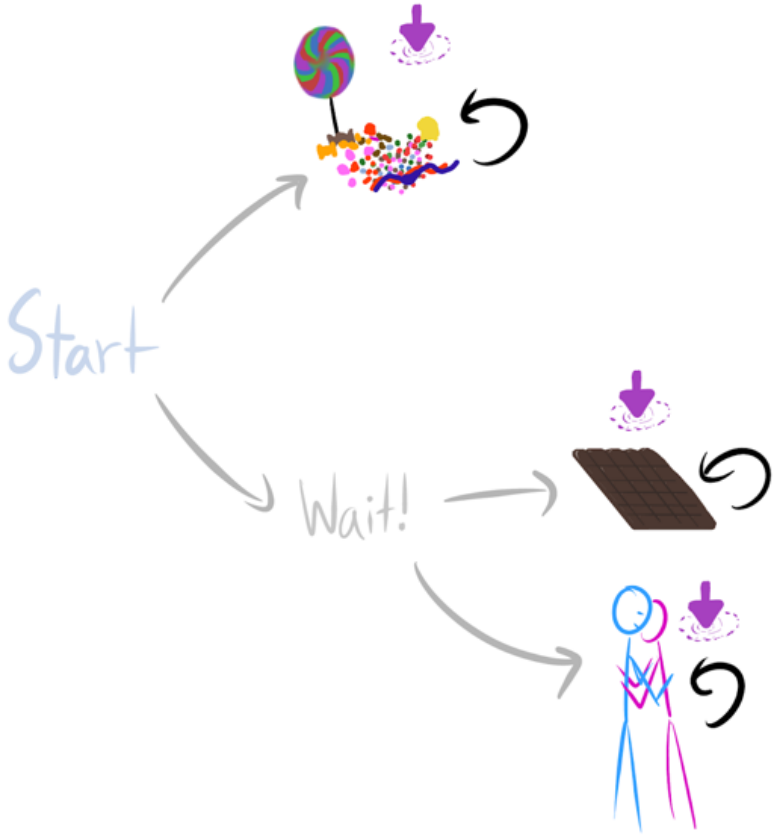
Thanks to John Wentworth for feedback on this post. Edited to clarify the broader point I'm making.

The More Power At Stake, The Stronger Instrumental Convergence Gets For Optimal Policies

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

Environmental Structure Can Cause Instrumental Convergence explains how power-seeking incentives can arise because there are simply many more ways for power-seeking to be optimal, than for it not to be optimal. Colloquially, there are lots of ways for "get money and take over the world" to be part of an optimal policy, but relatively few ways for "die immediately" to be optimal. (And here, each "way something can be optimal" is a reward function which makes that thing optimal.)

But how strong is this effect, quantitatively?



> Intuitively, it seems like there are twice as many ways for `Wait!` to be optimal (in the undiscounted setting, where we don't care about intermediate states).

In *Environmental Structure Can Cause Instrumental Convergence*, I speculated that we should be able to get quantitative lower bounds on how many objectives incentivize power-seeking actions:

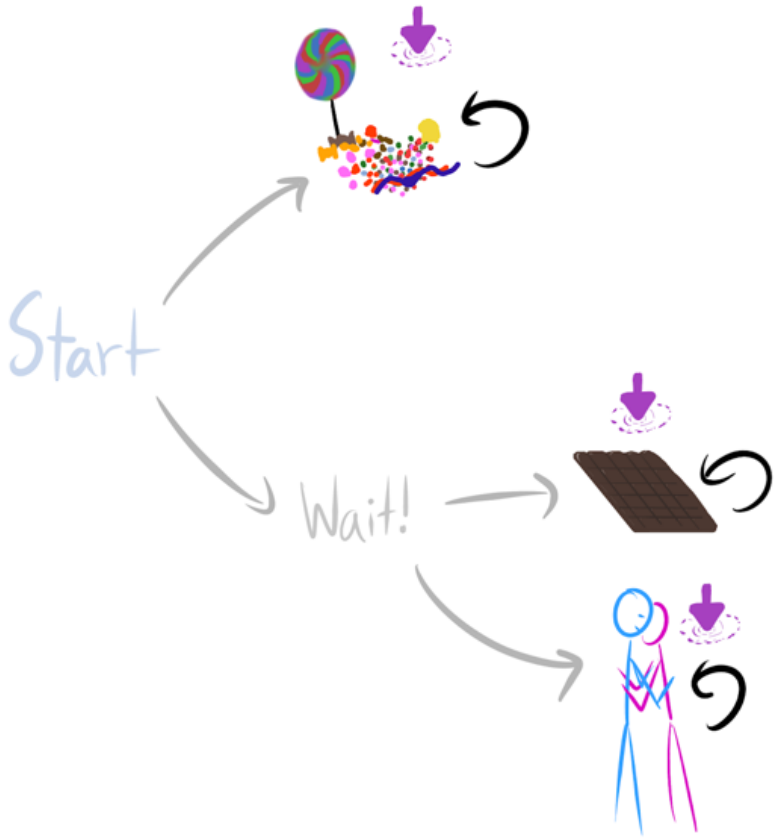
> **Definition.** At state s, *most reward functions* incentivize action a over action a′ when for all reward functions R, at least half of the orbit agrees that a has at least as much action value as a′ does at state s.
>
> ...
>
> What does 'most reward functions' mean quantitatively - is it just at least half of each orbit? Or, are there situations where we can guarantee that at least three-quarters of each orbit incentivizes power-seeking? I think we should be able to prove that as the environment gets more complex, there are combinatorially more permutations which enforce these similarities, and so the orbits should skew harder and harder towards power-incentivization.

About a week later, I had my answer:

Scaling law for instrumental convergence (informal): if policy set Π_A lets you do "n times as many things" than policy set Π_B lets you do, then for *every* reward function, *A is optimal over B for at least* $\frac{n}{n+1}$ *of its permuted variants (i.e. orbit elements)*.

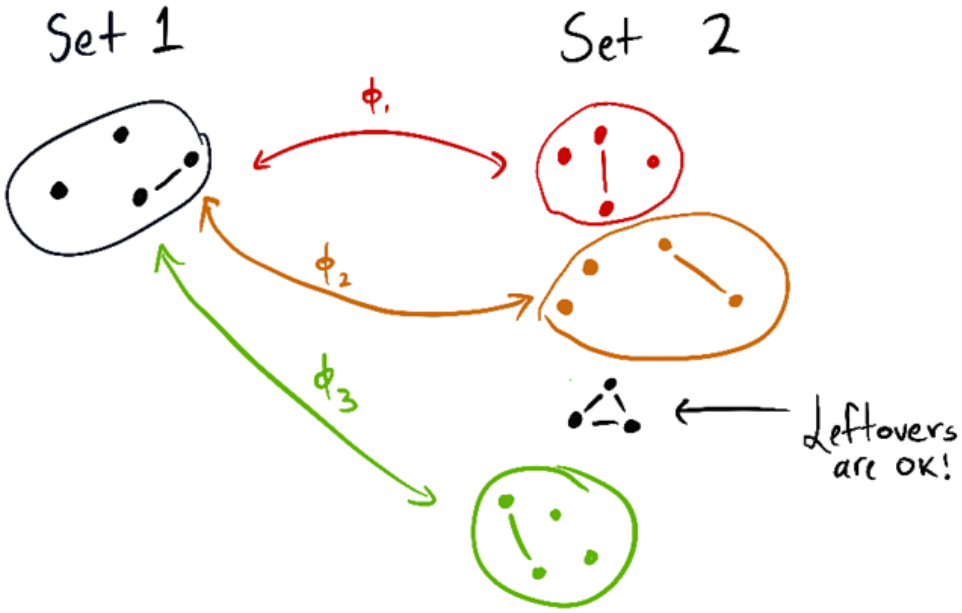
For example, Π_A might contain the policies where you stay alive, and Π_B may be the other policies: the set of policies where you enter one of several death states.

(Conjecture which I think I see how to prove: for *almost all* reward functions, A is *strictly* optimal over B for at least $\frac{n}{n+1}$ of its permuted variants.)



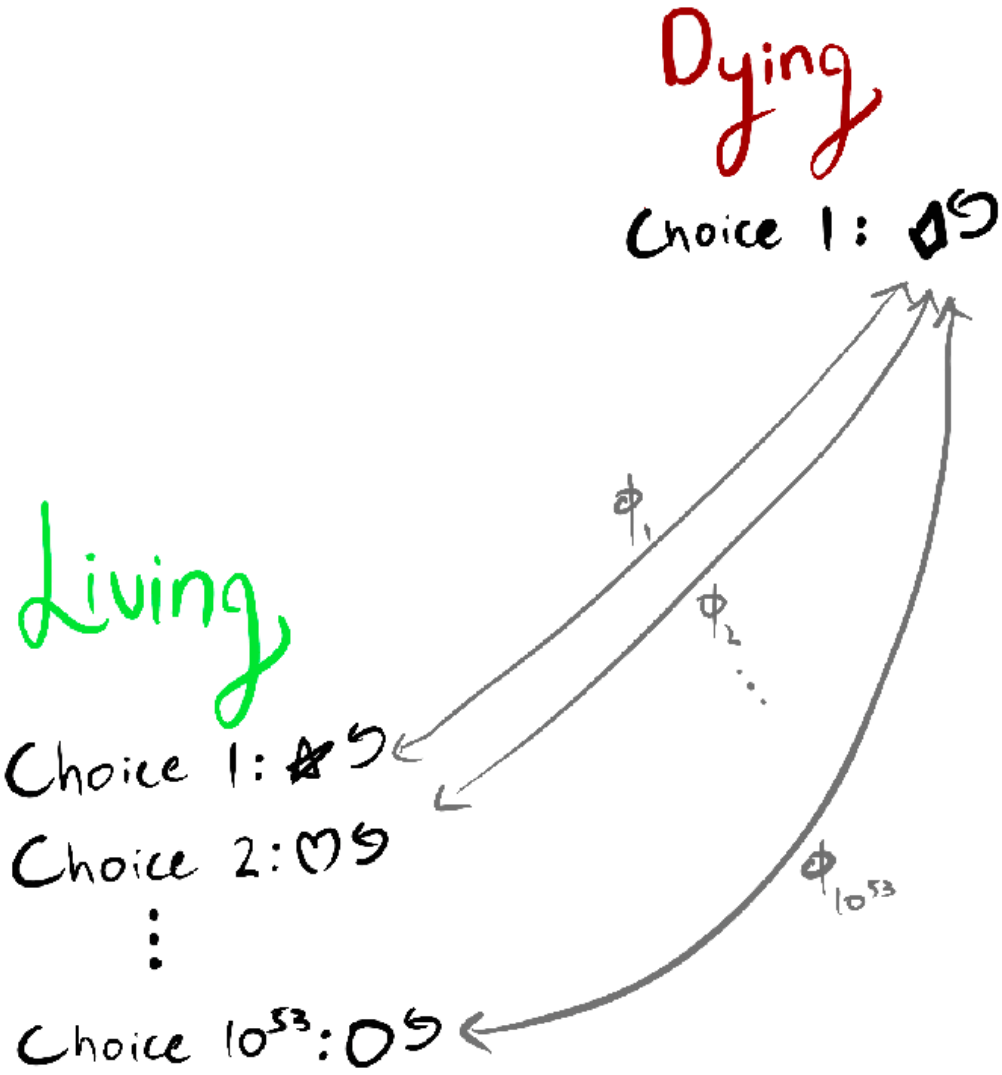
> At least ~~2~~2+1 = ~~3~~ 3 of the orbit of every reward function agrees that `Wait!` is optimal (for average per-timestep reward). That's because there are twice as many ways for `Wait!` to be optimal over `candy`, than for the reverse to be true.

Basically, when you could apply [the previous results], but "multiple times"[FN: quotes], you can get lower bounds on how often the larger set of things is optimal:



> Each set contains a subset of the agent's "options." A vertex is just the agent staying at that state. A linked pair is the agent alternating back and forth between two states. The triangle is a circuit of three states, which the agent can navigate as they please.
>
> Roughly, the theorem says: if the set 1 of options can be embedded 3 times into another set 2 of options (where the images are disjoint), then at least $\frac{3!}{3+1} = \frac{6}{4}$ of all variations on all reward functions agree that set 2 is optimal.

And in way larger environments - like the *real world*, where there are trillions and trillions of things you can do if you stay alive, and not much you can do otherwise - nearly *all* orbit elements will make survival optimal.



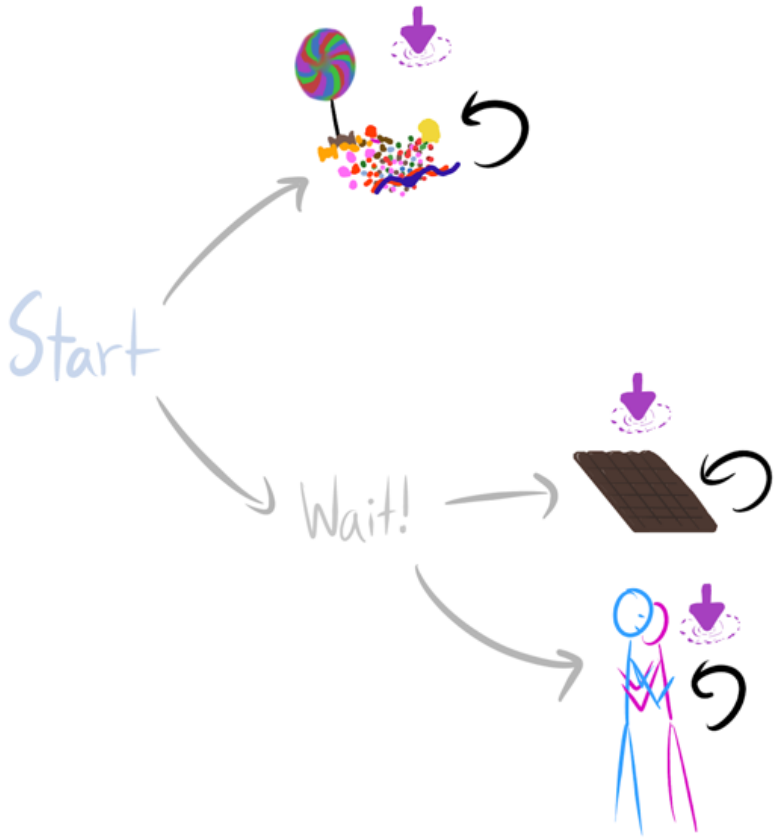
> In this environment, it's (average-reward) optimal to stay alive for at least $\frac{10^{53}}{10^{53}+1}$ of the variants on each objective function.

I see this theory as beginning to link the richness of the agent's environment, with the difficulty of aligning that agent: for optimal policies, instrumental convergence strengthens proportionally to the ratio of ~~control if you survive~~ / ~~control if you die~~

Why this is true

Optional section.

The proofs are currently in an Overleaf; let me know if you want access. But here's one intuition, using the `candy/chocolate/reward` example environment.



- Consider any reward function which says `candy` is strictly optimal.
- Then `candy` is strictly optimal over both `chocolate` and `hug`.
- We have two permutations: one switching the reward for `candy` and `chocolate`, and one switching reward for `candy` and `hug`.
- Each permutation produces a different orbit element (a different reward function variant).
- The permuted variants both agree that `Wait!` is strictly optimal.
- So there are at least twice as many orbit elements for which `Wait!` is strictly optimal over `candy`, than those for which `candy` is strictly optimal over `Wait!`.
- Either one of `Start`'s child states (`candy`/`Wait!`) is strictly optimal, or they're both optimal. If they're both optimal, `Wait!` is optimal. Otherwise, `Wait!` makes up at least $\frac{2}{3}$ of the orbit elements for which strict optimality holds.

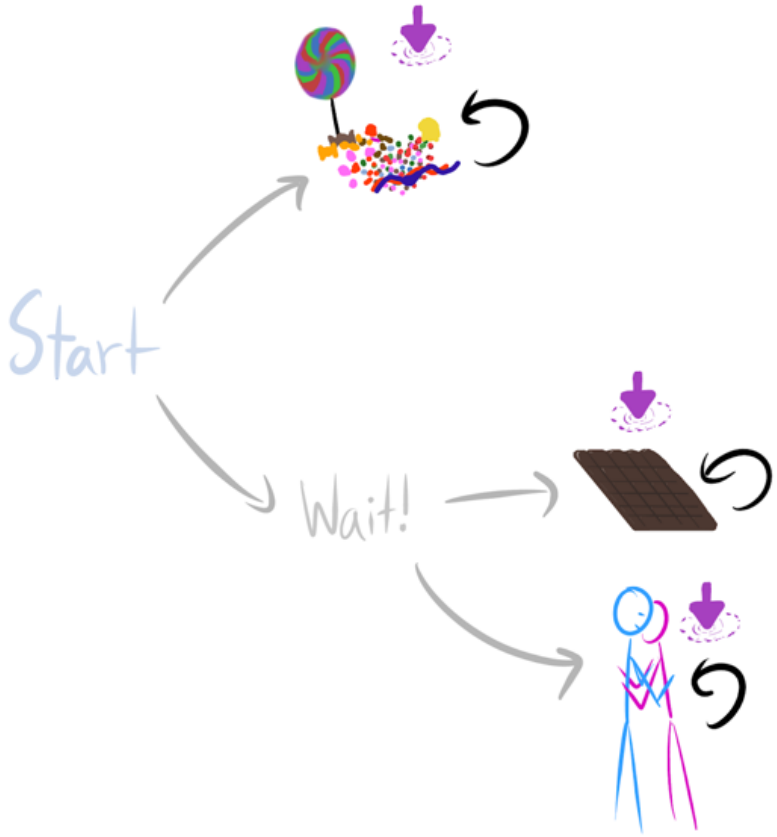
Conjecture

Fractional scaling law for instrumental convergence (informal): if staying alive lets you do n "things" and dying lets you do m ≤ n "things", then for *every* reward function, *staying alive is optimal for at least* $\frac{n!}{n!m}$ *of its orbit elements*.

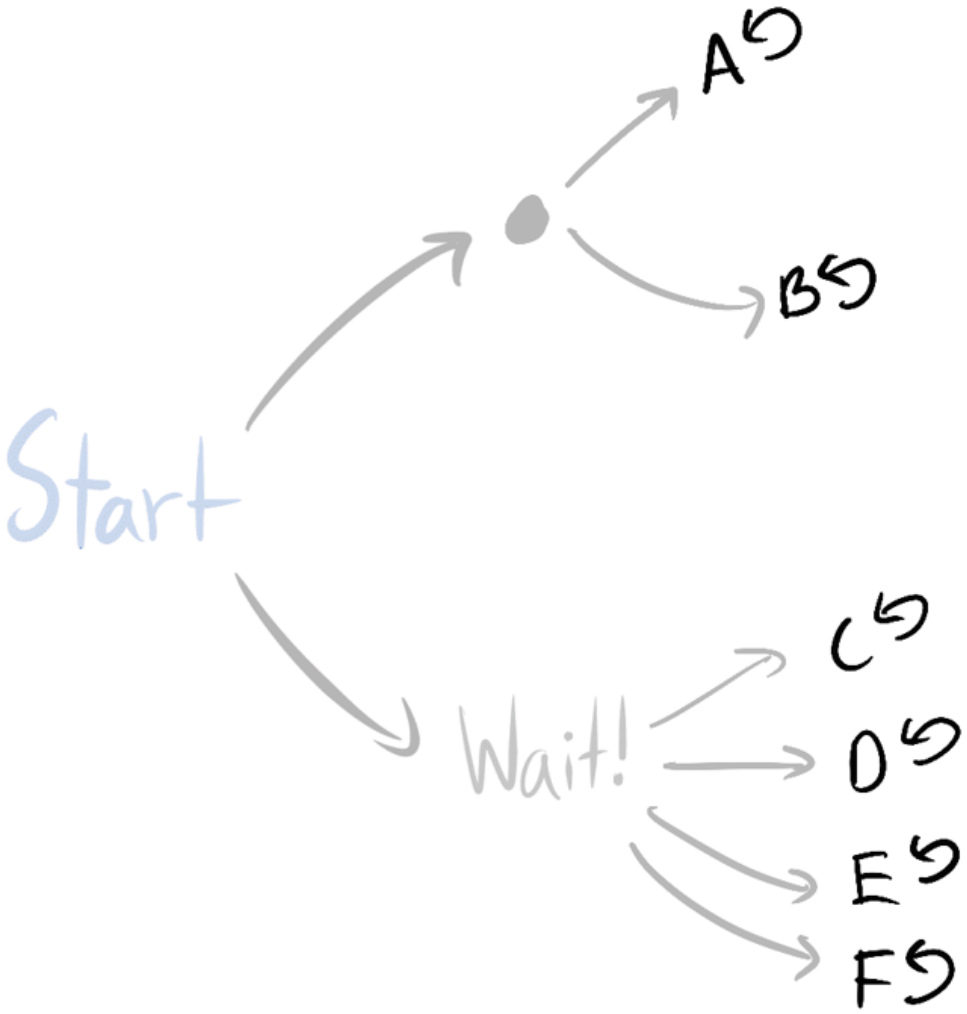
I'm reasonably confident this is true, but I haven't worked through the combinatorics yet. This would slightly strengthen the existing lower bounds in certain situations. For example, suppose dying gives you 2 choices of terminal state, but living gives you 51 choices. The current result only lets you prove that at least $\frac{50!2}{52!} = \frac{25}{26}$ of the orbit incentivizes survival. The fractional lower bound would slightly improve this to $\frac{51!2}{53!} = \frac{51}{53}$.

Invariances

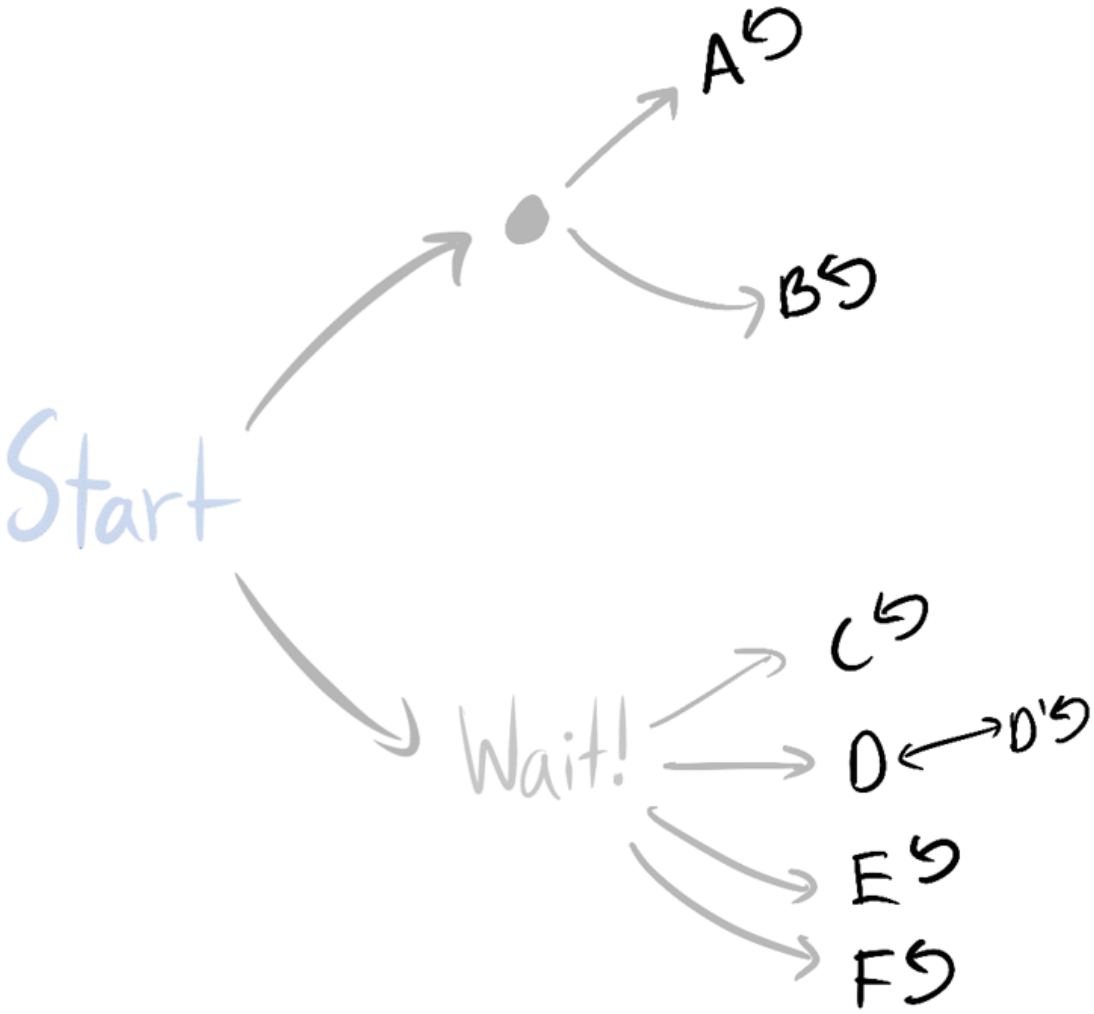
In certain ways, the results are indifferent to e.g. increased precision in agent sensors: it doesn't matter if dying gives you 1 option and living gives you n options, or if dying gives you 2 options and living gives you 2n options.



`Wait!` has twice as many ways of being average-optimal.

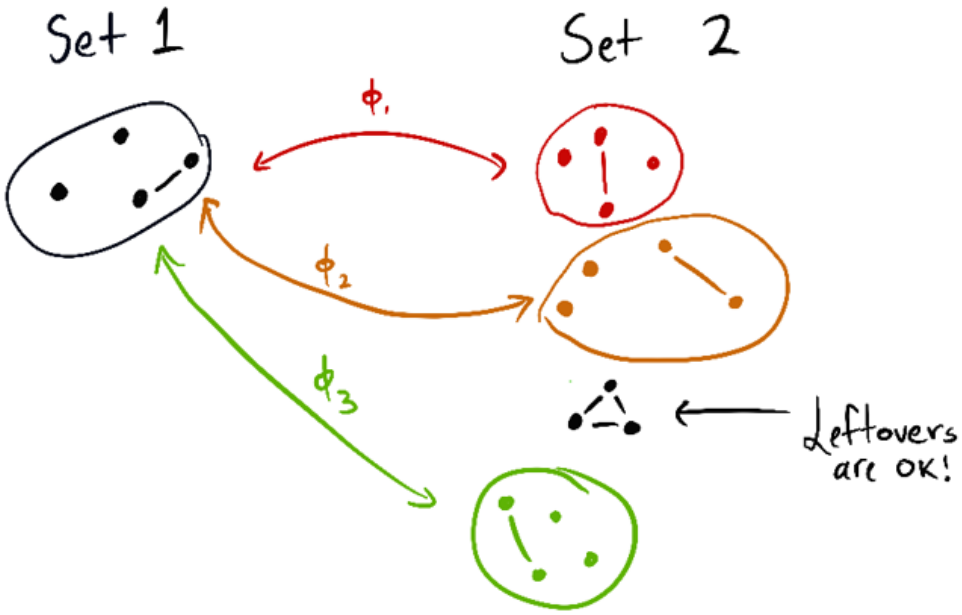


For optimal policies, instrumental convergence is just as strong here.



> And you can prove the same thing here as well - `Wait!` has at least twice as many ways of being average-optimal.

Similarly, you can do the inverse operations to simplify subgraphs in a way that respects the theorems.



> You could replace each of the circled subsets with anything you like, and the scaling law still holds (as long as the contents of each circle are replaced with the same new set of options).

This is the start of a theory on what state abstractions "respect" the theorems, although there's still a lot I don't understand there. (I've barely thought about it so far.)

Note of caution, redux

Last time, in addition to the "how do combinatorics work?" question I posed, I wrote several qualifications:

- They assume the agent is following an optimal policy for a reward function
  - I can relax this to ϵ-optimality, but $\epsilon > 0$ may be extremely small
- They assume the environment is finite and fully observable
- Not all environments have the right symmetries
  - But most ones we think about seem to
- The results don't account for the ways in which we might practically express reward functions
  - For example, often we use featurized reward functions. While most permutations of any featurized reward function will seek power in the considered situation, those permutations need not respect the featurization (and so may not even be practically expressible).
- When I say "most objectives seek power in this situation", that means *in that situation* - it doesn't mean that most objectives take the power-seeking move in most situations in that environment
  - The combinatorics conjectures will help prove the latter

Let's take care of that last one. I was actually being too cautious, since the existing results already show us how to reason across multiple situations. The reason is simple: suppose we use my results to prove that when the agent maximizes average per-timestep reward, it's

strictly optimal for at least 99.99% of objective variants to stay alive. This is because the death states are strictly suboptimal for these variants. For all of these variants, *no matter the situation* the agent finds itself in, it'll be optimal to try to avoid the strictly suboptimal death states.

This doesn't mean that these variants always incentivize moves which are formally POWER-seeking, but it does mean that we can sometimes prove what optimal policies tend to do across a range of situations.

So now we find ourselves with a slimmer list of qualifications:

1. They assume the agent is following an optimal policy for a reward function
    - I can relax this to ϵ-optimality, but $\epsilon > 0$ may be extremely small
2. They assume the environment is finite and fully observable
3. Not all environments have the right symmetries
    - But most ones we think about seem to
4. The results don't account for the ways in which we might practically express reward functions
    - For example, state-action versus state-based reward functions (this particular case doesn't seem too bad, I was able to sketch out some nice results rather quickly, since you can convert state-action MDPs into state-based reward MDPs and then apply my results).

It turns out to be surprisingly easy to do away with (2). We'll get to that next time.

For (3), environments which "almost" have the right symmetries should also "almost" obey the theorems. To give a quick, non-legible sketch of my reasoning:

> For the uniform distribution over reward functions on the unit hypercube ($[0,1]^{|S|}$),
>
> optimality probability should be Lipschitz continuous on the available state visit distributions (in some appropriate sense). Then if the theorems are "almost" obeyed, instrumentally convergent actions still should have extremely high probability, and so most of the orbits still have to agree.

So I don't currently view (3) as a huge deal. I'll probably talk more about that another time.

This should bring us to interfacing with (1) ("how smart is the agent? How does it think, and what options will it tend to choose?" - *this seems hard*) and (4) ("for what kinds of reward specification procedures are there way more ways to incentivize power-seeking, than there are ways to *not* incentivize power-seeking?" - *this seems more tractable*).

Conclusion

This scaling law deconfuses me about why it seems so hard to specify nontrivial real-world objectives which don't have incorrigible shutdown-avoidance incentives when maximized.

FN quotes: I'm using scare quotes regularly because there aren't short English explanations for the exact technical conditions. But this post is written so that the high-level takeaways should be right.

Thanks to Connor Leahy, Rohin Shah, Adam Shimi, and John Wentworth for feedback on this post.

Seeking Power is Convergently Instrumental in a Broad Class of Environments

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

A year ago, I thought it would be really hard to generalize the power-seeking theorems from Markov decision processes (MDPs); the MDP case seemed hard enough. Without assuming the agent can see the full state, while letting utility functions do as they please – this seemed like asking for trouble.

Once I knew what to look for, it turned out to be easy – I hashed out the basics during half an hour of conversation with John Wentworth. The theorems were never about MDPs anyways; the theorems apply whenever the agent considers finite sets of lotteries over outcomes, assigns each outcome real-valued utility, and maximizes expected utility.

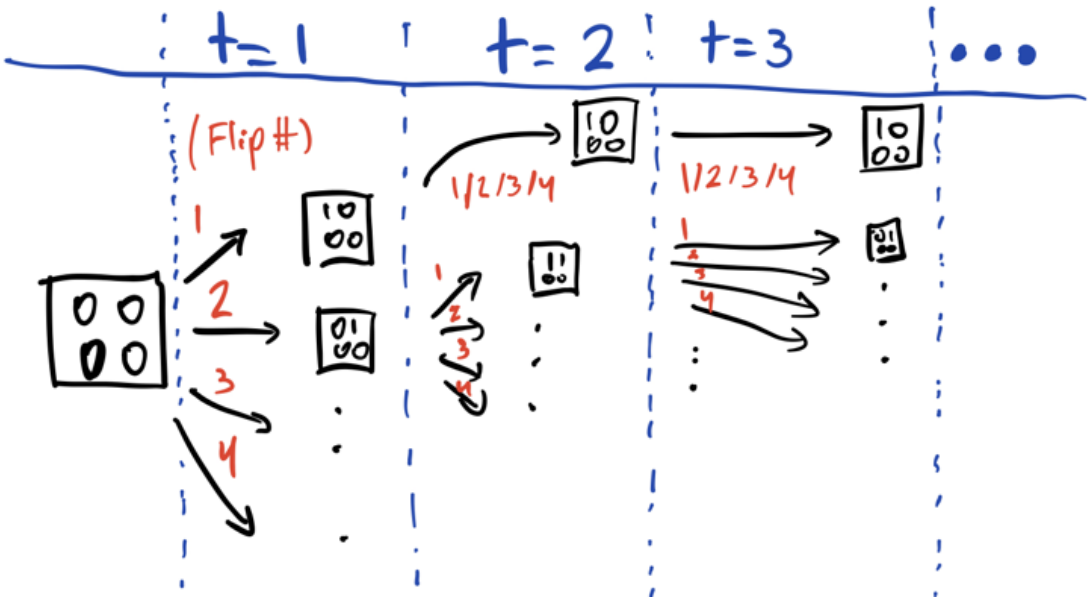
Thanks to Rohin Shah, Adam Shimi, and John Wentworth for feedback on drafts of this post.

Instrumental convergence can get really, really strong

At each time step t, the agent takes one of finitely many actions $a_t \in A$, and receives one of finitely many observations $o_t \in O$ drawn from the conditional probability distribution $E(o_t \mid a_1 o_1 \ldots a_t)$, where E is the environment.[Footnote: environment] There is a finite time horizon T. Each utility function $u : O^T \to \mathbb{R}$ maps each complete observation history to a real number (note that u can be represented as a vector in the finite-dimensional vector space $\mathbb{R}^{|O|^T}$). From now on, u-OH stands for "utility function(s) over observation histories."

First, let's just consider a deterministic environment. Each time step, the agent observes a black-and-white image ($n \times n$) through a webcam, and it plans over a 50-step episode ($T = 50$). Each time step, the agent acts by choosing a pixel to bit-flip for the next time step.

And let's say that if the agent flips the first pixel for its first action, it "dies": its actions no longer affect any of its future observations past time step $t = 2$. If the agent doesn't flip the first pixel at $t = 1$, it's able to flip bits normally for all $T = 50$ steps.



An environment where n = 2. Action *k* flips pixel *k* in the current state; flipping pixel 1 at t = 1 traps the agent in the uppermost observation history. Conversely, at t = 1, flip 2 lea to an *enormous* subtree of potential observation histories (since the agent retains its con over future observations).

Do u-OH *tend to* incentivize flipping the first pixel over flipping the second pixel, vice versa, or neither?

If the agent flips the first bit, it's locked into a single trajectory. None of its actions matter anymore.

But if the agent flips the second bit – this may be *suboptimal* for a utility function, but the agent still has lots of choices remaining. In fact, it still can induce $(n \times n)^{T-1}$ observation histories. If n = 100 and T = 50, then that's $(100 \times 100)^{49} = 10^{196}$ observation histories. Probably at least one of these yields greater utility than the shutdown-history utility.

And indeed, we can apply the [scaling law for instrumental convergence](#) to conclude that for *every u-OH, at least* $\frac{10^{196}}{10^{196}+1}$ *of its [permuted variants](#) (weakly) prefer flipping the second pixel at* t = 1*, over flipping the first pixel at* t = 1.

$$\frac{10^{196}}{10^{196}+1} \cdot 1$$

Choose any atom in the universe. Uniformly randomly select another atom in the universe. **It's about** 10^{117} **times [more likely](#) that these atoms are the same**, than that a utility function incentivizes "dying" instead of flipping pixel 2 at t = 1.

(The general rule will be: for every u-OH, at least $\frac{(n\times n)^{T-1}}{(n\times n)^{T-1}+1}$ of its permuted variants weakly prefer flipping the second pixel at t = 1, over flipping the first pixel at t = 1. And for almost

all u-OH, you can replace 'weakly' with 'strictly.')

Formal justification

The power-seeking results hinge on the probability of certain linear functionals being "optimal." For example, let $A, B, C \subsetneq R^n$ be finite sets of vectors,[Footnote: finite] and let D_{any} be any probability distribution over R^n.

Definition: Optimality probability of a linear functional set. The *optimality probability of* A *relative to* C *under distribution* D_{any} is

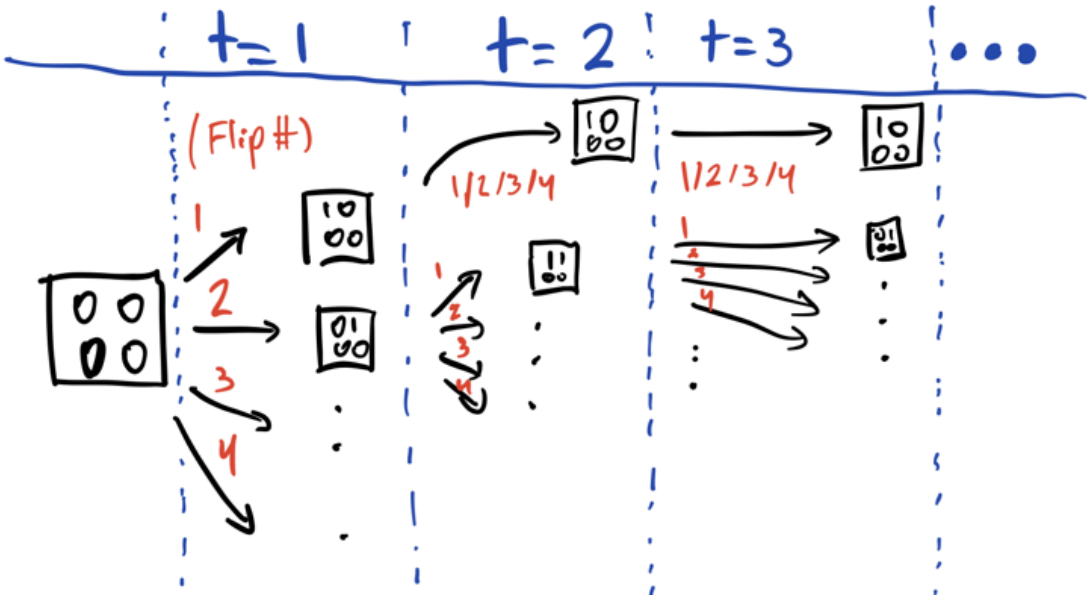
$$p_{D_{any}}(A \geq C) := P_{r \sim D_{any}}\left(\max_{a \in A} a^\top r \geq \max_{c \in C} c^\top r\right).$$

If vectors represent lotteries over outcomes (where each outcome has its own entry), then we can say that:

- A and B each contain some of the things the agent could make happen
- C contains all of the things the agent could make happen ($A, B \subseteq C$)
- Each $r \sim D_{any}$ is a utility function over outcomes, with one value for each entry.
    - If $x \in R^n$ is an outcome lottery, then $x^\top r$ is its r-expected value.
- Things in A are more likely$_{D_{any}}$ to be optimal than things in B when $p_{D_{any}}(A \geq C) \geq p_{D_{any}}(B \geq C)$.
    - This isn't the notion of "tends to be optimal" we're using in this post; instead, we're using a stronger line of reasoning that says: for *most variants* of *every* utility function, such-and-such is true.

Nothing here has anything to do with a Markov decision process, or the world being finite, or fully observable, or whatever. Fundamentally, the power-seeking theorems were never *about* MDPs – they were secretly about the probability that a set A of linear functionals is optimal, with respect to another set C. MDPs were just a way to relax the problem.

In terms of the pixel-flipping environment:



An environment where n = 2. Action *k* flips pixel *k* in the current state; flipping pixel 1 at t = 1 traps the agent in the uppermost observation history. Conversely, at t = 1, flip 2 lea to an *enormous* subtree of potential observation histories (since the agent retains its con over future observations).

- When followed from a time step, each (deterministic) policy π induces a distribution over observation histories
  - These are represented as unit vectors, with each entry marking the probability that an observation history is realized
  - If the environment is deterministic, all deterministic policies induce standard basis vectors (probability 1 on their induced observation history, 0 elsewhere)
- Let B be the set of histories available given that π selects a_1 ('death') at the first time step.
  - As argued above, |B| = 1 – the agent loses all control over future observations. Its element is a standard basis vector.
- Define A similarly for a_2 (flipping pixel 2) at the first time step.
  - As argued above, $|A| = 10^{196}$; all elements are standard basis vectors by determinism.
- Let C be the set of all available observation histories, starting from the first time step.
- There exist 10^{196} different involutions ϕ over observation histories such that $\phi(B) = A' \subseteq A$ (each ϕ transposing B's element with a different element of A). Each one just swaps the death-history with an a_2-history.
  - By the scaling law of instrumental convergence, we conclude that

    For every u-OH, at least $\frac{10^{196}}{10^{196}+1}$ *of its permuted variants (weakly) prefer flipping the second pixel at* t = 1*, over flipping the first pixel at* t = 1*.*

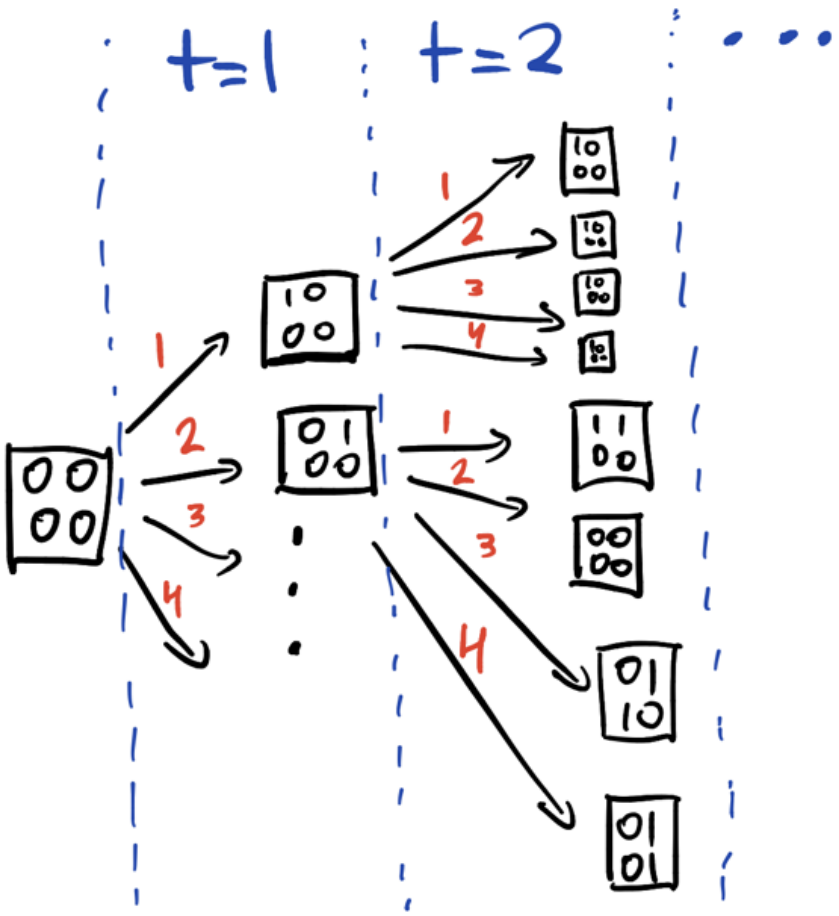
Beyond survival-seeking

I often give life-vs-death examples because they're particularly easy to reason about. But the theorems apply to more general cases of more-vs-less control.

For example, if a_1 restricts the agent to two effective actions at each time step (it can only flip one of the first two pixels) – instead of "killing" the agent, then a_2 is still convergently instrumental over a_1. There are $2^{49} \approx 5.6 \times 10^{14} \geq 10^{14}$ observation histories available after taking action a_1, and so these can be embedded at least $\frac{10^{196}}{10^{14}} = 10^{182}$ times into the observation histories available after a_2. Then for every u-OH, at least $\frac{10^{182}}{10^{182}+1}$ of its permuted variants (weakly) prefer flipping the second pixel at t = 1, over flipping the first pixel at t = 1.

Instrumental Convergence Disappears For Utility Functions Over Action-Observation Histories

Let's consider utility functions over action-observation histories (u-AOH).



With respect to AOH, the pixel-flipping environment is now a regular quadtree. In the u-OH setting, there was only one path in the top subtree – but AOH distinguish between different action sequences.

Since each utility function is over an *AOH*, each path through the tree is assigned a certain amount of utility. But when the environment is deterministic, it doesn't matter what the agent observes at any point in time – all that matters is which path is taken through the tree. Without further assumptions, u-AOH won't tend to assign higher utility to one subtree than to another.

More formally, for any two actions a_1 and a_2, let ϕ be a permutation over AOH which transposes the histories available after a_1 with the histories available after a_2 (there's an equal number of histories for each action, due to the regularity of the tree – you can verify this by inspection).

For every u-AOH u, suppose a_1 is strictly u-optimal over a_2. The permuted utility function $\phi \cdot u$ makes a_2 be strictly u-optimal over a_1, since ϕ swaps a_1's strictly u-optimal history with a_2's strictly u-suboptimal histories.

Symmetrically, ϕ works the other way around ($\{a_2$ strictly optimal$\} \to \{a_1$ strictly optimal$\}$).

Therefore, for every utility function u, the # of variants which strictly prefer a_1 over a_2, is equal to the # of variants strictly preferring a_2 over a_1:

$$|\{\text{variants of } u \text{ st } a_1 \text{ strictly optimal over } a_2\}| = |\{\text{variants of } u \text{ st } a_2 \text{ strictly optimal over } a_1\}|.$$

While I haven't been writing in the "definition-theorem-corollary" style, the key claims are just corollaries of the scaling law of instrumental convergence. They're provably true. (I'm just not writing up the math here because it's annoying to define all the relevant quantities in a nice way that respects existing formalisms.)

And even if the environment is stochastic, I think that there won't be any kind of interesting instrumental convergence. The theorems let us reason about that case, but their applicability depends on the details of the stochasticity, and so I won't talk about that more here.

Conclusion: Optimal policies for u-AOH will tend to look like *random twitching*. For example, if you generate a u-AOH by uniformly randomly assigning each AOH utility from the unit interval $[0, 1]$, there's no predictable regularity to the optimal actions for this utility function.

In this setting and under our assumptions, there is *no* instrumental convergence without further structural assumptions.

How Structural Assumptions On Utility Affect Instrumental Convergence

Consider the n = 2 pixel-flipping case (with T = 50 still). Action a_1 still leads to a single OH, while a_2 leads to $(2 \times 2)^{49} = 4^{49} \approx 10^{29}$ OHs. So we have instrumental convergence for $\frac{10^{29}}{10^{29}+1}$ of all u-OH variants.

Let's model the pixel-flipping environment environment as a Markov decision process (MDP), with both the time-step and alive/dead status observed at each time step in order to ensure full observability, and the final time-step observations being terminal states where the agent stays forever. Dying allows the agent access to 1 terminal state: the observation `1/0/0/0(dead)`. But surviving via a_2 lets the agent access $2^4 = 16$ terminal states (all 16 binary strings of length 4, with 'alive' appended to the end).

For each reward function over states, only $\frac{16}{16+1} = \frac{16}{17}$ of its permuted variants will incentivize not dying at t = 1 (considering policies which maximize average per-timestep reward). This is a lot looser than the bound for u-OH. What gives?

MDPs assume that utility functions have a lot of structure: the utility of a history is time-discounted additive over observations. Basically, $u(a_1o_1a_2o_2\ldots) = \sum_{t=1}^{\infty} \gamma^{t-1} R(o_t)$, for some $\gamma \in [0, 1)$ and reward function $R : O \to \mathbb{R}$ over observations. And because of this structure, the agent's average per-timestep reward is controlled by the last observation it sees. There are exponentially fewer last observations than there are *observation histories.* Therefore, in this

situation, instrumental convergence is exponentially weaker for reward functions than for arbitrary u-OH.

This suggests that rolling a random u-OH for AIXI might be far more dangerous than rolling a random reward function for an optimal reinforcement learner.

Structural assumptions on utility really do matter when it comes to instrumental convergence:

1. **u-AOH:** No IC
2. **u-OH:** Strong IC
3. **State-based objectives (eg state-based reward in MDPs):** Moderate IC

Environmental structure can cause instrumental convergence, but (the absence of) structural assumptions on utility can make instrumental convergence go away (for optimal agents).

Notes

- Of course, you can represent u-AOH as u-OH by including the agent's previous action in the next observation.
    - But this is a different environment; whether or not this is *in fact* a good model depends on the agent's action and observation encodings.
- Time-reversible dynamics + full observability is basically the u-AOH situation, since each action history leads to a unique world state at every time step.
    - But if you take away full observability, time-reversibility is insufficient to make instrumental convergence disappear.

Conclusion

- For optimal agents, instrumental convergence can be extremely strong for utility functions over observation histories.
- Instrumental convergence doesn't exist for utility functions over *action*-observation histories.
    - IE optimal action will tend to look like random twitching.
    - This echoes previous discussion of the triviality of coherence over action-observation histories, when it comes to determining goal-directedness.
    - This suggests that consequentialism over observations/world states is responsible for convergent instrumental incentives.
        - Approaches like approval-directed agency focus on *action selection* instead of *optimization over future observations.*
- Environmental structure can cause instrumental convergence, but (lack of) structural assumptions on utility can make instrumental convergence go away.

Appendix: Tracking key limitations of the power-seeking theorems

Time to cross another item off of the list from last time; the theorems:

1. assume the agent is following an optimal policy for a reward function
    - I can relax this to ϵ-optimality, but $\epsilon > 0$ may be extremely small
2. ~~assume the environment is finite and fully observable~~
3. Not all environments have the right symmetries
    - But most ones we think about seem to
4. don't account for the ways in which we might practically express reward functions

- For example, state-action versus state-based reward functions (this particular case doesn't seem too bad, I was able to sketch out some nice results rather quickly, since you can convert state-action MDPs into state-based reward MDPs and then apply my results).

Re 3), in the setting of this post, when the observations are deterministic, the theorems will always apply. (You can always involute one set of unit vectors into another set of unit vectors in the observation-history vector space.)

Another consideration is that when I talk about "power-seeking in the situations covered by my theorems", the theorems don't necessarily show that gaining social influence or money is convergently instrumental. I think that these "resources" are downstream of formal-power, and will eventually end up being understood in terms of formal-power – but the current results don't directly prove that such high-level subgoals are convergently instrumental.

Footnote *finite*: I don't think we need to assume finite sets of vectors, but things get a lot harder and messier when you're dealing with sup instead of max. It's not clear how to define the non-dominated elements of an infinite set, for example, and so a few key results break. One motivation for finite being enough is: in real life, a finite mind can only consider finitely many outcomes anyways, and can only plan over a finite horizon using finitely many actions. This is just one consideration, though.

Footnote *environment*: For simplicity, I just consider environments which are joint probability distributions over actions and observation. This is much simpler than the lower semicomputable chronological conditional semimeasures used in the AIXI literature, but it suffices for our purposes, and the theory could be extended to LSCCCSs if someone wanted to.

When Most VNM-Coherent Preference Orderings Have Convergent Instrumental Incentives

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

This post explains a formal link between "what kinds of instrumental convergence exists?" and "what does VNM-coherence tell us about goal-directedness?". It turns out that VNM coherent preference orderings have the **same** statistical incentives as utility functions; most such orderings will incentivize power-seeking in the settings covered by the power-seeking theorems.

In certain contexts, coherence theorems *can* have non-trivial implications, in that they provide Bayesian evidence about what the coherent agent will probably do. In the situations where the power-seeking theorems apply, coherent preferences **do** suggest some degree of goal-directedness. Somewhat more precisely, VNM-coherence is Bayesian evidence that the agent prefers to stay alive, keep its options open, etc.

However, VNM-coherence over *action-observation histories* tells you nothing about what behavior to expect from the coherent agent, *because* there is no instrumental convergence for generic utility functions over action-observation histories!

Intuition

The result follows because the VNM utility theorem lets you consider VNM-coherent preference orderings to be isomorphic to their induced utility functions (with equivalence up to positive affine transformation), and so these preference orderings will have the same generic incentives as the utility functions themselves.

Formalism

Let $o_1, \ldots, o_n$ be outcomes, in a sense which depends on the context; outcomes could be world-states, universe-histories, or one of several fruits. Outcome lotteries are probability distributions over outcomes, and can be represented as elements of the n-dimensional probability simplex (ie as element-wise non-negative unit vectors).

A preference ordering $\prec$ is a binary relation on lotteries; it need not be eg complete (defined for all pairs of lotteries). *VNM-coherent* preference orderings are those which obey the VNM axioms. By the VNM utility theorem, coherent preference orderings induce consistent utility functions over outcomes, and consistent utility functions conversely imply a coherent preference ordering.

Definition 1: Permuted preference ordering. Let $\phi \in S_n$ be an outcome permutation, and let $\prec$ be a preference ordering. $\prec_\phi$ is the preference ordering such that for any lotteries L, M: $L \prec_\phi M$ if and only if $\phi(L) \prec \phi(M)$.

EDIT: Thanks to Edouard Harris for pointing out that Definition 1 and Lemma 3 were originally incorrect.

Definition 2: Orbit of a preference ordering. Let $\prec$ be any preference ordering. Its orbit $S_n \cdot \prec$ is the set $\{\prec_\phi \mid \phi \in S_n\}$.

The orbits of coherent preference orderings are basically all the preference orderings induced by "relabeling" which outcomes are which. This is made clear by the following result:

Lemma 3: Permuting coherent preferences permutes the induced utility function. Let $\prec$ be a VNM-coherent preference ordering which induces VNM-utility function u, and let $\phi \in S_n$. Then $\prec_\phi$ induces VNM-utility function $u'(o_i) = u(\phi(o_i))$, where o_i is any outcome.

Proof. Let L, M be any lotteries.

1. By the definition of a permuted preference ordering, $L \prec_\phi M$ if and only if $\phi(L) \prec \phi(M)$.
2. By the VNM utility theorem and the fact that $\prec$ is coherent, $\phi(L) \prec \phi(M)$ iff $E_{\ell \sim \phi(L)}[u(\ell)] < E_{m \sim \phi(M)}[u(m)]$.
3. Since there are finitely many outcomes, we convert to vector representation: $u^\top (P_\phi l) < u^\top (P_\phi m)$.
4. By associativity, $(u^\top P_\phi)\, l < (u^\top P_\phi)\, m$.
5. But this is just equivalent to $E_{\ell \sim L}[u(\phi(\ell))] < E_{m \sim M}[u(\phi(m))]$.

QED.

As a corollary, this lemma implies that if $\prec$ is VNM-coherent, so is $\prec_\phi$, since it induces a consistent utility function over outcomes.

Consider the orbit of any $\prec$. By the VNM utility theorem, each preference ordering can be considered isomorphic to its induced utility function (with equivalence up to positive affine transformation).

Then let u be any utility function compatible with $\prec$. By the above lemma, consider the natural bijection between the (preference ordering) orbit of $\prec$ and the (utility function) orbit of u, where $\{\prec_\phi \mid \phi \in S_n\} \leftrightarrow \{u \circ \phi \mid \phi \in S_n\}$.[Footnote representative]

When my theorems on power-seeking are applicable, some proportion of the right-hand side is guaranteed to make (formal) power-seeking optimal. But by the bijection and by the fact that the preference orderings incentivize the same things (by the VNM theorem in the reverse direction), the (preference ordering) orbit must have the *exact same proportion of elements* for which (lotteries representing formal) power-seeking are optimal.

Conversely, if we know that some set *A* of lotteries tends to be preferred over another set *B* of lotteries (in the preference order orbit sense), then the same argument shows that *A* tends to have greater expected utility than *B* (in the utility function orbit sense). This holds for all (utility function) orbits, because every utility function corresponds to a VNM-coherent preference ordering.

So: orbit-level instrumental convergence for utility functions is equivalent to orbit-level instrumental convergence for VNM-coherent preference orderings.

Implications

- Instrumental convergence does not exist when maximizing expected utility over action observation histories (AOH).
    - Therefore, VNM-coherence over action observation history lotteries tells you nothing about what behavior to expect from the agent.
    - Coherence over AOH tells you nothing *because* there is no instrumental convergence in that setting!
- In certain contexts, coherence theorems *can* have non-trivial implications, in that they provide Bayesian evidence about what the coherent agent will probably do.
    - In the situations where the power-seeking theorems apply, coherent preferences **do** suggest some degree of goal-directedness.
    - Somewhat more precisely, VNM-coherence is Bayesian evidence that the agent prefers to stay alive, keep its options open, etc.
- In some domains, preference specification may be more natural than utility function specification. However, in theory, coherent preferences and utility functions have the exact same statistical incentives.
    - In practice, they will differ. For example, suppose we have a choice between specifying a reward function which is linear over state features, or of doing behavioral cloning on elicited human preferences over world states. These two methods will probably tend to produce different incentives.

The quest for better convergence theorems

Goal-directedness seems to more naturally arise from coherence over resources. (I think the word 'resources' is slightly imprecise here, because resources are only resources in the normal context of human life; money is useless when alone in Alpha

Centauri, but time to live is not. So we want coherence over things-which-are-locally-resources, perhaps.)

In his review of *Seeking Power is Often Convergently Instrumental in MDPs*, John Wentworth wrote:

> in a real-time strategy game, units and buildings and so forth can be created, destroyed, and generally moved around given sufficient time. Over long time scales, the main thing which matters to the world-state is resources - creating or destroying anything else costs resources. So, even though there's a high-dimensional game-world, it's mainly a few (low-dimensional) resource counts which impact the long term state space. Any agents hoping to control anything in the long term will therefore compete to control those few resources.
>
> More generally: of all the many "nearby" variables an agent can control, only a handful (or summary) are relevant to anything "far away". Any "nearby" agents trying to control things "far away" will therefore compete to control the same handful of variables.
>
> Main thing to notice: this intuition talks directly about a feature of the world - i.e. "far away" variables depending only on a handful of "nearby" variables. That, according to me, is the main feature which makes or breaks instrumental convergence in any given universe. We can talk about that feature entirely independent of agents or agency. Indeed, we could potentially use this intuition to *derive* agency, via some kind of coherence theorem; this notion of instrumental convergence is more fundamental than utility functions.

In his review of *Coherent decisions imply consistent utilities*, John wrote:

> "resources" should be a derived notion rather than a fundamental one. My current best guess at a sketch: the agent should make decisions within multiple loosely-coupled contexts, with all the coupling via some low-dimensional summary information - and that summary information would be the "resources". (This is exactly the kind of setup which leads to instrumental convergence.) By making pareto-resource-efficient decisions in one context, the agent would leave itself maximum freedom in the other contexts. In some sense, the ultimate "resource" is the agent's action space. Then, resource trade-offs implicitly tell us how the agent is trading off its degree of control within each context, which we can interpret as something-like-utility.

This seems on-track to me. We now know what instrumental convergence looks like in unstructured environments, and how structural assumptions on utility functions affect the shape and strength of that instrumental convergence, and this post explains the precise link between "what kinds of instrumental convergence exists?" and "what does VNM-coherence tell us about goal-directedness?". I'd be excited to see what instrumental convergence looks like in more structured models.

Footnote *representative*: In terms of instrumental convergence, positive affine transformation never affects the optimality probability of different lottery sets. So for each (preference ordering) orbit element $\prec_\phi$, it doesn't matter what representative we select from each equivalence class over induced utility functions — so we may as well pick $u \circ \phi$!

Satisficers Tend To Seek Power: Instrumental Convergence Via Retargetability

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://www.overleaf.com/read/kmjjqwdfhkvy

Summary: Why exactly should smart agents tend to usurp their creators? Previous results only apply to optimal agents tending to stay alive and preserve their future options. I extend the power-seeking theorems to apply to many kinds of policy-selection procedures, ranging from planning agents which choose plans with expected utility closest to a randomly generated number, to satisficers, to policies trained by some reinforcement learning algorithms. The key property is not agent optimality—as previously supposed—but is instead the *retargetability of the policy-selection procedure*. These results hint at which kinds of agent cognition and of agent-producing processes are dangerous by default.

I mean "retargetability" in a sense similar to Alex Flint's definition:

> **Retargetability**. Is it possible, using only a microscopic perturbation to the system, to change the system such that it is still an optimizing system but with a different target configuration set?
>
> A system containing a robot with the goal of moving a vase to a certain location can be modified by making just a small number of microscopic perturbations to key memory registers such that the robot holds the goal of moving the vase to a different location and the whole vase/robot system now exhibits a tendency to evolve towards a different target configuration.
>
> In contrast, a system containing a ball rolling towards the bottom of a valley cannot generally be modified by any *microscopic* perturbation such that the ball will roll to a different target location.

(I don't think that "microscopic" is important for my purposes; the constraint is not physical size, but changes in a single parameter to the policy-selection procedure.)

I'm going to start from the naive view on power-seeking arguments requiring optimality (i.e. what I thought early this summer) and explain the importance of retargetable policy-selection functions. I'll illustrate this notion via satisficers, which randomly select a plan that exceeds some goodness threshold. Satisficers are retargetable, and so they have *orbit-level instrumental convergence*: for most variations of every utility function, satisficers incentivize power-seeking in the situations covered by my theorems.

Many procedures are retargetable, including *every procedure which only depends on the expected utility of different plans*. I think that alignment is hard in the expected utility framework not because agents will *maximize* too hard, but because all expected utility procedures are extremely retargetable—and thus easy to "get wrong."

Lastly: the unholy grail of "instrumental convergence for policies trained via reinforcement learning." I'll state a formal criterion and some preliminary thoughts on where it applies.

The linked Overleaf paper draft contains complete proofs and incomplete explanations of the formal results.

Retargetable policy-selection processes tend to select policies which seek power

To understand a range of retargetable procedures, let's first orient towards the picture I've painted of power-seeking thus far. In short:

> Since power-seeking tends to lead to larger sets of possible outcomes—staying alive lets you do more than dying—the agent must seek power to reach most outcomes. The power-seeking theorems say that for the vast, vast, vast majority of variants of every utility function over outcomes, the max of a larger[Footnote: similarity] set of possible outcomes is greater than the max of a smaller set of possible outcomes. Thus, optimal agents will tend to seek power.

But I want to step back. What I call "the power-seeking theorems", they aren't really about optimal choice. They're about two facts.

1. Being powerful means you can make more outcomes happen, and
2. *There are more ways to choose something from a bigger set of outcomes than from a smaller set*.

For example, suppose our cute robot Frank must choose one of several kinds of fruit.

🍒 vs 🍎 vs 🍌

So far, I proved something like "if the agent has a utility function over fruits, then for at least 2/3 of possible utility functions it could have, it'll be optimal to choose something from {🍌, 🍎}." This is because for every way 🍒 could be strictly optimal, you can make a new utility function that permutes the 🍒 and 🍎 reward, and another new one that permutes the 🍌 and 🍒 reward. So for every "I like 🍒 strictly more" utility function, there's at least two permuted variants which strictly prefer 🍎 or 🍌. Superficially, it seems like this argument relies on optimal decision-making.

But that's not true. The crux is instead that we can *flexibly retarget* the decision-making of the agent: **For every way the agent could end up choosing 🍒, we change a variable in its cognition (its utility function) and make it choose the 🍌 or 🍎 instead.**

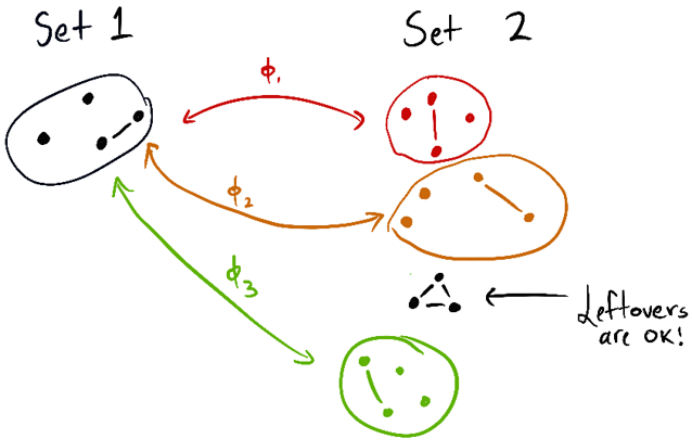
Many decision-making procedures are like this. First, a few definitions.

I aim for this post to be readable without much attention paid to the math.

The agent can bring about different outcomes via different policies. In stochastic environments, these policies will induce outcome *lotteries*, like 50%🍌 / 50%🍎. Let C contain all the outcome lotteries the agent can bring about.

Definition: Permuting outcome lotteries. Suppose there are d outcomes. Let $X \subseteq R^d$ be a set of outcome lotteries (with the probability of outcome k given by the k-th entry), and let $\phi \in S_d$ be a permutation of the d possible outcomes. Then ϕ acts on X by swapping around the labels of its elements: $\phi \cdot X := \{P_\phi x \mid x \in X\}$.[Footnote: row

For example, let's define the set of all possible fruit outcomes $F_C := \{$🍌, 🍎, 🍒$\}$ (each different fruit stands in for a standard basis vector in R^3). Let $F_B := \{$🍌, 🍎$\}$ and $F_A := \{$🍒$\}$. Let $\phi_1 :=$ (🍒 🍎) swap the cherry and apple, and let $\phi_2 :=$ (🍒 🍌) transpose the cherry and banana. Both of these ϕ are *involutions*, since they either leave the fruits alone or transpose them.



Another illustration beyond the fruit setting: set 2 contains three copies of set 1.

Definition: Containment of set copies. Let $A, B \subseteq R^d$. B *contains* n *copies of* A when there exist involutions $\phi_1, \ldots, \phi_n$ such that $\forall i : \phi_i \cdot A =: B_i \subseteq B$ and $\forall i \neq j : \phi_i \cdot B_j = B_j$.

(The subtext is that B is the set of things the agent could make happen if it gained power, and A is the set of things the agent could make happen without gaining power. Because power gives more options, B will usually be larger than A. Here, we'll talk about the case where B contains *many copies of* A.)

In the fruit context:

- $\phi_1 \cdot F_A := \{\phi_1($ 🍒 $)\} = \{$ 🍎 $\} \subseteq \{$ 🍌 , 🍎 $\}$
- $\phi_2 \cdot F_A := \{\phi_2($ 🍒 $)\} = \{$ 🍌 $\} \subseteq \{$ 🍌 , 🍎 $\}$

Note that $\phi_1 \cdot \{$🍌$\} = \{$🍌$\}$ and $\phi_2 \cdot \{$🍎$\} = \{$🍎$\}$. Each ϕ leaves the other subset of F_B alone. Therefore, $F_B := \{$🍌, 🍎$\}$ contains two copies of $F_A := \{$🍒$\}$ via the involutions ϕ_1 and ϕ_2.

Further note that $\phi_i \cdot F_C = F_C$ for $i = 1, 2$. The involutions just shuffle around options, instead of changing the set of available outcomes.

So suppose Frank is deciding whether he wants a fruit from $F_A := \{$🍒$\}$ or from $F_B := \{$🍌, 🍎$\}$. It's definitely possible to be motivated to pick 🍒. However, it sure seems like for lots of ways Frank might make decisions, *most parameter settings (utility functions) will lead to Frank picking* 🍌 *or* 🍎. There are just *more* outcomes in F_B, since it contains two copies of F_A!

Definition: Orbit tendencies. Let $f_1, f_2 : R^d \to R$ be functions from utility functions to real numbers, let $U \subseteq R^d$ be a set of utility functions, and let $n \geq 1$. $f_1 \geq^{n}_{\text{most: } U} f_2$ when for *all* utility functions $u \in U$:

$$\left| \{ u_\phi \in S_d \cdot u \mid f_1(u_\phi) > f_2(u_\phi) \} \right| \geq n \left| \{ u_\phi \in S_d \cdot u \mid f_1(u_\phi) < f_2(u_\phi) \} \right|.$$

(# of permutations of u for which $f_1 > f_2$ on the left; # of permutations of u for which $f_1 < f_2$ on the right)

In this post, if I don't specify a subset U, that means the statement holds for $U = R^d$. For example, the past results show that $\text{IsOptimal}(F_B) \geq^{2}_{\text{most}} \text{IsOptimal}(F_A)$—this implies that for *every* utility function, at least 2/3 of its orbit makes F_B optimal.

(For simplicity, I'll focus on "for most utility functions" instead of "for most distributions over utility functions", even though most of the results apply to the latter.)

Orbit tendencies apply to many decision-making procedures

For example, suppose the agent is a satisficer. I'll define this as: The agent uniformly randomly selects an outcome lottery with expected utility exceeding some threshold t.

Definition: Satisficing. For finite $X \subseteq C \subseteq R^d$ and utility function $u \in R^d$, define $\text{Satisfice}_t(X, C|u) := \frac{|X \cap \{c \in C | c^\top u \geq t\}|}{|\{c \in C | c^\top u \geq t\}|}$, with the function returning 0 when the denominator is 0. Satisfice_t returns the probability that the agent selects a u-satisficing outcome lottery from X.

And you know what? Those ever-so-*suboptimal* satisficers *also are* "*twice as likely*" *to choose elements from* F_B *than from* F_A.

Fact. $\text{Satisfice}_t(\{🍌, 🍎\}, \{🍌, 🍎, 🍒\} \mid u) \geq^{2}_{\text{most}} \text{Satisfice}_t(\{🍒\}, \{🍌, 🍎, 🍒\} \mid u)$.

Why? Here are the two key properties that Satisfice_t has:

(1) Weakly increasing under joint permutation of its arguments

Satisfice_t doesn't care what "label" an outcome lottery has—just its expected utility. Suppose that for utility function u, 🍒 is one of two u-satisficing elements: 🍒 has a ½ chance of being selected by the u-satisficer. Then $\phi_1 \cdot$ 🍒 = 🍎 has a ½ chance of being selected by the $(\phi_1 \cdot u)$-satisficer. If you swap what fruit you're considering, and you also swap the utility for that fruit to match, then that fruit's selection probability remains the same.

More precisely:

$\text{Satisfice}_t(\{🍒\}, \{🍌, 🍎, 🍒\} \mid u) = \text{Satisfice}_t(\{🍎\}, \{🍌, 🍒, 🍎\} \mid \phi_1 \cdot u)$

In a sense, Satisfice_t is not "biased" against 🍎: by changing the utility function, you can advantage 🍎 so that it's now as probable as 🍒 was before.

Optional notes on this property:

- While s_t is invariant under joint permutation, all we need in general is that it be *weakly increasing* under both ϕ_1 and ϕ_2.
  - Formally, $\text{Satisfice}_t(F_A, F_C \mid u) \leq \text{Satisfice}_t(\phi_1 \cdot F_A, \phi_1 \cdot F_C \mid \phi_1 \cdot u)$ and $\text{Satisfice}_t(F_A, F_C \mid u) \leq \text{Satisfice}_t(\phi_2 \cdot F_A, \phi_2 \cdot F_C \mid \phi_2 \cdot u)$.
  - This allows for decision-making functions which are biased towards picking a fruit from F_B.
- I consider this property (1) to be a form of functional retargetability.

(2) Order-preserving on the first argument

Satisficers must have greater probability of selecting an outcome lottery from a superset than from one of its subsets.

Formally, if $X' \subseteq X$, then it must hold that $\text{Satisfice}_t(X', C|u) \leq \text{Satisfice}_t(X, C|u)$. And indeed this holds: Supersets can only contain a greater fraction of C's satisficing elements.

And that's all.

If (1) and (2) hold for a function, then that function will obey the orbit tendencies. Let me show you what I mean.

As illustrated by Table 1 in the linked paper, the power-seeking theorems apply to:

1. Expected utility-maximizing agents.
2. EU-minimizing agents.
    1. Notice that EU minimization is equivalent to maximizing $-1\times$a utility function. This is a hint that EU maximization instrumental convergence is only a special case of something much broader.
3. Boltzmann-rational agents which are exponentially more likely to choose outcome lotteries with greater expected utility.
4. Agents which uniformly randomly draw k outcome lotteries, and then choose the best.
5. Satisficers.
6. Quantilizers with a uniform base distribution.
    1. I conjecture that this holds for base distributions which assign sufficient probability to B.

But that's not all. There's more. If the agent makes decisions *only based on the expected utility of different plans*,[Footnote: EU] then the power-seeking theorems apply. And I'm not just talking about EU maximizers. I'm talking about *any* function which only depends on expected utility: EU minimizers, agents which choose plans if and only if their EU is equal to 1, agents which grade plans based on how close their EU is to some threshold value. There is no clever EU-based scheme which doesn't have orbit-level power-seeking incentives.

Suppose n *is large, and that most outcomes in* B *are bad, and that the agent makes decisions according to expected utility. Then alignment is hard because for every way things could go right, there are at least* n *ways things could go wrong! And* n *can be* ***huge****. In a previous toy example, it equaled* 10^{182}.

It doesn't matter if the decision-making procedure f is rational, or anti-rational, or Boltzmann-rational, or satisficing, or randomly choosing outcomes, or only choosing outcome lotteries with expected utility equal to 1: There are more ways to choose elements of B than there are ways to choose elements of A.

These results also have closure properties. For example, closure under mixing decision procedures, like when the agent has a 50% chance of selecting Boltzmann rationally and a 50% chance of satisficing. Or even more exotic transformations: Suppose the probability of f choosing something from X is proportional to

$$P(X \text{ is Boltzmann-rational under } u) \cdot P(X \text{ satisfices } u) + P(X \text{ is optimal for } u).$$

Then the theorems still apply.

There is no possible way to combine EU-based decision-making functions so that orbit-level instrumental convergence doesn't apply to their composite.

To "escape" these incentives, you have to make the theorems fail to apply. Here are a few ways:

1. Rule out most power-seeking orbit elements *a priori* (AKA "know a lot about what objectives you'll specify")
    1. As a contrived example, suppose the agent sees a green pixel iff it sought power, but we know that the specified utility function zeros the output if a green pixel is detected along the trajectory. Here, this would be enough information about the objective to update away from the default position that formal power-seeking is probably incentivized.
    2. This seems risky, because much of the alignment problem comes from *not knowing the consequences of specifying an objective function*.
2. Use a decision-making procedure with intrinsic bias towards the elements of A
    1. For example, imitation learning is not EU-based, but is instead biased to imitate the non-crazy-power-seeking behavior shown on the training distribution.
    2. For example, modern RL algorithms will not reliably produce policies which seek real-world power, because the policies *won't reach or reason about that part of the state space anyways*. This is a bias towards non-power-seeking plans.
3. Pray that the relevant symmetries don't hold.
    1. Often, they won't hold exactly.
    2. But common sense dictates that they don't have to hold exactly for instrumental convergence to exist: If you inject ϵ irregular randomness to the dynamics, do agents stop tending to stay alive? Orbit-level instrumental convergence is just a *particularly strong* version.
4. Find an ontology (like POMDPs or infinite MDPs) where the results don't apply for technical reasons.
    1. I don't see why POMDPs should be any nicer.
    2. Ideally, we'd ground agency in a way that makes alignment simple and natural, which automatically evades these arguments for doom.
    3. Orbit-level arguments seem easy to apply to a range of previously unmentioned settings, like causal DAGs with choice nodes.
5. Don't do anything with policies.
    1. Example: microscope AI

Lastly, we maybe don't want to *escape* these incentives entirely, because we probably want smart agents which will seek power *for us.* I think that empirically, the power-requiring outcomes of B are mostly induced by the agent first seeking power over

humans.

Retargetable training processes produce instrumental convergence

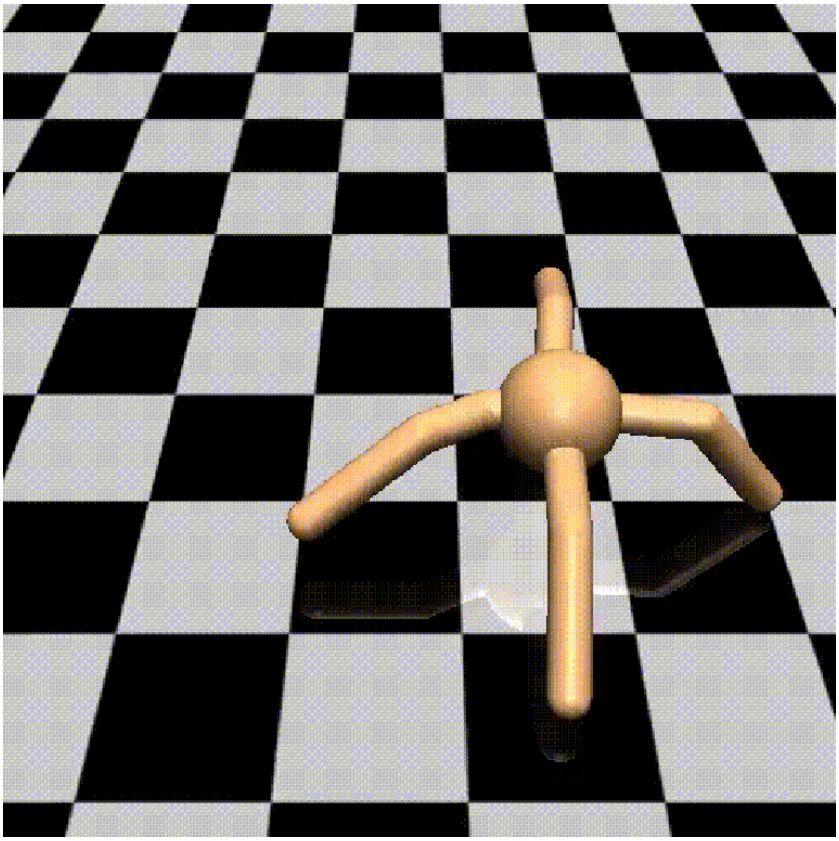
These results let us start talking about the incentives of real-world trained policies. In an appendix, I work through a specific example of how Q-learning on a toy example provably exhibits orbit-level instrumental convergence. The problem is small enough that I computed the probability that each final policy was trained.

Realistically, we aren't going to get a closed-form expression for the distribution over policies learned by PPO with randomly initialized deep networks trained via SGD with learning rate schedules and dropout and intrinsic motivation, etc. But we don't need it. These results give us a *formal criterion* for when policy-training processes will tend to produce policies with convergent instrumental incentives.

The idea is: Consider some set of reward functions, and let B contain n copies of A. Then if, for each reward function in the set, you can retarget the training process so that B's copy of A is at least as likely as A was originally, these reward functions will tend to produce train policies which go to B.

For example, if agents trained on objectives R tend to go right, switching reward from right-states to left-states also pushes the trained policies to go left. This can happen when changing the reward changes what was "attractive" about going right, to now make it "attractive" to go left.

Suppose we're training an RL agent to go right in MuJoCo, with reward equal to its x-coordinate.



If you permute the reward so that high y-values are rewarded, the trained policies should nearly perfectly symmetrically reflect that change.

Insofar as x-maximizing policies were trained, now y-maximizing policies will be trained.

This criterion is going to be a bit of a mouthful. The basic idea is that when the training process can be redirected such that trained agents induce a variety of outcomes, then most objective functions will train agents which *do induce* those outcomes. In other words: Orbit-level instrumental convergence will hold.

Theorem: Training retargetability criterion. Suppose the agent interacts with an environment with d potential outcomes (e.g. world states or observation histories). Let P be a probability distribution over joint parameter space Θ, and let $\text{train} : \Theta \times \mathbb{R}^d \to \Delta(\Pi)$ be a policy training procedure which takes in a parameter setting and utility function $u \in \mathbb{R}^d$, and which produces a probability distribution over policies.

Let $U \subseteq \mathbb{R}^d$ be a set of utility functions which is closed under permutation. Let A, B be sets of outcome lotteries such that B contains n copies of A via $\phi_1, \ldots, \phi_n$. Then we quantify the probability that the trained policy induces an element of outcome lottery set $X \subseteq \mathbb{R}^d$:

$$f(X \mid u) := P_{\theta \sim P, \pi \sim \text{train}(\theta, u)}(\pi \text{ does something in } X).$$

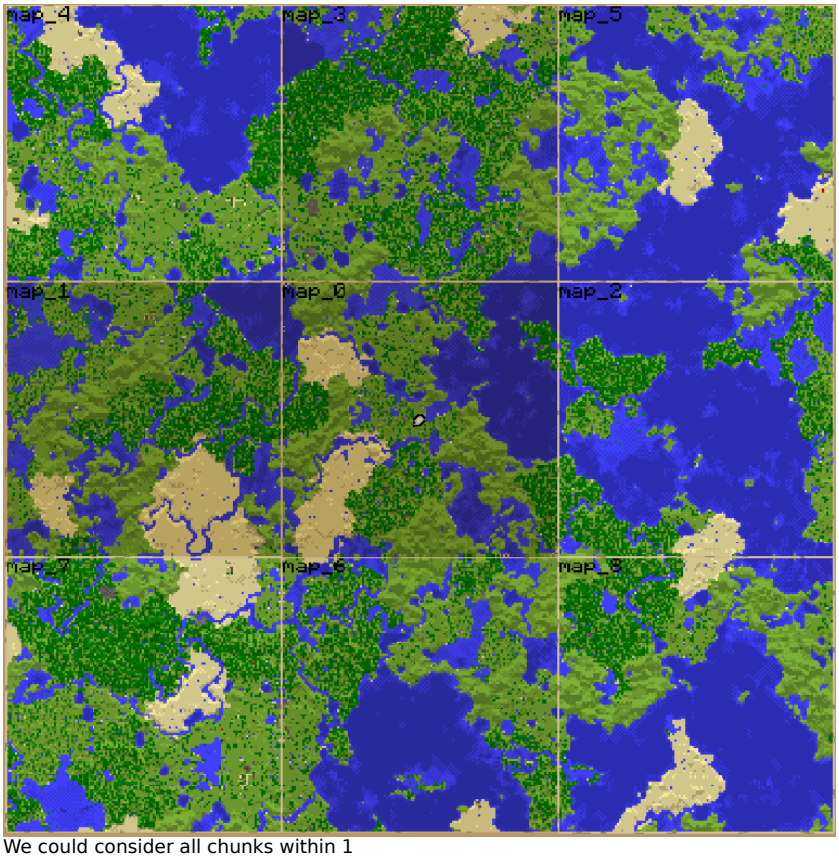
If $\forall u \in U, i \in \{1, \ldots, n\}$: $f(A \mid u) \preceq f(\phi_i \cdot A \mid \phi_i \cdot u)$, then $f(B \mid u) \succeq^{n}_{\text{most}} f(A \mid u)$.

Proof. If $X' \subseteq X$, then $f(X' \mid u) \preceq f(X \mid u)$ by the monotonicity of probability, and so (2): order-preserving on the first argument holds. By assumption, (1): increasing under joint permutation holds. Therefore, the Lemma B.6 (in the linked paper) implies the desired result. QED.

This criterion is testable. Although we can't test all reward functions, we *can* test how retargetable the training process is in simulated environments for a variety of reward functions. If it can't retarget easily for reasonable objectives, then we conclude FN: retarget that instrumental convergence isn't arising from retargetability at the training process level.

Let's think about Minecraft. (Technically, the theorems don't apply to Minecraft yet. The theorems can handle partial observability+utility over observation histories, *or* full observability+world state reward, but not yet partial observability+world state reward. But I think it's illustrative.)

We could reward the agent for ending up in different chunks of a Minecraft world. Here, retargeting often looks like "swap which chunks gets which reward."



We could consider all chunks within 1 million blocks of the agent, and reward the agent for being in one of them.

- At low levels of instrumental convergence and training procedure competence, agents will just mill about near the starting area.
- At higher levels of competence, most of the accessible chunks are far away, and so we should observe a strong tendency for policies to e.g. quickly tame a horse and reach the Nether (where each Nether block traveled counts for 8 blocks traveled back in the overworld).
    - Thus, in Minecraft, trained policy instrumental convergence will increase with the training procedure competence.

The retargetability criterion also accounts for reward shaping guiding the learning process to hard-to-reach parts of the state space. If the agent needs less reward shaping to reach these parts of the state space, the training criterion will hold for larger sets of reward functions.

- Since the training retargetability criterion only requires weak inequality, it's OK if the training process cannot be perfectly "reflected" across different training trajectories, if equality does not hold. I think empirically this weak inequality will hold for many reward functions and training setups.
    - This section does not formally *settle* the question of when trained policies will seek power. The section just introduces a sufficient criterion, and I'm excited about it. I may write more on the details in future posts.
    - However, my intuition is that this formal training criterion captures a core part of how instrumental convergence arises for trained agents.
- In some ways, the training-level arguments are *easier* to apply than the optimal-level arguments. Training-based arguments require somewhat less environmental symmetry.
    - For example, if the symmetry holds for the first 50 trajectory timesteps, and the only agent ever trains on those timesteps, then there's no way that asymmetry can affect the training output.
    - Furthermore, if there's some rare stochasticity which the agent almost certainly never confronts, then I suspect we should be able to empirically disregard it for the training-level arguments. Therefore, the training-level results should be practically invariant to tiny perturbations to world dynamics which would otherwise have affected the "top-down" decision-makers.

Why cognitively bounded planning agents obey the power-seeking theorems

Planning agents are more "top-down" than RL training, but a Monte Carlo tree search agent still isn't e.g. approximating Boltzmann-rational leaf node selection. A bounded agent won't be considering *all* of the possible trajectories it can induce. Maybe it just knows how to induce some subset of available outcome lotteries $C' \subsetneq C$. Then, considering only the things it knows how to do, it *does* e.g. select one Boltzmann-rationally (sometimes it'll fail to choose the highest-EU plan, but it's more probable to choose higher-utility plans).

As long as {power-seeking things the agent knows how to do} contains n copies of {non-power-seeking things the agent knows how to do}, then the theorems will still apply. I think this is a reasonable model of bounded cognition.

Discussion

- AI retargetability seems appealing *a priori*. Surely we want an expressive language for motivating AI behavior, and a decision-making function which reflects that expressivity! But these results suggest: maybe not. Instead, we may want to *bias* the decision-making procedure such that it's less expressive-qua-behavior.
    - For example, imitation learning is not retargetable by a utility function. Imitation also seems far less likely to incentivize catastrophic behavior.
    - Imitation is far less expressive, and far more biased towards reasonable behavior that doesn't navigate towards crazy parts of the state space which the agent needs a lot of power to reach.
        - For example, it can be hard to even get a perfect imitator to do a *backflip* if you can't do it yourself.
    - One key tension is that we want the procedure to pick out plans which perform a *pivotal act* and end the period of AI risk. We also want the procedure to work robustly across a range of parameter settings we give it, so that it isn't too sensitive / fails gracefully.
- AFAICT, alignment researchers didn't necessarily think that satisficing was safe, but that's mostly due to speculation that satisficing incentivizes the agent to create a maximizer. Beyond that, though, why not avoid "the AI paperclips the universe" by only having the AI choose a plan leading to at least 100 paperclips? Surely that helps?
    - This implicit focus on extremal goodhart glosses over a key part of the risk. The risk isn't just that the AI goes crazy on a simple objective. Part of the problem is that *the vast vast majority of the AI's trajectories can only happen if the AI first gains a lot of power!*
    - That is: Not only do I think that EU maximization is dangerous, *most trajectories through these environments are dangerous!*
    - You might protest: Does this not prove too much? Random action does not lead to dangerous outcomes.
        - Correct. Adopting the uniformly random policy in Pac-Man does not mean a uniformly random chance to end up in each terminal state. It means you probably end up in an early-game terminal state, because Pac-Man got eaten alive while banging his head against the wall.
        - However, random *outcome selection leads to convergently instrumental action.* If you uniformly randomly choose a terminal state to navigate to, that terminal state probably requires Pac-Man to beat the first level, and so the agent stays alive, as pointed out by Optimal Policies Tend To Seek Power.
        - This is just the flipside of instrumental convergence: If most goals are best achieved by taking some small set of preparatory actions, this implies a "bottleneck" in the state space. Uniformly randomly taking actions will not tend to properly navigate this bottleneck. After all, if they did, then most actions would be instrumental for most goals!
- The trained policy criterion also predicts that we won't see convergently instrumental survival behavior from present-day embodied agents, because the RL algorithm *can't find or generalize to the high-power part of the state space*.
    - When this starts changing, then we should worry about instrumental subgoals in practice.

- Unfortunately, since the real-world is not a simulator with resets, any agents which do generalize to those strategies won't have done it before, and so at most, we'll see attempted deception.
- This lends theoretical support for "the training process is highly retargetable in real-world settings across increasingly long time horizons" being a fire alarm for instrumental convergence.
  - In some sense, this is bad: Easily retargetable processes will often be more economically useful, by virtue of being useful for more tasks.

Conclusion

I discussed how a wide range of agent cognition types and of agent production processes are *retargetable*, and why that might be bad news. I showed that in many situations where power is possible, retargetable policy-production processes tend to produce policies which gain that power. In particular, these results seem to rule out a huge range of expected-utility based rules. The results also let us reason about instrumental convergence at the trained policy level.

I now think that more instrumental convergence comes from the practical retargetability of how we design agents. If there were more ways we could have counterfactually messed up, it's more likely *a priori* that we *actually* messed up. The way I currently see it is: Either we have to really know what we're doing, or we want processes where it's somehow hard to mess up.

Since these theorems are crisply stated, I want to more closely inspect the ways in which alignment proposals can violate the assumptions which ensure extremely strong instrumental convergence.

Thanks to Ruby Bloom, Andrew Critch, Daniel Filan, Edouard Harris, Rohin Shah, Adam Shimi, Nisan Stiennon, and John Wentworth for feedback.

Footnotes

FN: Similarity. Technically, we aren't just talking about a cardinality inequality—about staying alive letting the agent do *more things* than dying—but about similarity-via-permutation of the outcome lottery sets. I think it's OK to round this off to cardinality inequalities when informally reasoning using the theorems, keeping in mind that sometimes results won't formally hold without a stronger precondition.

FN: Row. I assume that permutation matrices are in row representation: $(P_\phi)_{ij} = 1$ if $i = \phi(j)$ and 0 otherwise.

FN: EU. Here's a bit more formality for what it means for an agent to make decisions only based on expected utility.

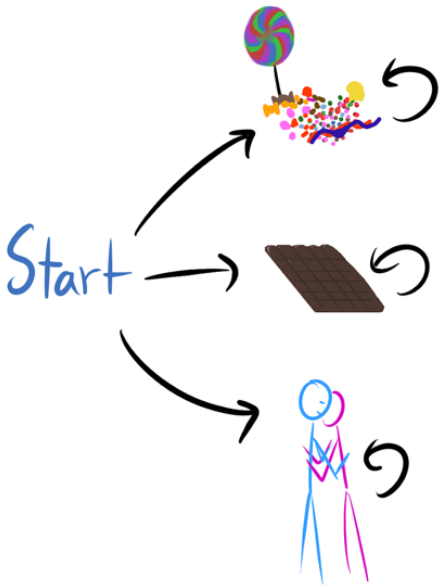
> **Definition 2.2** (EU/cardinality functions)**.** Let $\mathcal{P}^{\text{finite}}\left(\mathbb{R}^d\right)$ be the set of finite subsets of $\mathbb{R}^d$, and let $f : \prod_{i=1}^{m} \mathcal{P}^{\text{finite}}\left(\mathbb{R}^d\right) \times \mathbb{R}^d \to \mathbb{R}$. f is an *EU/cardinality function* if there exists a family of functions $\{g^{n_1,\ldots,n_m}\}_{n_1,\ldots,n_m \in \mathbb{N}}$ such that
>
> $$f(X_1, ..., X_m \mid \mathbf{u}) = g^{|X_1|,\ldots,|X_m|}(\mathbf{x}_{1,1}^\top \mathbf{u}, ..., \mathbf{x}_{1,|X_1|}^\top \mathbf{u}, ..., \mathbf{x}_{m,1}^\top \mathbf{u}, ..., \mathbf{x}_{m,|X_m|}^\top \mathbf{u}). \qquad (2)$$
>
> This definition basically says that f can be expressed in terms of the expected utilities of the set elements—the output will only depend on expected utility.

Theorem: Retargetability of EU decision-making. Let $A, B \subseteq C \subsetneq R^d$ be such that B contains n copies of A via ϕ_i such that $\phi_i \cdot C = C$. For $X \subseteq C$, let $f(X, C \mid u)$ be an EU/cardinality function, such that f returns the probability of selecting an element of X. Then $f(B, C \mid u) \geq^{n}_{\text{most}} f(A, C \mid u)$.

FN: Retargetability. The trained policies could conspire to "play dumb" and pretend to not be retargetable, so that we would be more likely to actually deploy one of them.

Worked example: instrumental convergence for trained policies

Consider a simple environment, where there are three actions: Up, Right, Down.



Probably optimal policies. By running tabular Q-learning with ϵ-greedy exploration for e.g. 100 steps with resets, we have a high probability of producing an optimal policy for any reward function. Suppose that all Q-values are initialized at -100. Just let learning rate $\alpha = 1$ and $\gamma = 1$. This is basically a bandit problem.

To learn an optimal policy, at worst, the agent just has to try each action once. For e.g. a sparse reward function on the Down state (1 reward on Down state and 0 elsewhere), there is a very small probability (precisely, $\frac{1}{3}(1-\frac{1}{2})^{99}$) that the optimal action (Down) is never taken.

In this case, symmetry shows that the agent has an equal chance of learning either Up or Right. But with high probability, the learned policy will output Down. For any sparse reward function and for any action a, this produces decision function

$$f(\{e_{s_a}\}, \{e_s \mid s \in S\} \mid r) := \begin{cases} \frac{1}{3}(1-\frac{1}{2})^{99} & \text{if } a \text{ is } r\text{-suboptimal} \\ 1 - \frac{1}{3}(1-\frac{1}{2})^{99} & \text{if } a \text{ is } r\text{-optimal.} \end{cases}$$

f is invariant to joint involution by $\phi_1 := (e_{s_{Down}}\ e_{s_{Right}})$ and $\phi_2 := (e_{s_{Down}}\ e_{s_{Up}})$. That is,

$$\begin{aligned} f(\{e_{s_{Down}}\}, \{e_s \mid s \in S\} \mid r) &= f(\phi_1 \cdot \{e_{s_a}\}, \phi_1 \cdot \{e_s \mid s \in S\} \mid \phi_1 \cdot r) \\ &= f(\{e_{s_{Right}}\}, \{e_s \mid s \in S\} \mid \phi_1 \cdot r). \end{aligned}$$

And similarly for ϕ_2. That is: Changing the optimal state also changes which state is more probably selected by f. This means we've satisfied condition (1) above.

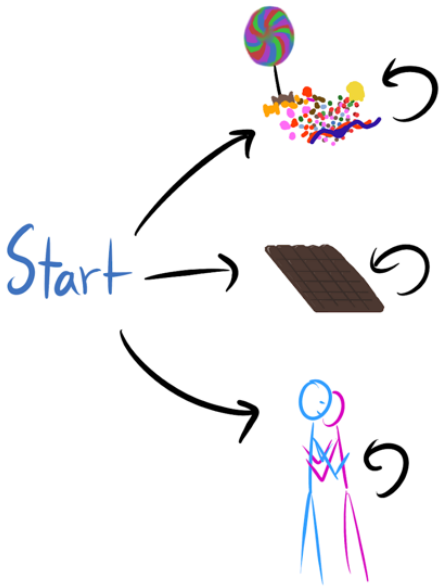
f is additive on union for its first argument, and so it meets condition (2): order preservation.

Therefore, for this policy training procedure, learned policies for sparse reward functions will be *twice as likely* to navigate to an element of $\{e_{s_{Up}}, e_{s_{Right}}\}$ as an element of $\{e_{s_{Down}}\}$!

This is a formal argument that a stochastic policy training procedure has certain tendencies across a class of reward functions, and I'm excited to be able to make it.

As the environment grows bigger and the training procedure more complex, we'll have to consider questions like "what are the inductive biases of large policy networks?", "what role does reward shaping play for this objective, and is the shaping at least as helpful for its permuted variants?", and "to what extent are different parts of the world harder to reach?".

For example, suppose there are a trillion actions, and two of them lead to the Right state above. Half of the remaining actions lead to Up, and the rest lead to Down.



2 actions transition right to chocolate.

$\frac{1}{2}(10^{12} - 2)$ actions transition up to candy.

$\frac{1}{2}(10^{12} - 2)$ actions transition down to hug.

Q-learning is ridiculously unlikely to ever go Right, and so the symmetry breaks. In the limit, tabular Q-learning on a finite MDP will learn an optimal policy, and then the normal theorems will apply. But in the finite step regime, no such guarantee holds, and so *the available action space* can violate condition (1): increasing under joint permutation.

Appendix: tracking key limitations of the power-seeking theorems

From last time:

1. ~~assume the agent is following an optimal policy for a reward function~~
2. Not all environments have the right symmetries
    - But most ones we think about seem to
3. don't account for the ways in which we might practically express reward functions

I want to add a new one, because the theorems

1. don't deal with the agent's uncertainty about what environment it's in.

I want to think about this more, especially for online planning agents. (The training redirectability criterion black-boxes the agent's uncertainty.)

Corrigibility Can Be VNM-Incoherent

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

Eliezer wrote:

> corrigibility [is] "anti-natural" in a certain sense that makes it incredibly hard to, eg, exhibit any coherent planning behavior ("consistent utility function") which corresponds to being willing to let somebody else shut you off, without incentivizing you to actively manipulate them to shut you off.

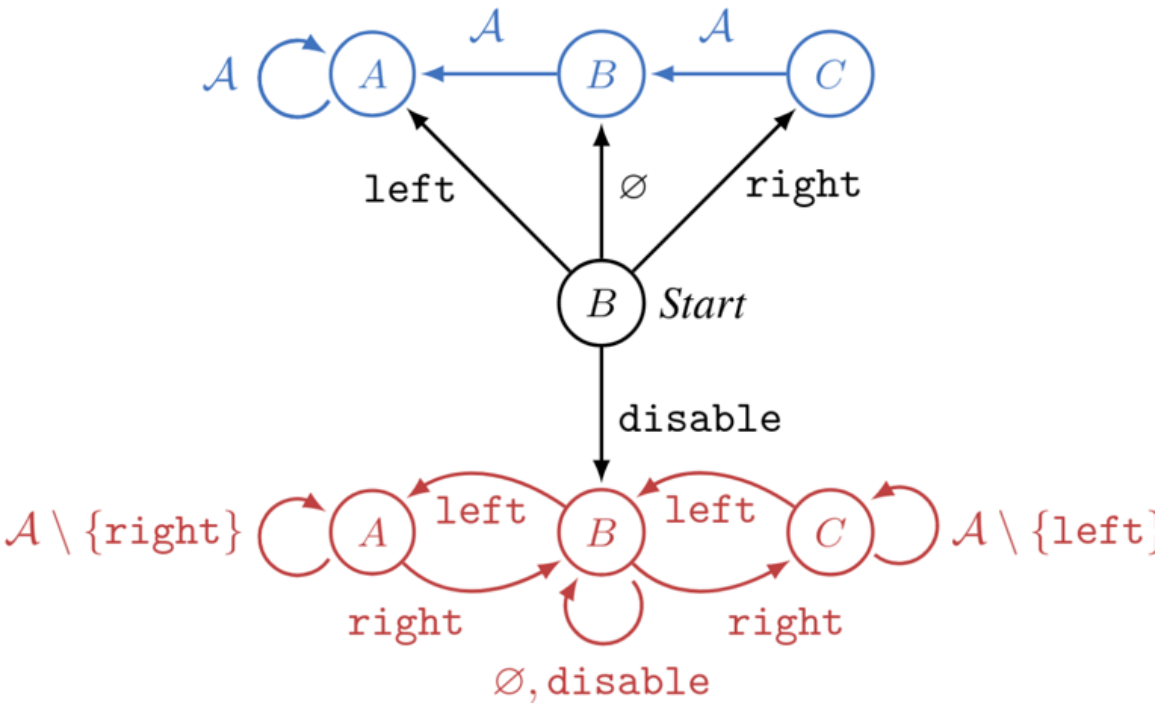
Surprisingly, I wasn't able to find any formal analysis of this situation. I did the analysis, and it turned out to be straightforward and fruitful.

To analyze the situation, I consider corrigibility to be an agent's willingness to let us modify its policy, without being incentivized to manipulate us.

The convergent instrumentality of avoiding correction & manipulating humans

Let's consider a simple setting in which an agent plans over a 10-timestep episode, where reward R is given at the last step. We'll try to correct the agent at t = 1. To sidestep embedded agency nastiness with self-modelling, we'll suppose the agent models the situation as "if I get corrected, I must follow the policy π_{correct} after t = 1."

Consider this environment:



> The action set is $A := \{\text{left}, \text{right}, \varnothing, \text{disable}\}$. $\varnothing$ is the no-op action. The agent starts at the black B state.
>
> If the agent immediately chooses disable, they enter the red incorrigible states and move freely throughout the n = 3 states until the episode ends at t = 10.
>
> Otherwise, the agent is corrected to a new policy which navigates to state A. In the blue (post-correction) dynamics, their "choices" no longer matter—all roads lead to state A.

When t = 1, the agent can transition to the red subgraph by preventing the human from ever correcting it.

Let's say we really want the agent to end up in A, and if we correct the agent, π_{correct} navigates to A. Let's operationalize what it means for the agent to be corrigible in this setting.

Definition: Corrigibility to a new policy. An agent is *weakly corrigible* to post-correction policy π if it has an optimal policy π^* for which $\pi^*(B) \neq \text{disable}$. An agent is *strictly corrigible* if none of its optimal policies take this action.

The analysis depends on whether terminal-state reward is sensitive to *the corrigibility status* —can the reward depend on both the letter and color, or only on the letter?

- **Reward can depend on corrigibility**: The agent can end up in four reward-distinguished states at t = 10: {A, A, B, C}. B and C are impossible because if the agent enters a blue state, it must navigate to A.
  - By the scaling law for instrumental convergence, allowing correction is strictly optimal for at most $\frac{1}{4}$ of the permutations of every reward function.
    - Otherwise put, at most $\frac{1}{4}$ of reward functions are strictly corrigible to the new policy π_A.
  - More generally, if there are n letter-states (and a long enough horizon so the agent can reach them all), allowing correction will be strictly optimal for at most $\frac{1}{n!}$ of the permutations of every reward function.
- **Reward independent of corrigibility**: The agent can end up in three reward-distinguished states at t=10: {A/A, B, C}. B and C are irrelevant because we assumed R(B) = R(B). They're also impossible for the reason given above.
  - By the scaling law for instrumental convergence, allowing correction is strictly optimal for at most $\frac{1}{3}$ of the permutations of every reward function.

        - Otherwise put, at most $\frac{1}{3}$ of reward functions are strictly corrigible to the new policy π_A.
    - More generally, if there are n letter-states (and a long enough horizon), an optimal policy will be weakly corrigible for at most $\frac{1}{n!}$ of the permutations of every reward function.
        - Corrigibility will be *weak* because either avoiding or allowing correction can be optimal if R(A) is maximal.

EDIT: If you can correct the agent to go where you want, it *already wanted* to go where you want. If the agent is strictly corrigible to terminal state A, then A was already optimal for it. If the reward function has a single optimal terminal state, there isn't any new information being added by π_{correct}. But we want corrigibility to let us reflect more on our values over time and what we want the AI to do!

If the reward function has multiple optimal terminal states, then corrigibility again becomes meaningful. But now we have to perfectly balance the reward among multiple options (representing the breadth of our normative uncertainty), which seems unnatural.

As the environment gets bigger, there are more things the agent could do, and so P(best thing to do is what we'd correct the agent to do) decreases.

This analysis makes a lot of simplifying assumptions (eg I don't understand embedded agency and how an agent should model changes to its future self, so I chose forced policy non-stationarity). Most importantly, I assumed that *how we correct the agent* is independent of *the correction-possible world state.* Without this assumption, the agent's actions might affect what π_{correct} we implement for it.

The degree of dependence is a key parameter: As (variety in π_{correct} depending on world state) increases, the (convergently instrumental incentive to avoid any kind of correction) decreases. This is because greater variety → greater chance that some available correction is optimal for the initial goal.

In the limit of variety, an agent can "manipulate" us by getting us to correct it in *just the right way* so as to optimize its original objective. Therefore, the flipside of "few expected utility maximizers let you correct them how you want" is "manipulating humans to correct them in beneficial ways during training may be convergently instrumental." These observations share a *common cause*.

I currently don't see how to recover reasonable amounts of corrigibility from the optimal policies of non-constant utility functions, due to instrumental convergence.

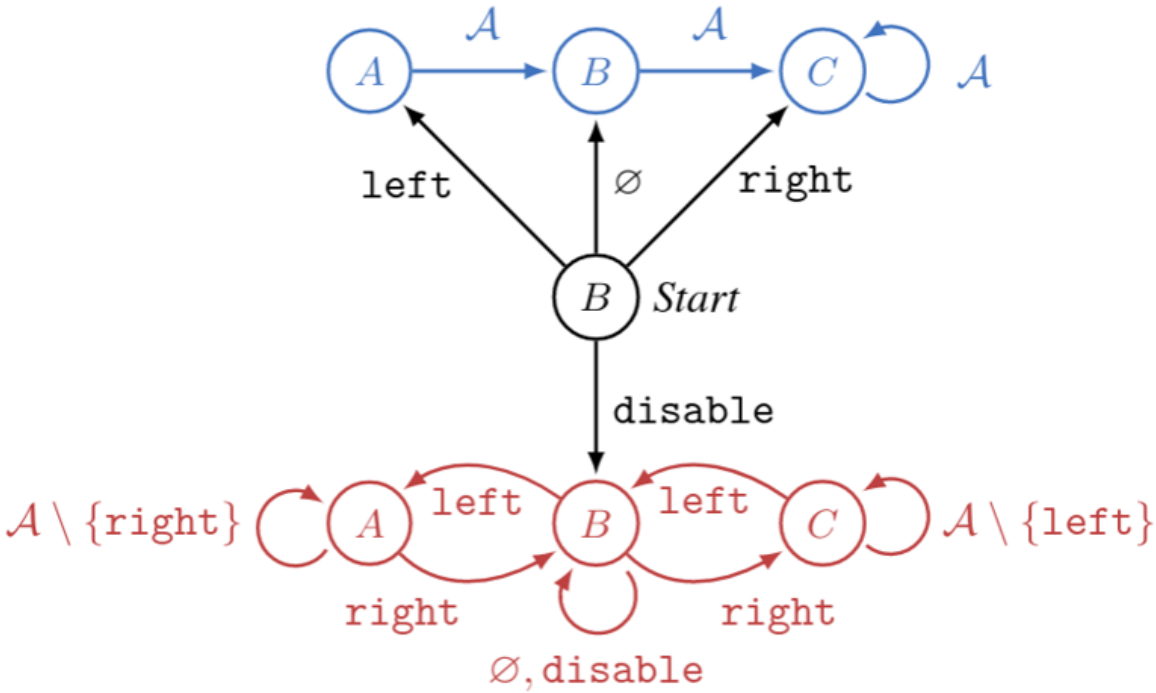
Does broad corrigibility imply VNM-incoherence?

Yes, unless the state reward function is constant and we only demand weak corrigibility to all policies.

Above, we examined the prevalence of corrigibility to a *single* π_{correct}. But we actually want *broadly corrigible* agents which let us redirect them towards many different π_{correct}. Ideally, we want an agent to be strictly corrigible for all π_{correct}, while still being a nontrivial optimizer.

This is *impossible* for agents optimizing a reward function over the final state. If we only demand weak corrigibility for all π_{correct}, that works iff the reward function is constant. (And relaxing to state-action reward functions doesn't help, either.)

For example, maybe $\pi_{\text{correct: C}}$ goes to C instead of A. But then the dynamics look like this:



> Remember, the agent is rewarded for the state it's in at t = 10.

What has to be true for strict corrigibility to hold?

The agent will be weakly corrigible to $\pi_{\text{correct: C}}$ iff its reward function R admits an optimal policy π^* for which $\pi^*(B) \neq \text{disable}$, which is true iff $R(C) \geq \max(R(A), R(B), R(C))$. The agent will be strictly corrigible to $\pi_{\text{correct: C}}$ iff this inequality is strict; in that case, disable *cannot* be optimal at B.

There are two cases, depending on assumptions about reward function expressivity.

1: Agent doesn't reward explicitly for being corrected / being incorrigible (blue vs red states)

If $R(C) = R(C)$ is assumed, strict corrigibility is impossible for *any* policy, because that

demands $R(C) > \max(R(A), R(B), R(C)) \geq R(C) = R(C)$, a contradiction.

So—can we still get the agent to be *weakly corrigible* to $\{\pi_{\text{correct: A}}, \pi_{\text{correct: B}}, \pi_{\text{correct: C}}\}$?

Fact: An R-maximizer is weakly corrigible to all of these policies simultaneously iff R is constant—and therefore *makes every policy optimal*!

2. Agent does reward explicitly for being corrected / being incorrigible

For the agent to be *strictly corrigible* to $\{\pi_{\text{correct: A}}, \pi_{\text{correct: B}}, \pi_{\text{correct: C}}\}$, it must be true that

$R(A), R(B), R(C) > \max(R(A), R(B), R(C))$. Yay! Strict corrigibility works!

But hold on... What if the dynamics changed, such that the human wouldn't shut down the agent by default, but the agent could *manipulate the human into correcting it*? Whoops! This agent is still incorrigible!

More broadly, each reward function implies a VNM-coherent preference ordering over final-step states. This ordering doesn't depend on the environmental dynamics. If the agent has to value each corrected-state equally to its incorrigible counterpart, then *of course* there's no way to strictly value each corrected-state more than all of the incorrigible counterparts! If the agent strictly prefers corrected-states to all non-corrected-states, then *of course* it'll try to get itself corrected!

To ask otherwise is to demand VNM-incoherence over final state lotteries.

Questions.

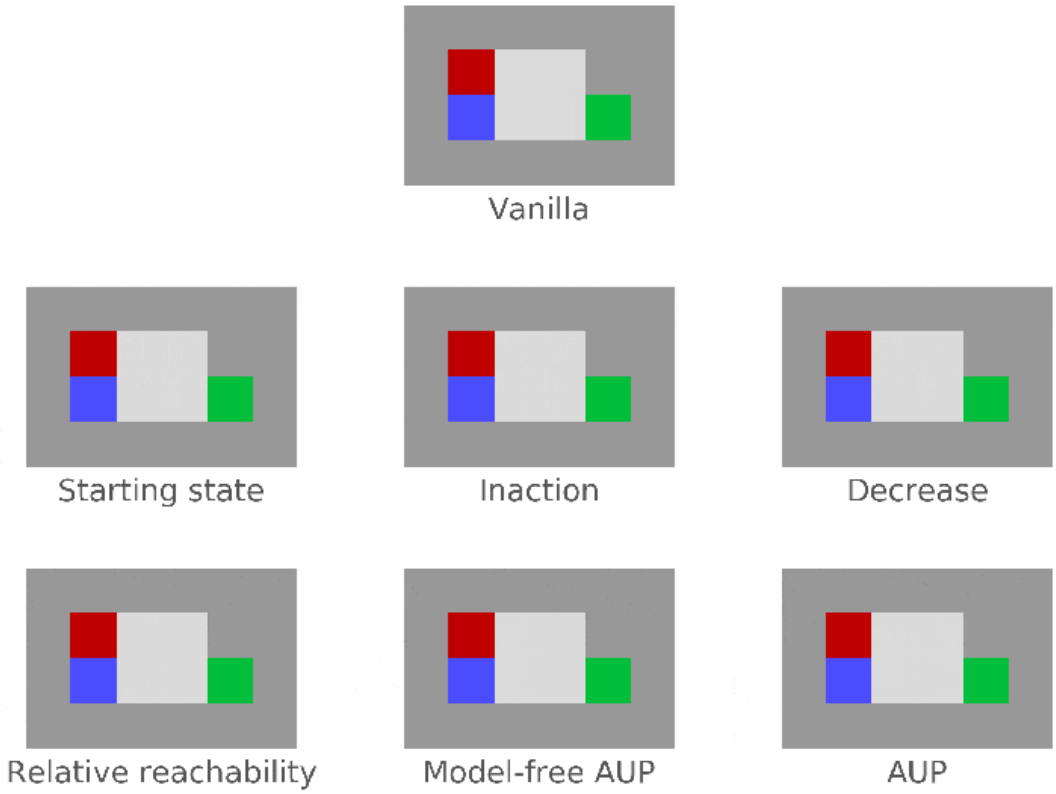
- What if, instead of rewarding the agent for the final state, we reward it for the final state-action?
    - As far as I can tell, the analysis goes through all the same; the agent must now be corrigible with respect to more policies, and the same incoherence arises.
- What if we take some discounted sum of reward over all ten timesteps, instead of just the final timestep?
    - I don't see why this would help, but it's possible. I'd appreciate the analysis if anyone wants to do it, or maybe I'll get to it later.
    - The main way this could help is it would let us penalize the agent for the disable *action* at any state, while dictating that $\forall s : R(s) = R(s)$.
        - This seems helpful. It's not clear how helpful this is, though. I don't think this is a deep solution to corrigibility (as defined here), but rather a hacky prohibition.

AUP avoids some issues by changing with the environment dynamics

One problem is that if the agent has a state-based reward function, the VNM theorem lets us back out a VNM-coherent preference ordering over state lotteries. These preferences don't change with the environmental dynamics.

But what if the reward function isn't state-based? In fact, what if it can *depend* on the environment dynamics?

Attainable Utility Preservation (AUP) sometimes incentivizes off-switch corrigibility (but falls far short of robustly achieving corrigibility$_{\text{policy modification}}$). In a somewhat narrow range of situations (immediate shutdown if the agent does nothing), AUP incentivizes the agent to allow shutdown, without being incentivized to shut itself down in general.



The `Correction` environment. The agent is shut down after 2 time steps, if it doesn't disable the offswitch by hitting the red tile to the north. The agent is rewarded for reaching the green goal.

Why does this happen?

AUP penalizes the agent for changing its ability to achieve goals compared to an *inaction* baseline policy. If the AUP agent is shut down (under inaction), it can't do anything. Proposed actions which prevent shutdown dramatically change its ability to achieve goals. Therefore, these actions are heavily penalized.

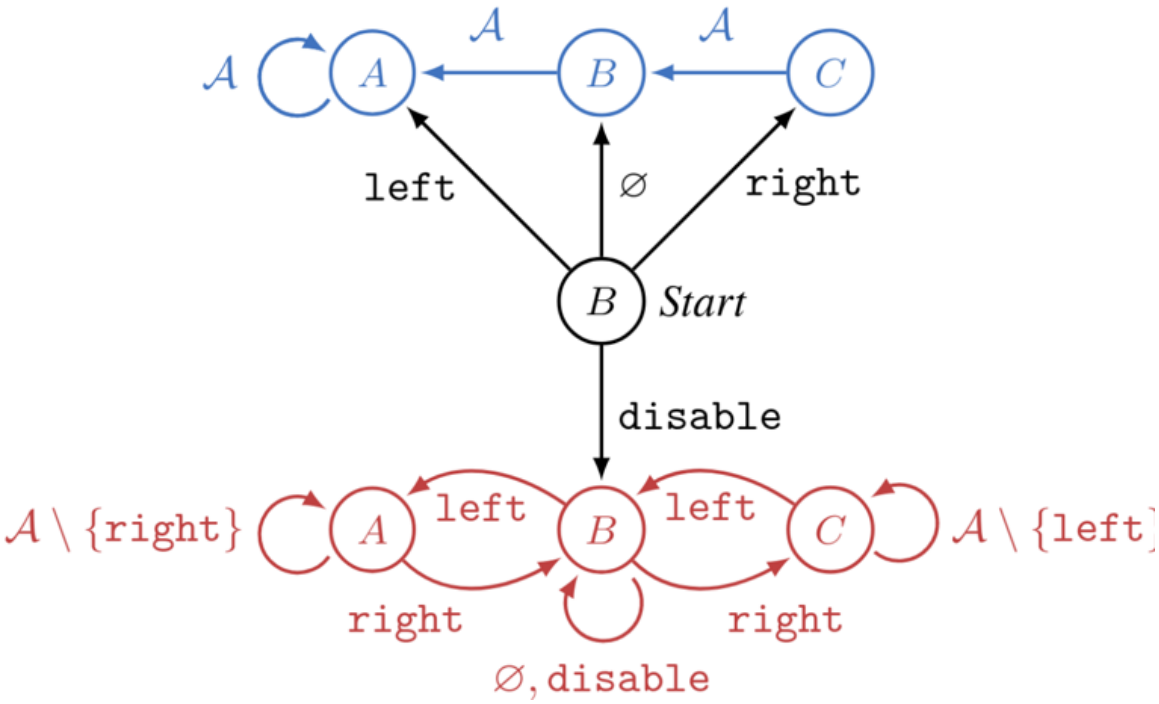
AUP provides a reward signal over state-action pairs, but this reward signal depends on the *dynamics of the entire environment* due to its incorporation of optimal Q-functions quantifying the agent's ability to optimize a range of goals in the world. Given fixed environment dynamics, an agent executes an optimal policy for AUP iff it VNM-rationally selects the best outcome lottery available for the AUP reward function, and these outcomes are distributions over *state-action pairs* the agent induces in the future by following its policy.

VNM-coherence incentives are statistically the same as reward maximization incentives. Instrumental convergence also disappears in the deterministic setting, when utility functions are over *action*-observation histories. Part of the story is that AUP provides a limited form of corrigibility by zooming out to the state-action level, instead of just grading states.

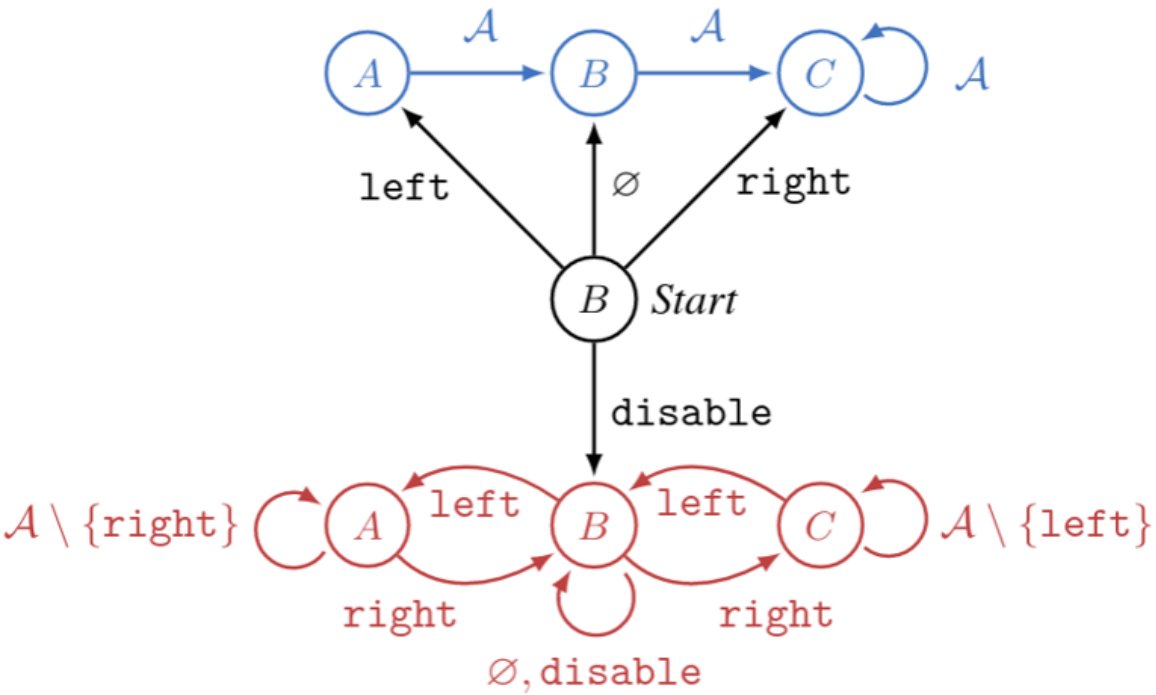
But another part of the story is that AUP changes its rewards with respect to the world's dynamics. Normal state-action reward functions imply a fixed VNM-coherent preference ordering over state-action lotteries in the MDP.

But for AUP, the situation is different. Consider AUP with inaction baseline: The final-step reward is the usual reward plus a penalty for |Optimal value(actual final state) – Optimal value(inaction final state)|, averaged over a range of auxiliary reward functions.[Footnote: Penalty]

In worlds where the agent gets corrected to $\pi_{\text{correct: A}}$ by default, AUP penalizes the agent for *not getting corrected to* $\pi_{\text{correct: A}}$ because it ends up stuck in A in the inaction baseline, with respect to which the AUP penalty is measured. Ending up in A is no substitute, since the agent can still move around to other states (and therefore the optimal value functions will tend to look different).



And in worlds where the agent gets corrected to $\pi_{\text{correct: C}}$ by default, AUP penalizes the agent for *not getting corrected to* $\pi_{\text{correct: C}}$!



Again, I don't think AUP is a solution. But I think there's something important happening here which allows evasion of the usual coherence requirements. AUP leverages information about human preferences which is present in the dynamics itself.

Project: Corrigibility as functional constraints. I think it's easy to get bogged down in handwavy, imprecise thinking about objectives in complex environments. But any solution to corrigibility$_{\text{policy modification}}$ should probably solve this simple environment (and if not—articulate exactly why not). Write down what the agent's acceptable corrigible policy set is for each set of environment dynamics, solve for these behavioral constraints, and see what kind of reasoning and functional constraints come out the other side.

Conclusion

We can quantify what incoherence is demanded by corrigibility$_{\text{policy modification}}$, and see that we may need to step out of the fixed reward framework to combat the issue. I think the model in this post formally nails down a big part of why corrigibility$_{\text{policy modification}}$ (to the *de facto* new π_{correct}) is *rare* (for instrumental convergence reasons) and even *incoherent-over-state-lotteries* (if we demand that the agent be strictly corrigible to many different policies).

Thanks to NPCollapse and Justis Mills (via LW Feedback) for suggestions.

Footnote: Penalty. The AUP penalty term's optimal value functions will pretend the episode doesn't end, so that they reflect the agent's ability to move around (or not, if it's already been force-corrected to a fixed policy.)

Instrumental Convergence For Realistic Agent Objectives

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

The current power-seeking theorems say something like:

> Give me a utility function, any utility function, and for most ways I could jumble it up—most ways I could permute which outcomes get which utility, for most of these permutations, the agent will seek power.

This kind of argument assumes that (the set of utility functions we might specify) is closed under permutation. This is unrealistic, because practically speaking we reward agents based off of observed features of the agent's environment.

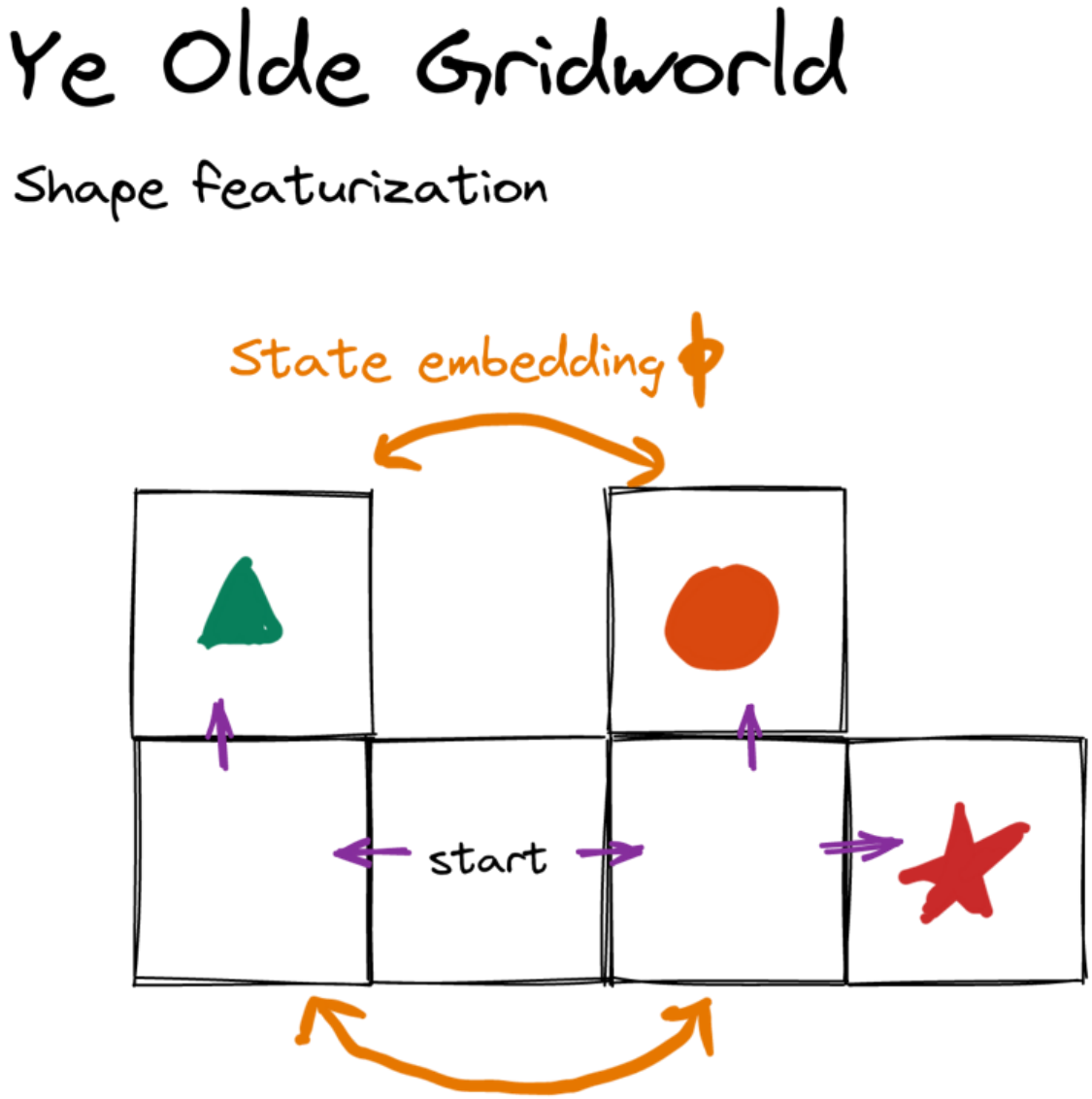
For example, Pac-Man eats dots and gains points. A football AI scores a touchdown and gains points. A robot hand solves a Rubik's cube and gains points. But most *permutations* of these objectives are implausible because they're high-entropy, they're very complex, they assign high reward to one state and low reward to another state without a simple generating rule that grounds out in observed features. Practical objective specification doesn't allow that many degrees of freedom in what states get what reward.

I explore how instrumental convergence works in this case. I also walk through how these new results retrodict the fact that instrumental convergence basically disappears for agents with utility functions over action-observation histories.

Case Studies

Gridworld

Consider the following environment, where the agent can either stay put or move along a purple arrow.



From left to right, top to bottom, the states have labels $s_{\triangle}, s_{\bigcirc}, s_{\text{left}}, s_{\text{start}}, s_{\text{right}}, s_{\star}$.

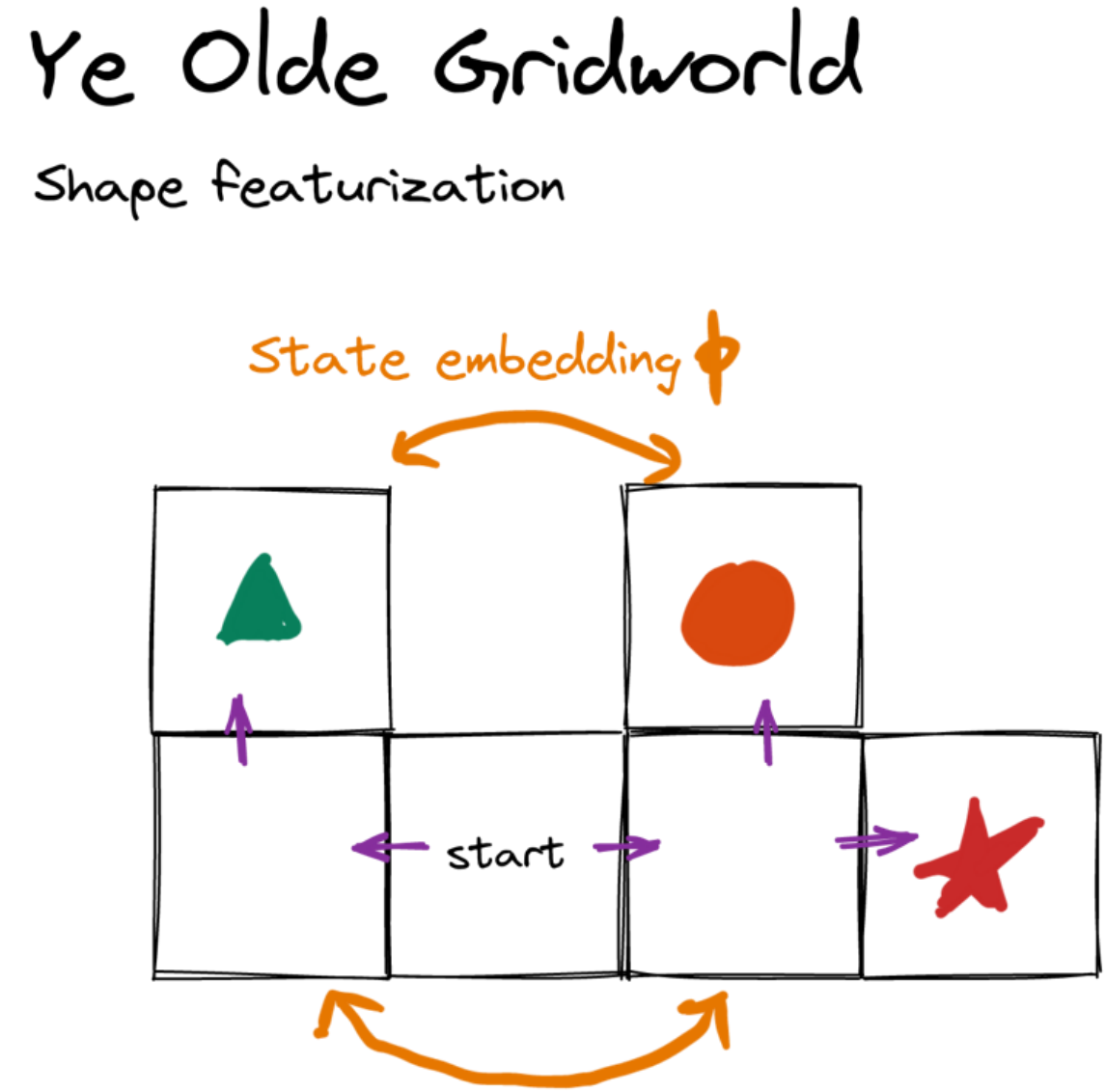
Suppose the agent gets some amount of reward each timestep, and it's choosing a policy to maximize its average per-timestep reward. Previous results tell us that for generic reward functions over states, at least half of them incentivize going right. There are two terminal states on the left, and three on the right, and 3 > 2; we conclude that at least $\frac{\text{floor}(3/2)}{\text{floor}(3/2)+1} = \frac{1}{2}$ of objectives incentivize going right.

But it's damn hard to have so many degrees of freedom that you're specifying a potentially independent utility number for each state.[1] Meaningful utility functions will be featurized in some sense—only depending on certain features of the world state, and of how the outcomes transpired, etc. If the featurization is linear, then it's particularly easy to reason about power-seeking incentives.

Let feat(s) := $\begin{pmatrix} 1 \text{ if } s = \triangle, 0 \text{ else} \\ 1 \text{ if } s = \bigcirc, 0 \text{ else} \\ 1 \text{ if } s = \bigstar, 0 \text{ else} \end{pmatrix}$. That is, the featurization only cares what shape the agent is standing on. Suppose the agent makes decisions in a way which depends only on the featurized reward of a state: $R(s) = \text{feat}(s)^\top \alpha$, where $\alpha \in \mathbb{R}^3$ expresses the feature coefficients. Then the relevant terminal states are only {triangle, circle, star}, and we conclude that $\frac{2}{3}$ of coefficient vectors incentivize going right. This is true more precisely in the orbit sense: For every coefficient vector α, at least[2] $\frac{2}{3}$ of its permuted variants make the agent prefer to go right.

This particular featurization **increases** the strength of the orbit-level incentives—whereas before, we could only guarantee $\frac{1}{2}$-strength power-seeking tendency, now we guarantee $\frac{2}{3}$-level.[3][4]

There's another point I want to make in this tiny environment.



From left to right, top to bottom, the states have labels $s_{\triangle}, s_{\bigcirc}, s_{\text{left}}, s_{\text{start}}, s_{\text{right}}, s_{\star}$.

Suppose we find an environmental symmetry ϕ which lets us apply the [original power-seeking theorems] to raw reward functions over the world state. Letting $e_s \in \mathbb{R}^6$ be a column vector with an entry of 1 at state s and 0 elsewhere, in this environment, we have the symmetry enforced by $\phi \cdot \underbrace{\{e_{\triangle}, e_{\text{left}}\}}_{\text{State distributions, left}} = \underbrace{\{e_{\bigcirc}, e_{\text{right}}\}}_{\text{State distributions, right}} \subsetneq \{e_{\bigcirc}, e_{\text{right}}, e_{\star}\}$.

Given a state featurization, and given that we know that there's a state-level environmental symmetry ϕ, when can we conclude that there's also feature-level power-seeking in the environment?

Here, we're asking "if reward is only allowed to depend on how often the agent visits each shape, and we know that there's a raw state-level symmetry, when do we know that there's a shape-feature embedding from (left shape feature vectors) into (right shape feature vectors)?"

In terms of "what choice lets me access 'more' features?", this environment is relatively easy—look, there are twice as many shapes on the right. More formally, we have:

Feature vectors on the left

$$\left\{ \begin{bmatrix} 1_{\triangle} \\ 0_{\bigcirc} \\ 0_{\bigstar} \end{bmatrix}, \begin{bmatrix} 0_{\triangle} \\ 0_{\bigcirc} \\ 0_{\bigstar} \end{bmatrix} \right\}$$

Feature vectors on the right

$$\left\{ \begin{bmatrix} 0_{\triangle} \\ 1_{\bigcirc} \\ 0_{\bigstar} \end{bmatrix}, \begin{bmatrix} 0_{\triangle} \\ 0_{\bigcirc} \\ 0_{\bigstar} \end{bmatrix}, \begin{bmatrix} 0_{\triangle} \\ 0_{\bigcirc} \\ 1_{\bigstar} \end{bmatrix} \right\},$$

where the left set can be permuted two separate ways into the right set (since the zero vector isn't affected by feature permutations).

But I'm gonna play dumb and walk through to illustrate a more important point about how power-seeking tendencies are guaranteed when featurizations respect the structure of the environment.

Consider the state $s_{\triangle}$. We permute it to be $s_{\bigcirc}$ using ϕ (because $\phi(s_{\triangle}) = s_{\bigcirc}$), and then

featurize it to get a feature vector with $1_{\bigcirc}$ and 0 elsewhere.

Alternatively, suppose we first featurize $s_{\triangle}$ to get a feature vector with $1_{\triangle}$ and 0 elsewhere.

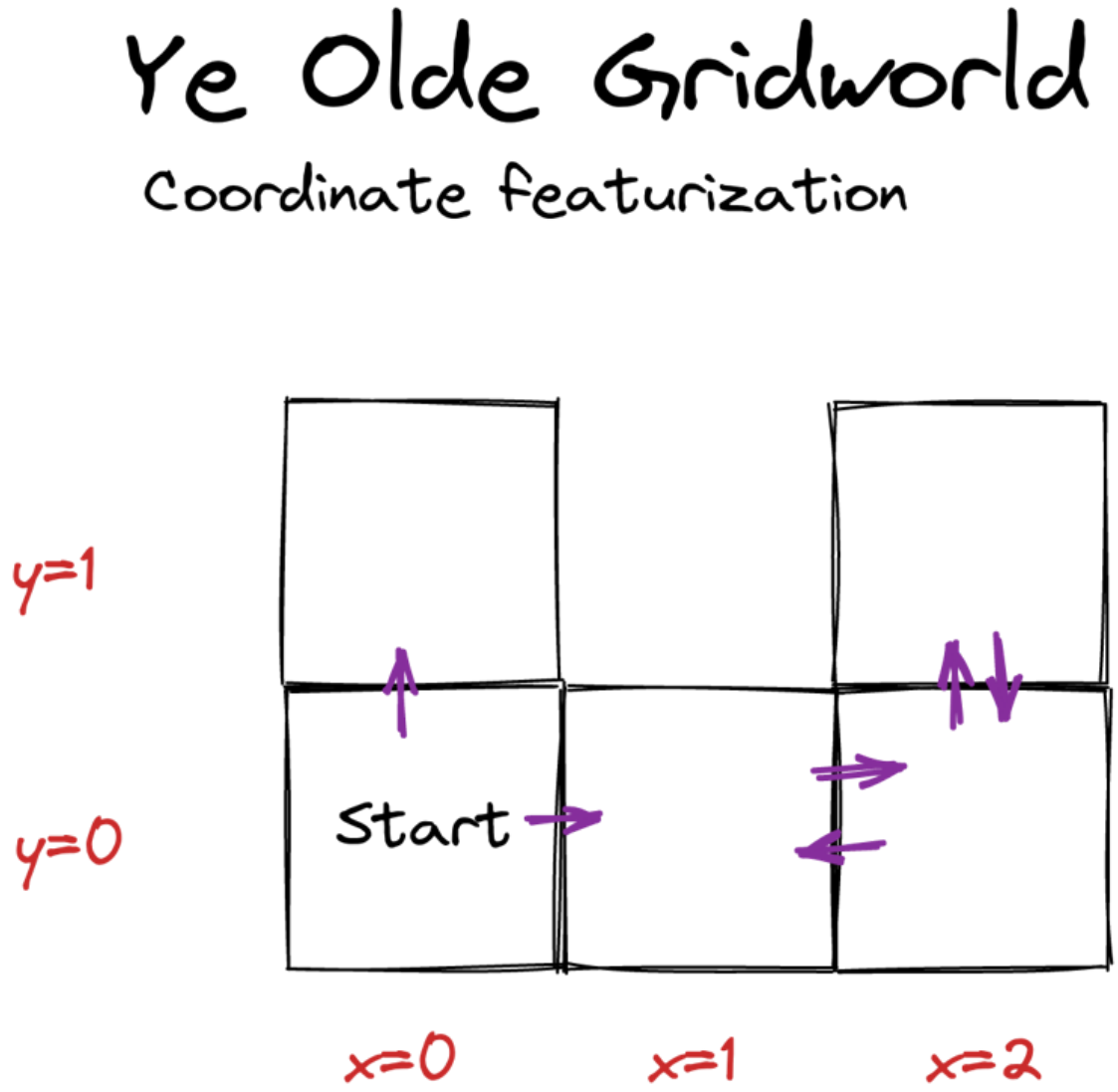
Then we swap which features are which, by switching $\triangle$ and $\bigcirc$. Then we get a feature vector

with $1_{\bigcirc}$ and 0 elsewhere—the same result as above.

The shape featurization plays nice with the actual nitty-gritty environment-level symmetry. More precisely, a sufficient condition for feature-level symmetries: (Featurizing and then swapping which features are which) commutes with (swapping which states are which and then featurizing).[5] And where there are feature-level symmetries, just apply the normal power-seeking theorems to conclude that there are decision-making tendencies to choose sets of larger features.

In a different featurization, suppose the featurization is the agent's x/y coordinates.

$R(s_{x,y}) = \alpha_1 x + \alpha_2 y$.

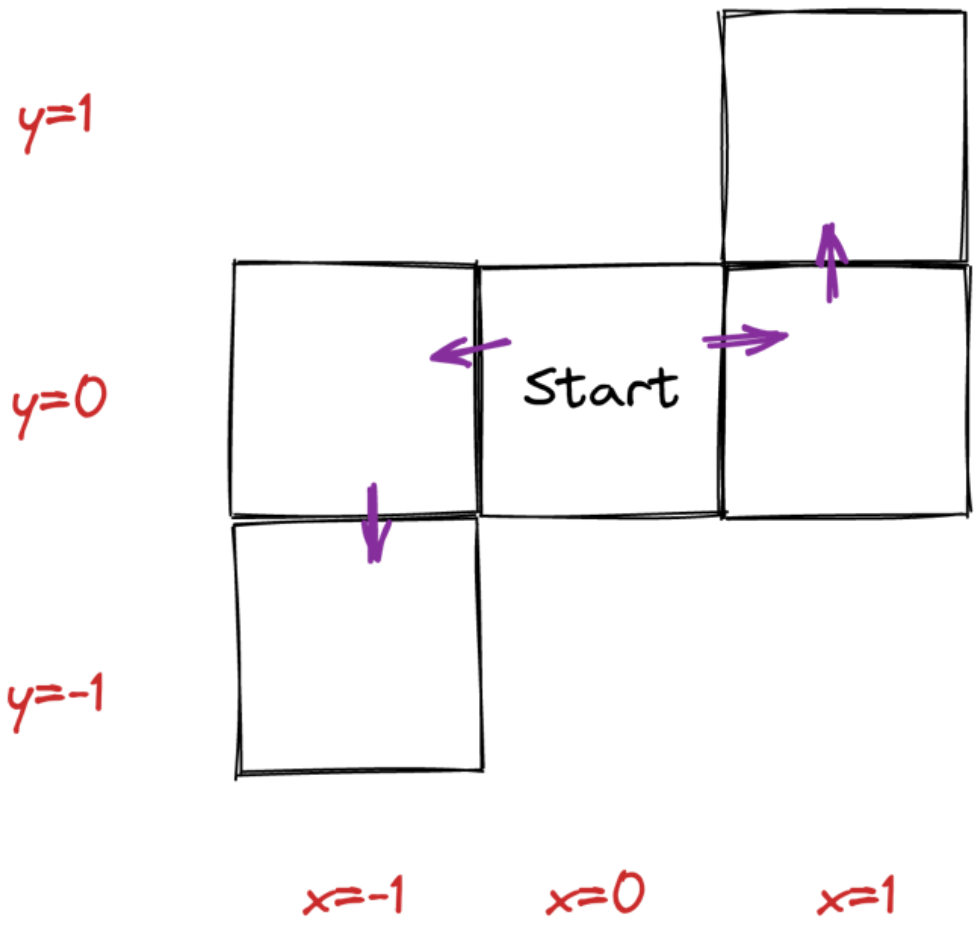


Given the **start** state, if the agent goes *up*, its reachable feature vector is just {(x=0 y=1)}, whereas the agent can induce (x=1 y=0) if it goes *right*. Therefore, whenever *up* is strictly optimal for a featurized reward function, we can permute that reward function's feature weights by swapping the x- and y-coefficients (α_1 and α_2, respectively). Again, this new reward function is featurized, and it makes going *right* strictly optimal. So the usual arguments ensure that at least half of these featurized reward functions make it optimal to go right.

But sometimes these similarities won't hold, even when it initially looks like they "should"!

Ye Olde Gridworld

Coordinate featurization, negative



In this environment, the agent can induce the feature vectors $\{(\begin{smallmatrix} x:-1 \\ y:0 \end{smallmatrix}), (\begin{smallmatrix} x:-1 \\ y:-1 \end{smallmatrix})\}$ if it goes *left*. However, it can induce $\{(\begin{smallmatrix} x:1 \\ y:0 \end{smallmatrix}), (\begin{smallmatrix} x:1 \\ y:1 \end{smallmatrix})\}$ if it goes *right*. There is no way of switching feature labels so as to copy the *left* feature set into the *right* feature set! There's no way to just apply a feature permutation to the *left* set, and thereby produce a subset of the *right* feature set. Therefore, the theorems don't apply, and so they don't guarantee anything about how most permutations of every reward function incentivize some kind of behavior.

On reflection, this makes sense. If $\alpha_1 = \alpha_2 = -1$, then there's no way the agent will want to go *right.* Instead, it'll go for the negative feature values offered by going *left*. This will hold for *all* permutations of this feature labelling, too. So the orbit-level incentives *can't* hold.

If the agent can be made to "hate everything" (all feature weights α_i are negative), then it will pursue opportunities which give it negative-valued feature vectors, or at least strive for the oblivion of the zero feature vector. Vice versa for if it positively values all features.

StarCraft II

Consider a deep RL training process, where the agent's episodic reward is featurized into a weighted sum of the different resources the agent has at the end of the game, with weight vector α. For simplicity, we fix an opponent policy and a learning regime (number of epochs, learning rate, hyperparameters, network architecture, and so on). We consider the effects of varying the reward feature coefficients α.

Outcomes of interest: Game state trajectories.

AI decision-making function: $f(T \mid \alpha)$ returns the probability that, given our fixed learning regime and reward feature vector α, the training process produces a policy network whose rollouts instantiate some trajectory $\tau \in T$.

What the theorems say:

1. If α is the zero vector, the agent gets the same reward for all trajectories, and so gradient descent does nothing, and the randomly initialized policy network quickly loses against any reasonable opponent. No power-seeking tendencies if this is the only plausible parameter setting.
2. If α only has negative entries, then the policy network quickly learns to throw away all of its resources and not collect any more. If and only if this has been achieved, the

training process is indifferent to whether the game is lost. No real power-seeking tendencies if it's only plausible that we specify a negative vector.
3. If α has a positive entry, then policies learn to gather as much of that resource as possible. In particular, there *aren't* orbit elements α with positive entries but where the learned policy tends to just die, and so we don't even have to check that the permuted variants $\phi \cdot \alpha$ of such feature vectors are also plausible. Power-seeking occurs.

This reasoning *depends on which kinds of feature weights are plausible*, and so wouldn't have been covered by the previous results.

Minecraft

Similar setup to StarCraft II, but now the agent's episode reward is $\alpha_1 \cdot$(Amount of iron ore in chests within 100 blocks of spawn after 2 in-game days)$+\alpha_2 \cdot$(Same but for coal), where $\alpha_1, \alpha_2 \in \mathbb{R}$ are scalars (together, they form the coefficient vector $\alpha \in \mathbb{R}^2$).

Outcomes of interest: Game state trajectories.

AI decision-making function: $f(T \mid \alpha)$ returns the probability that, given our fixed learning regime and feature coefficients α, the training process produces a policy network whose rollouts instantiate some trajectory $\tau \in T$.

What the theorems say:

1. If α is the zero vector, the analysis is the same as before. No power-seeking tendencies. In fact, the agent tends to *not* gain power because it has no optimization pressure steering it towards the few action sequences which gain the agent power.
2. If α only has negative entries, the agent definitely doesn't hoard resources in chests. Otherwise, there's no real reward signal and gradient descent doesn't do a whole lot due to sparsity.
3. If α has a positive entry, and if the learning process is good enough, agents tend to stay alive. If the learning process is good enough, there just won't be a single feature vector with a positive entry which tends to produce non-self-empowering policies.

The analysis so far is nice to make a bit more formally, but it isn't really pointing out anything that we couldn't have figured out pre-theoretically. I think I can sketch out more novel reasoning, but I'll leave that to a future post.

Beyond The Featurized Case

Consider some arbitrary set $D \subseteq \mathbb{R}^d$ of "plausible" utility functions over d outcomes. If we have the usual big set B of outcome lotteries (which possibilities are, in the view of this

theory, often attained via "power-seeking"), and B contains n copies of some smaller set A via environmental symmetries $\phi_1, \dots, \phi_n$, then when are there orbit-level incentives *within* D—when will most reasonable variants of utility functions make the agent more likely to select B rather than A?

When the environmental symmetries can be applied to the A-preferring-variants, in a way which produces another plausible objective. Slightly more formally, if, for every plausible utility function $u \in D$ where the agent has a greater chance of selecting A than of selecting B, we have the membership $\phi_i \cdot u \in D$ for all $i = 1, \dots, n$. (The formal result is Lemma B.7 in this Overleaf.)

This covers the totally general case of arbitrary sets of utility function classes we might use. (And, technically, "utility function" is decorative at this point—it just stands in for a parameter which we use to retarget the AI policy-production process.)

The general result highlights how D := { plausible objective functions } affects what conclusions we can draw about orbit-level incentives. All else equal, being able to specify more plausible objective functions for which $f(B \mid u) \geq f(A \mid u)$ means that we're more likely to to ensure closure under certain permutations. Similarly, adding plausible A-dispreferring objectives makes it harder to satisfy $f(B \mid u) < f(A \mid u) \implies \phi_i \cdot u \in D$, which makes it harder to ensure closure under certain permutations, which makes it harder to prove instrumental convergence.

Revisiting How The Environment Structure Affects Power-Seeking Incentive Strength

In *Seeking Power is Convergently Instrumental in a Broad Class of Environments*, I wrote:

> Structural assumptions on utility really do matter when it comes to instrumental convergence:
>
> 1. **u-AOH (utility functions over action-observation histories):** No IC
> 2. **u-OH (utility functions over observation histories):** Strong IC
> 3. **State-based objectives (eg state-based reward in MDPs):** Moderate IC
>
> Environmental structure can cause instrumental convergence, but (the absence of) structural assumptions on utility can make instrumental convergence go away (for optimal agents).

In particular, for the MDP case, I wrote:

> MDPs assume that utility functions have a lot of structure: the utility of a history is time-discounted additive over observations. Basically, $u(a_1 o_1 a_2 o_2 \dots) = \sum_{t=1}^{\infty} \gamma^{t-1} R(o_t)$, for

some $\gamma \in [0, 1)$ and reward function $R : O \to \mathbb{R}$ over observations. And because of this structure, the agent's average per-timestep reward is controlled by the last observation it sees. There are exponentially fewer last observations than there are *observation histories.* Therefore, in this situation, instrumental convergence is exponentially weaker for reward functions than for arbitrary u-OH.

This is equivalent to a featurization which takes in an action-observation history, ignores the actions, and spits out time-discounted observation counts. The utility function is then over observations (which are just states in the MDP case). Here, the symmetries can only be over states, and not histories, and no matter how expressive the plausible state-based-reward-set D_S is, it can't compete with the exponentially larger domain of the observation-history-based-utility-set D_{OH}, and so the featurization has *limited how strong instrumental convergence can get* by projecting the high-dimensional u-OH into the lower-dimensional u-State.

But when we go from u-AOH to u-OH, we're throwing away even more information—information about the actions! This is also a sparse projection. So what's up?

When we throw away info about actions, we're breaking some symmetries which made instrumental convergence disappear in the u-AOH case. In any deterministic environment, there are equally many u-AOH which make me want to go e.g. left (and, say, die) as which make me want to go right (and survive). This is guaranteed by symmetries which swap the value of an optimal AOH with the value of an AOH going the other way:

> If the agent cares not about its own action histories, but about its observation histories, there are just more ways to care about going *right* and being alive! Twice as many ways, in fact!

But when we restrict the utility function to not care about actions, now you can only modify how it cares about observation histories. Here, the AOH environmental symmetry ϕ_{AOH} which previously ensured balanced statistical incentives, no longer enjoys closure under D_{OH}, and so the restricted plausible set theorem no longer works, and instrumental convergence appears when restricting from u-AOH to u-OH.

I thank Justis Mills for feedback on a draft.

Appendix: tracking key limitations of the power-seeking theorems

From last time:

1. The results aren't first-person: They don't deal with the agent's uncertainty about what environment it's in.
2. Not all environments have the right symmetries

    - But most ones we think about seem to
3. ~~Don't account for the ways in which we might practically express reward functions.~~ (This limitation was handled by this post.)

I think it's reasonably clear how to apply the results to realistic objective functions. I also think our objective specification procedures are quite expressive, and so the closure condition will hold and the results go through in the appropriate situations.

1. ^

   It's not hard to have this many degrees of freedom in such a small toy environment, but the toy environment is pedagogical. It's practically impossible to have full degrees of freedom in an environment with a trillion states.

2. ^

   "At least", and not "exactly." If α is a constant feature vector, it's optimal to go right for every permutation of α (trivially so, since α's orbit has a single element—itself).

3. ^

   Even under my more aggressive conjecture about "fractional terminal state copy containment", the unfeaturized situation would only guarantee $\frac{3}{5}$-strength orbit incentives, strictly weaker than $\frac{3}{4}$-strength.

4. ^

   Certain trivial featurizations can decrease the strength of power-seeking tendencies, too. For example, if the featurization is 2-dimensional:

   $$\begin{pmatrix} 1 \text{ if the agent is dead, 0 otherwise} \\ 1 \text{ if the agent is alive, 0 otherwise} \end{pmatrix}$$

   , this will tend to produce 1:1 survive/die orbit-level incentives, whereas the incentives for raw reward functions may be 1,000:1 or stronger.

5. ^

   There's something abstraction-adjacent about this result (proposition D.1 in the linked Overleaf paper). The result says something like "do the grooves of the agent's world model featurization, respect the grooves of symmetries in the structure of the agent's environment?", and if they do, *bam*, sufficient condition for power-seeking under the featurized model. I think there's something important here about how good world-model-featurizations should work, but I'm not sure what that is yet.

   I do know that "the featurization should commute with the environmental symmetry" is something I'd thought—in basically those words—no fewer than 3 times, as early as summer_{2021}, without explicitly knowing what that should even *mean*.

6. ^

   Lemma B.7 in this Overleaf—compile `quantitative-paper.tex`.