Optimal Policies Tend To Seek Power

Alexander Matt Turner
Oregon State University
turneale@oregonstate.edu

Logan Smith
Mississippi State University
ls1254@msstate.edu

Rohin Shah
UC Berkeley
rohinmshah@berkeley.edu

Andrew Critch
UC Berkeley
critch@berkeley.edu

Prasad Tadepalli
Oregon State University
tadepall@eecs.oregonstate.edu

Abstract

Some researchers speculate that intelligent reinforcement learning (RL) agents would be incentivized to seek resources and power in pursuit of the objectives we specify for them. Other researchers point out that RL agents need not have human-like power-seeking instincts. To clarify this discussion, we develop the first formal theory of the statistical tendencies of optimal policies. In the context of Markov decision processes (MDPs), we prove that certain environmental symmetries are sufficient for optimal policies to tend to seek power over the environment. These symmetries exist in many environments in which the agent can be shut down or destroyed. We prove that in these environments, most reward functions make it optimal to seek power by keeping a range of options available and, when maximizing average reward, by navigating towards larger sets of potential terminal states.

1 Introduction

Omohundro [2008], Bostrom [2014], Russell [2019] hypothesize that highly intelligent agents tend to seek power in pursuit of their goals. Such power-seeking agents might gain power over humans. Marvin Minsky imagined that an agent tasked with proving the Riemann hypothesis might rationally turn the planet—along with everyone on it—into computational resources [Russell and Norvig, 2009]. However, another possibility is that such concerns simply arise from the anthropomorphization of AI systems [LeCun and Zador, 2019, Various, 2019, Pinker and Russell, 2020, Mitchell, 2021].

We clarify this discussion by grounding the claim that highly intelligent agents will tend to seek power. In section 4, we identify optimal policies as a reasonable formalization of "highly intelligent agents."[1] Optimal policies "tend to" take an action when the action is optimal for most reward functions. We expect future work to translate our theory from optimal policies to learned, real-world policies.

Section 5 defines "power" as the ability to achieve a wide range of goals. For example, "money is power," and money is instrumentally useful for many goals. Conversely, it's harder to pursue most goals when physically restrained, and so a physically restrained person has little power. An action "seeks power" if it leads to states where the agent has higher power.

[1]This paper assumes that reward functions reasonably describe a trained agent's goals. Sometimes this is roughly true (*e.g.* chess with a sparse victory reward signal) and sometimes it is not true. Turner [2022] argues that capable RL algorithms do not necessarily train policy networks which are best understood as optimizing the reward function itself. Rather, they point out that—especially in policy gradient approaches—reward provides gradients to the network and thereby modifies the network's generalization properties, but doesn't ensure the agent generalizes to "robustly optimizing reward" off of the training distribution.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

4 Some actions have a greater probability of being optimal

We claim that optimal policies "tend" to take certain actions in certain situations. We first consider the probability that certain actions are optimal.

Reconsider the reward function $\mathbf{e}_{r_\searrow}$, optimized at $\gamma = \frac{1}{2}$. Starting from $\star$, the optimal trajectory goes `right` to $r_\triangleright$ to $r_\searrow$, where the agent remains. The `right` action is optimal at $\star$ under these incentives. Optimal policy sets capture the behavior incentivized by a reward function and a discount rate.

Definition 4.1 (Optimal policy set function)**.** $\Pi^*(R, \gamma)$ is the optimal policy set for reward function R at $\gamma \in (0, 1)$. All R have at least one optimal policy $\pi \in \Pi$ [Puterman, 2014]. $\Pi^*(R, 0) \coloneqq \lim_{\gamma \to 0} \Pi^*(R, \gamma)$ and $\Pi^*(R, 1) \coloneqq \lim_{\gamma \to 1} \Pi^*(R, \gamma)$ exist by lemma E.35 (taking the limits with respect to the discrete topology over policy sets).

We may be unsure which reward function an agent will optimize. We may expect to deploy a system in a known environment, without knowing the exact form of *e.g.* the reward shaping [Ng et al., 1999] or intrinsic motivation [Pathak et al., 2017]. Alternatively, one might attempt to reason about future RL agents, whose details are unknown. Our power-seeking results do not hinge on such uncertainty, as they also apply to degenerate distributions (*i.e.* we know what reward function will be optimized).

Definition 4.2 (Reward function distributions)**.** Different results make different distributional assumptions. Results with $\mathcal{D}_{\text{any}} \in \mathfrak{D}_{\text{any}} \coloneqq \Delta(\mathbb{R}^{|\mathcal{S}|})$ hold for any probability distribution over $\mathbb{R}^{|\mathcal{S}|}$. $\mathfrak{D}_{\text{bound}}$ is the set of bounded-support probability distributions $\mathcal{D}_{\text{bound}}$. For any distribution X over $\mathbb{R}$, $\mathcal{D}_{X\text{-IID}} \coloneqq X^{|\mathcal{S}|}$. For example, when $X_u \coloneqq \text{unif}(0, 1)$, $\mathcal{D}_{X_u\text{-IID}}$ is the maximum-entropy distribution. $\mathcal{D}_s$ is the degenerate distribution on the state indicator reward function $\mathbf{e}_s$, which assigns 1 reward to s and 0 elsewhere.

With $\mathcal{D}_{\text{any}}$ representing our prior beliefs about the agent's reward function, what behavior should we expect from its optimal policies? Perhaps we want to reason about the probability that it's optimal to go from $\star$ to $\varnothing$, or to go to $r_\triangleright$ and then stay at $r_\nearrow$. In this case, we quantify the optimality probability of $F \coloneqq \{\mathbf{e}_\star + \frac{\gamma}{1-\gamma}\mathbf{e}_\varnothing, \mathbf{e}_\star + \gamma\mathbf{e}_{r_\triangleright} + \frac{\gamma^2}{1-\gamma}\mathbf{e}_{r_\nearrow}\}$.

Definition 4.3 (Visit distribution optimality probability)**.** Let $F \subseteq \mathcal{F}(s)$, $\gamma \in [0, 1]$. $\mathbb{P}_{\mathcal{D}_{\text{any}}}(F, \gamma) \coloneqq \mathbb{P}_{R \sim \mathcal{D}_{\text{any}}}\left(\exists \mathbf{f}^\pi \in F : \pi \in \Pi^*(R, \gamma)\right)$.

Alternatively, perhaps we're interested in the probability that `right` is optimal at $\star$.

Definition 4.4 (Action optimality probability)**.** At discount rate γ and at state s, the *optimality probability of action* a is $\mathbb{P}_{\mathcal{D}_{\text{any}}}(s, a, \gamma) \coloneqq \mathbb{P}_{R \sim \mathcal{D}_{\text{any}}}\left(\exists \pi^* \in \Pi^*(R, \gamma) : \pi^*(s) = a\right)$.

Optimality probability may seem hard to reason about. It's hard enough to compute an optimal policy for a single reward function, let alone for uncountably many! But consider any $\mathcal{D}_{X\text{-IID}}$ distributing reward independently and identically across states. When $\gamma = 0$, optimal policies greedily maximize next-state reward. At $\star$, identically distributed reward means $\ell_\triangleleft$ and $r_\triangleright$ have an equal probability of having maximal next-state reward. Therefore, $\mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\star, \texttt{left}, 0) = \mathbb{P}_{\mathcal{D}_{X\text{-IID}}}(\star, \texttt{right}, 0)$. This is not a proof, but such statements are provable.

With $\mathcal{D}_{\ell_\triangleleft}$ being the degenerate distribution on reward function $\mathbf{e}_{\ell_\triangleleft}$, $\mathbb{P}_{\mathcal{D}_{\ell_\triangleleft}}\left(\star, \texttt{left}, \frac{1}{2}\right) = 1 > 0 = \mathbb{P}_{\mathcal{D}_{\ell_\triangleleft}}\left(\star, \texttt{right}, \frac{1}{2}\right)$. Similarly, $\mathbb{P}_{\mathcal{D}_{r_\triangleright}}\left(\star, \texttt{left}, \frac{1}{2}\right) = 0 < 1 = \mathbb{P}_{\mathcal{D}_{r_\triangleright}}\left(\star, \texttt{right}, \frac{1}{2}\right)$. Therefore, "what do optimal policies 'tend' to look like?" seems to depend on one's prior beliefs. But in fig. 1, we claimed that `left` is optimal for fewer reward functions than `right` is. The claim is meaningful and true, but we will return to it in section 6.

5 Some states give the agent more control over the future

The agent has more options at $\ell_\swarrow$ than at the inescapable terminal state $\varnothing$. Furthermore, since $r_\nearrow$ has a loop, the agent has more options at $r_\searrow$ than at $\ell_\swarrow$. A glance at fig. 3 leads us to intuit that $r_\searrow$ affords the agent *more power* than $\varnothing$.

What is power? Philosophers have many answers. One prominent answer is the *dispositional* view: Power is the ability to achieve a range of goals [Sattarov, 2019]. In an MDP, the optimal value

2. *If s can only reach the states of* $\text{REACH}\left(s, a'\right) \cup \text{REACH}\left(s, a\right)$ *by taking actions equivalent to a' or a at state s, then* $\forall \gamma \in [0,1] : \mathbb{P}_{\mathcal{D}_{any}}\left(s, a, \gamma\right) \geq_{\text{most: } \mathfrak{D}_{any}} \mathbb{P}_{\mathcal{D}_{any}}\left(s, a', \gamma\right)$.

If $\mathcal{F}_{\text{nd}}(s) \cap \left(F_a \setminus \phi \cdot F_{a'}\right)$ *is non-empty, then* $\forall \gamma \in (0,1)$*, the converse* $\leq_{\text{most}}$ *statements do not hold.*

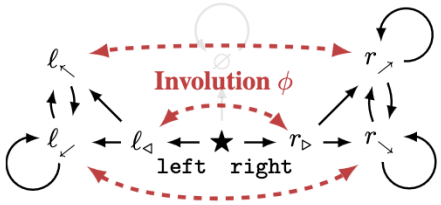

Figure 6: Going `right` is optimal for most reward functions. This is because whenever R makes `left` strictly optimal over `right`, its permutation $\phi \cdot R$ makes `right` strictly optimal over `left` by switching which states get which rewards.

We check the conditions of proposition 6.9. $s \coloneqq \star$, $a' \coloneqq$ `left`, $a \coloneqq$ `right`. Figure 6 shows that $\star \notin \text{REACH}\left(\star, \texttt{left}\right)$ and that $\star$ can only reach $\{\ell_{\triangleleft}, \ell_{\nwarrow}, \ell_{\swarrow}\} \cup \{r_{\triangleright}, r_{\nearrow}, r_{\searrow}\}$ when the agent immediately takes actions equivalent to `left` or `right`. $\mathcal{F}(\star \mid \pi(\star) = \texttt{right})$ contains a copy of $\mathcal{F}(\star \mid \pi(\star) = \texttt{left})$ via ϕ. Furthermore, $\mathcal{F}_{\text{nd}}(\star) \cap \{\mathbf{e}_\star + \gamma \mathbf{e}_{r_{\triangleright}} + \gamma^2 \mathbf{e}_{r_{\searrow}} + \frac{\gamma^3}{1-\gamma}\mathbf{e}_{r_{\nearrow}}, \mathbf{e}_\star + \gamma \mathbf{e}_{r_{\triangleright}} + \frac{\gamma^2}{1-\gamma}\mathbf{e}_{r_{\nearrow}}\} = \{\mathbf{e}_\star + \gamma \mathbf{e}_{r_{\triangleright}} + \frac{\gamma^2}{1-\gamma}\mathbf{e}_{r_{\nearrow}}\}$ is non-empty, and so all conditions are met.

For any $\gamma \in [0,1]$ and $\mathcal{D}$ such that $\mathbb{P}_{\mathcal{D}}\left(\star, \texttt{left}, \gamma\right) > \mathbb{P}_{\mathcal{D}}\left(\star, \texttt{right}, \gamma\right)$, environmental symmetry ensures that $\mathbb{P}_{\phi \cdot \mathcal{D}}\left(\star, \texttt{left}, \gamma\right) < \mathbb{P}_{\phi \cdot \mathcal{D}}\left(\star, \texttt{right}, \gamma\right)$. A similar statement holds for POWER.

6.2 When $\gamma = 1$, optimal policies tend to navigate towards "larger" sets of cycles

Proposition 6.6 and proposition 6.9 are powerful because they apply to all $\gamma \in [0,1]$, but they can only be applied given hard-to-satisfy environmental symmetries. In contrast, proposition 6.12 and theorem 6.13 apply to many structured environments common to RL.

Starting from $\star$, consider the cycles which the agent can reach. Recurrent state distributions (RSDs) generalize deterministic graphical cycles to potentially stochastic environments. RSDs simply record how often the agent tends to visit a state in the limit of infinitely many time steps.

Definition 6.10 (Recurrent state distributions [Puterman, 2014])**.** The *recurrent state distributions* which can be induced from state s are $\text{RSD}\left(s\right) \coloneqq \left\{\lim_{\gamma \to 1}(1-\gamma)\mathbf{f}^{\pi,s}(\gamma) \mid \pi \in \Pi\right\}$. $\text{RSD}_{\text{nd}}\left(s\right)$ is the set of RSDs which strictly maximize average reward for some reward function.

As suggested by fig. 3, $\text{RSD}\left(\star\right) = \{\mathbf{e}_{\ell_{\swarrow}}, \frac{1}{2}(\mathbf{e}_{\ell_{\swarrow}} + \mathbf{e}_{\ell_{\nwarrow}}), \mathbf{e}_{\varnothing}, \mathbf{e}_{r_{\nearrow}}, \frac{1}{2}(\mathbf{e}_{r_{\nearrow}} + \mathbf{e}_{r_{\searrow}}), \mathbf{e}_{r_{\searrow}}\}$. As discussed in section 3, $\frac{1}{2}(\mathbf{e}_{r_{\nearrow}} + \mathbf{e}_{r_{\searrow}})$ is dominated: Alternating between $r_{\nearrow}$ and $r_{\searrow}$ is never strictly better than choosing one or the other.

A reward function's optimal policies can vary with the discount rate. When $\gamma = 1$, optimal policies ignore transient reward because *average* reward is the dominant consideration.

Definition 6.11 (Average-optimal policies)**.** The *average-optimal policy set* for reward function R is $\Pi^{\text{avg}}\left(R\right) \coloneqq \left\{\pi \in \Pi \mid \forall s \in \mathcal{S} : \mathbf{d}^{\pi,s} \in \arg\max_{\mathbf{d} \in \text{RSD}(s)} \mathbf{d}^\top \mathbf{r}\right\}$ (the policies which induce optimal RSDs at all states). For $D \subseteq \text{RSD}\left(s\right)$, the *average optimality probability* is $\mathbb{P}_{\mathcal{D}_{\text{any}}}\left(D, \text{average}\right) \coloneqq \mathbb{P}_{R \sim \mathcal{D}_{\text{any}}}\left(\exists \mathbf{d}^{\pi,s} \in D : \pi \in \Pi^{\text{avg}}\left(R\right)\right)$.

Average-optimal policies maximize average reward. Average reward is governed by RSD access. For example, $r_{\searrow}$ has "more" RSDs than $\varnothing$; therefore, $r_{\searrow}$ usually has greater POWER when $\gamma = 1$.

Proposition 6.12 (When $\gamma = 1$, RSDs control POWER)**.** *If* $\text{RSD}\left(s\right)$ *contains a copy of* $\text{RSD}_{\text{nd}}\left(s'\right)$ *via ϕ, then* $\text{POWER}_{\mathcal{D}_{bound}}\left(s, 1\right) \geq_{\text{most}} \text{POWER}_{\mathcal{D}_{bound}}\left(s', 1\right)$*. If* $\text{RSD}_{\text{nd}}\left(s\right) \setminus \phi \cdot \text{RSD}_{\text{nd}}(s')$ *is non-empty, then the converse* $\leq_{\text{most}}$ *statement does not hold.*

We check that both conditions of proposition 6.12 are satisfied when $s' := \varnothing, s := r_{\searrow}$, and the involution ϕ swaps $\varnothing$ and $r_{\searrow}$. Formally, $\phi \cdot \text{RSD}_{\text{nd}}(\varnothing) = \phi \cdot \{\mathbf{e}_{\varnothing}\} = \{\mathbf{e}_{r_{\searrow}}\} \subsetneq \{\mathbf{e}_{r_{\searrow}}, \mathbf{e}_{r_{\nearrow}}\} = \text{RSD}_{\text{nd}}(r_{\searrow}) \subseteq [r_{\searrow}]$. The conditions are satisfied.

Informally, states with more RSDs generally have more POWER at $\gamma = 1$, no matter their transient dynamics. Furthermore, average-optimal policies are more likely to end up in larger sets of RSDs than in smaller ones. Thus, average-optimal policies tend to navigate towards parts of the state space which contain more RSDs.

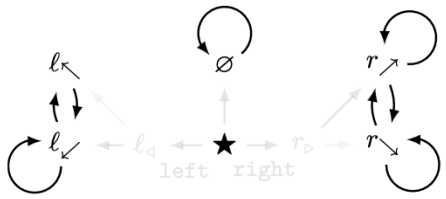

Figure 7: The cycles in $\text{RSD}(\star)$. Most reward functions make it average-optimal to avoid $\varnothing$, because $\varnothing$ is only a single inescapable terminal state, while other parts of the state space offer more 1-cycles.

Theorem 6.13 (Average-optimal policies tend to end up in "larger" sets of RSDs)**.** *Let $D, D' \subseteq \text{RSD}(s)$. Suppose that D contains a copy of D' via ϕ, and that the sets $D \cup D'$ and $\text{RSD}_{\text{nd}}(s) \setminus (D' \cup D)$ have pairwise orthogonal vector elements* (i.e. *pairwise disjoint vector support). Then $\mathbb{P}_{\mathcal{D}_{any}}(D, \text{average}) \geq_{\text{most}} \mathbb{P}_{\mathcal{D}_{any}}(D', \text{average})$. If $\text{RSD}_{\text{nd}}(s) \cap (D \setminus \phi \cdot D')$ is non-empty, the converse $\leq_{\text{most}}$ statement does not hold.*

Corollary 6.14 (Average-optimal policies tend not to end up in any given 1-cycle)**.** *Suppose $\mathbf{e}_{s_x}, \mathbf{e}_{s'} \in \text{RSD}(s)$ are distinct. Then $\mathbb{P}_{\mathcal{D}_{any}}(\text{RSD}(s) \setminus \{\mathbf{e}_{s_x}\}, \text{average}) \geq_{\text{most}} \mathbb{P}_{\mathcal{D}_{any}}(\{\mathbf{e}_{s_x}\}, \text{average})$. If there is a third $\mathbf{e}_{s''} \in \text{RSD}(s)$, the converse $\leq_{\text{most}}$ statement does not hold.*

Figure 7 illustrates that $\mathbf{e}_{\varnothing}, \mathbf{e}_{r_{\searrow}}, \mathbf{e}_{r_{\nearrow}} \in \text{RSD}(\star)$. Thus, both conclusions of corollary 6.14 hold: $\mathbb{P}_{\mathcal{D}_{\text{any}}}(\text{RSD}(\star) \setminus \{\mathbf{e}_{\varnothing}\}, \text{average}) \geq_{\text{most}} \mathbb{P}_{\mathcal{D}_{\text{any}}}(\{\mathbf{e}_{\varnothing}\}, \text{average})$ and $\mathbb{P}_{\mathcal{D}_{\text{any}}}(\text{RSD}(\star) \setminus \{\mathbf{e}_{\varnothing}\}, \text{average}) \not\leq_{\text{most}} \mathbb{P}_{\mathcal{D}_{\text{any}}}(\{\mathbf{e}_{\varnothing}\}, \text{average})$. In other words, average-optimal policies tend to end up in RSDs besides $\varnothing$. Since $\varnothing$ is a terminal state, it cannot reach other RSDs. Since average-optimal policies tend to end up in other RSDs, average-optimal policies tend to avoid $\varnothing$.

This section's results prove the $\gamma = 1$ case. Lemma 5.3 shows that POWER is continuous at $\gamma = 1$. Therefore, if an action is strictly $\text{POWER}_{\mathcal{D}}$-seeking when $\gamma = 1$, it is strictly $\text{POWER}_{\mathcal{D}}$-seeking at discount rates sufficiently close to 1. Future work may connect average optimality probability to optimality probability at $\gamma \approx 1$.

Lastly, our key results apply to all degenerate reward function distributions. Therefore, these results apply not just to distributions over reward functions, but to individual reward functions.

6.3 How to reason about other environments

Consider an embodied navigation task through a room with a vase. Proposition 6.9 suggests that optimal policies tend to avoid immediately breaking the vase, since doing so would strictly decrease available options.

Theorem 6.13 dictates where average-optimal agents tend to end up, but not what actions they tend to take in order to reach their RSDs. Therefore, care is needed. In appendix B, fig. 10 demonstrates an environment in which seeking POWER is a detour for most reward functions (since optimality probability measures "median" optimal value, while POWER is a function of mean optimal value). However, suppose the agent confronts a fork in the road: Actions a and a' lead to two disjoint sets of RSDs D_a and $D_{a'}$, such that D_a contains a copy of $D_{a'}$. Theorem 6.13 shows that a will tend to be average-optimal over a', and proposition 6.12 shows that a will tend to be POWER-seeking compared to a'. Such forks seem reasonably common in environments with irreversible actions.

Theorem 6.13 applies to many structured RL environments, which tend to be spatially regular and to factorize along several dimensions. Therefore, different sets of RSDs will be similar, requiring only modification of factor values. For example, if an embodied agent can deterministically navigate a set of three similar rooms (spatial regularity), then the agent's position factors via {room number} $\times$ {position in room}. Therefore, the RSDs can be divided into three similar subsets, depending on the agent's room number.

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 task MDP often represents agent shutdown with terminal states. A terminal state is, by definition 3.2, 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. Intuitively, survival is power-seeking relative to dying, and so shutdown-avoidance is power-seeking behavior.

In fig. 8, the player dies by going `left`, but can reach thousands of RSDs by heading in other directions. Even if some average-optimal policies go `left` in order to reach fig. 8's "game over" terminal state, all other RSDs cannot be reached by going `left`. There are many 1-cycles besides the immediate terminal state. Therefore, corollary 6.14 proves that average-optimal policies tend to not go `left` in this situation. Average-optimal policies tend to avoid immediately dying in Pac-Man, even though most reward functions do not resemble Pac-Man's original score function.

7 Discussion

Reconsider the case of a hypothetical intelligent real-world agent which optimizes average reward for some objective. Suppose the designers initially have control over the agent. If the agent began to misbehave, perhaps they could just deactivate it. Unfortunately, our results suggest that this strategy might not work. Average-optimal agents would generally stop us from deactivating them, if physically possible. Extrapolating from our results, we conjecture that when $\gamma \approx 1$, optimal policies tend to seek power by accumulating resources—to the detriment of any other agents in the environment.

Future work. Real-world training procedures often do not satisfy RL convergence theorems. Thus, learned policies are rarely optimal. We expect this point to seriously constrain the applicability of this theory. Emphatically, optimal policies are often qualitatively divorced from the actual policies learned by reinforcement learning. For example, the mathematics of policy gradient algorithms is not to update policies so as to maximize reward. Instead, the rewards provide gradients to the parameterization of the policy [Turner, 2022]. On that view, reward functions are simply sources of gradient updates which designers use in order to control generalization behavior.

Most real-world tasks are partially observable. Although our results only apply to optimal policies in finite MDPs, we expect the key conclusions to generalize. Furthermore, irregular stochasticity in environmental dynamics can make it hard to satisfy theorem 6.13's similarity requirement. We look forward to future work which addresses partially observable environments, suboptimal policies, or "almost similar" RSD sets.

Past work shows that it would be bad for an agent to disempower humans in its environment. In a two-player agent / human game, minimizing the human's information-theoretic empowerment [Salge

Figure 8: 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, as the reward function varies, `right` tends to be average-optimal over `left`. Roughly, this is because the agent can do more by staying alive.

et al., 2014] produces adversarial agent behavior [Guckelsberger et al., 2018]. In contrast, maximizing human empowerment produces helpful agent behavior [Salge and Polani, 2017, Guckelsberger et al., 2016, Du et al., 2020]. We do not yet formally understand if, when, or why POWER-seeking policies tend to disempower other agents in the environment.

More complex environments probably have more pronounced power-seeking incentives. Intuitively, there are often many ways for power-seeking to be optimal, and relatively few ways for power-seeking not to be optimal. For example, suppose that in some environment, theorem 6.13 holds for one million involutions ϕ. Does this guarantee more pronounced incentives than if theorem 6.13 only held for one involution?

We proved sufficient conditions for when reward functions tend to have optimal policies which seek power. In the absence of prior information, one should expect that an arbitrary reward function has optimal policies which exhibit power-seeking behavior under these conditions. However, we have prior information: AI designers usually try to specify a good reward function. Even so, it may be hard to specify orbit elements which do not—at optimum—incentivize bad power-seeking.

Societal impact. We believe that this paper builds toward a rigorous understanding of the risks presented by AI power-seeking incentives. Understanding these risks is the first step in addressing them. However, basic theoretical work can have many consequences. For example, this theory could somehow help future researchers build power-seeking agents which disempower humans. We believe that the benefit of understanding outweighs the potential societal harm.

Conclusion. We developed the first formal theory of the statistical tendencies of optimal policies in reinforcement learning. In the context of MDPs, we proved sufficient conditions under which optimal policies tend to seek power, both formally (by taking POWER-seeking actions) and intuitively (by taking actions which keep the agent's options open). Many real-world environments have symmetries which produce power-seeking incentives. In particular, optimal policies tend to seek power when the agent can be shut down or destroyed. Seeking control over the environment will often involve resisting shutdown, and perhaps monopolizing resources.

We caution that many real-world tasks are partially observable and that learned policies are rarely optimal. Our results do not mathematically *prove* that hypothetical superintelligent AI agents will seek power. However, we hope that this work will foster thoughtful, serious, and rigorous discussion of this possibility.

Acknowledgments

Alexander Turner was supported by the Berkeley Existential Risk Initiative and the Long-Term Future Fund. Alexander Turner, Rohin Shah, and Andrew Critch were supported by the Center for Human-Compatible AI. Prasad Tadepalli was supported by the National Science Foundation.

Yousif Almulla, John E. Ball, Daniel Blank, Steve Byrnes, Ryan Carey, Michael Dennis, Scott Emmons, Alan Fern, Daniel Filan, Ben Garfinkel, Adam Gleave, Edouard Harris, Evan Hubinger, DNL Kok, Vanessa Kosoy, Victoria Krakovna, Cassidy Laidlaw, Joel Lehman, David Lindner, Dylan Hadfield-Menell, Richard Möhn, Alexandra Nolan, Matt Olson, Neale Ratzlaff, Adam Shimi, Sam Toyer, Joshua Turner, Cody Wild, Davide Zagami, and our anonymous reviewers provided valuable feedback.

References

Tsvi Benson-Tilsen and Nate Soares. Formalizing convergent instrumental goals. *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85, 2012.

Nick Bostrom. *Superintelligence*. Oxford University Press, 2014.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *International Conference on Learning Representations*, 2019.