

# Power-seeking can be probable and predictive for trained agents

Victoria Krakovna<sup>1</sup> and Janos Kramar<sup>1</sup>

<sup>1</sup>DeepMind

Power-seeking behavior is a key source of risk from advanced AI, but our theoretical understanding of this phenomenon is relatively limited. Building on existing theoretical results demonstrating power-seeking incentives for most reward functions, we investigate how the training process affects power-seeking incentives and show that they are still likely to hold for trained agents under some simplifying assumptions. We formally define the training-compatible goal set (the set of goals consistent with the training rewards) and assume that the trained agent learns a goal from this set. In a setting where the trained agent faces a choice to shut down or avoid shutdown in a new situation, we prove that the agent is likely to avoid shutdown. Thus, we show that power-seeking incentives can be probable (likely to arise for trained agents) and predictive (allowing us to predict undesirable behavior in new situations).

## 1. Introduction

Power-seeking behavior is a major source of risk from advanced AI and a key element of many threat models in AI alignment [Carlsmith, 2022, Cotra, 2022, Ngo, 2022]. Existing theoretical results [Turner et al., 2021, Turner and Tadepalli, 2022] show that most reward functions incentivize reinforcement learning agents to take power-seeking actions. This is concerning, but does not immediately imply that a trained agent will seek power, since it may not learn to optimize the training reward [Turner, 2022], and the goals the agent learns are not chosen at random from the set of all possible rewards, but are shaped by the training process to reflect our preferences. In this work, we investigate how the training process affects power-seeking incentives and show that they are still likely to hold for trained agents under some assumptions (e.g. that the agent learns a goal during the training process).

Suppose an agent is trained using reinforcement learning with reward function  $\theta^*$ . We assume that the agent learns a *goal* during the training process: a set of internal representations of favored and disfavored outcomes (state features), as defined in Ngo [2022]. For simplicity, we assume this is equivalent to learning a reward function, which is not necessarily the same as the training reward function  $\theta^*$ . We consider the set of reward functions that are consistent with the training rewards received by the agent, in the sense that the agent’s behavior on the training data is optimal for these reward functions. We call this the *training-compatible goal set*, and we expect that the agent is likely to learn a reward function from this set.

We make another simplifying assumption that the training process will randomly select a goal for the agent to learn that is consistent with the training rewards, i.e. uniformly drawn from the training-compatible goal set. Then we will argue that the power-seeking results apply under these conditions, and thus are useful for predicting undesirable behavior by the trained agent in new situations. We aim to show that power-seeking incentives can be probable (likely to arise for trained agents) and predictive (allowing us to predict undesirable behavior in new situations) [Shah, 2023].

We will begin by reviewing some necessary definitions and results from the power-seeking literature in Section 2. We formally define the training-compatible goal set and give an example in the CoinRun environment in Section 3. Then in Section 4 we consider a setting where the trained agent faces a

choice to shut down or avoid shutdown in a new situation, and apply the power-seeking result to the training-compatible goal set to show that the agent is likely to avoid shutdown.

To satisfy the conditions of the power-seeking theorem, we show that the agent can be retargeted away from shutdown without affecting rewards received on the training data (Theorem 2). This can be done by switching the rewards of the shutdown state and a reachable recurrent state, as the recurrent state can provide repeated rewards, while the shutdown state provides less reward since it can only be visited once, assuming a high enough discount factor (Proposition 3). As the discount factor increases, more recurrent states can be retargeted to, which implies that a higher proportion of training-comptatible goals leads to avoiding shutdown in a new situation.

## 2. Preliminaries from the power-seeking literature

We will use definitions and results from the paper “Parametrically retargetable decision-makers tend to seek power” (here abbreviated as RDSP) [Turner and Tadepalli, 2022], with notation and explanations modified as needed for our purposes.

### Notation and assumptions:

- The environment is an MDP with finite state space  $\mathcal{S}$ , finite action space  $\mathcal{A}$ , and discount rate  $\gamma$ .
- Let  $\theta$  be a  $d$ -dimensional state reward vector, where  $d$  is the size of the state space  $\mathcal{S}$ , and let  $\Theta$  be a set of reward vectors.
- Let  $r^\theta(s)$  be the reward assigned by  $\theta$  to state  $s$ .
- Let  $A_0, A_1$  be disjoint action sets.
- Let  $f$  be an algorithm that produces an optimal policy  $f(\theta)$  on the training data given rewards  $\theta$ , and let  $f_s(A_i|\theta)$  be the probability that this policy chooses an action from set  $A_i$  in a given state  $s$ .

**Definition 1 (Orbit of a reward vector - Def 3.1 in RDSP).** Let  $S_d$  be the symmetric group consisting of all permutations of  $d$  items. The orbit of  $\theta$  inside  $\Theta$  is the set of all permutations of the entries of  $\theta$  that are also in  $\Theta$ :  $\text{Orbit}_\Theta(\theta) := (S_d \cdot \theta) \cap \Theta$ .

**Definition 2 (Orbit subset where an action set is preferred - from Def 3.5 in RDSP).** Let

$$\text{Orbit}_{\Theta,s,A_i > A_j}(\theta) := \{\theta' \in \text{Orbit}_\Theta(\theta) | f_s(A_i|\theta') > f_s(A_j|\theta')\}.$$

This is the subset of  $\text{Orbit}_\Theta(\theta)$  that results in  $f_s$  choosing  $A_i$  over  $A_j$ .

**Definition 3 (Preference for an action set  $A_1$  - Def 3.2 in RDSP).** The function  $f_s$  chooses action set  $A_1$  over  $A_0$  for the  $n$ -majority of elements  $\theta$  in each orbit, denoted as  $f_s(A_1|\theta) \geq_{\text{most:}\Theta}^n f_s(A_0|\theta)$ , iff the following inequality holds for all  $\theta \in \Theta$ :

$$|\text{Orbit}_{\Theta,s,A_1 > A_0}(\theta)| \geq n |\text{Orbit}_{\Theta,s,A_0 > A_1}(\theta)|$$

**Definition 4 (Multiply retargetable function from  $A_0$  to  $A_1$  - Def 3.5 in RDSP).** The function  $f_s$  is a multiply retargetable function from  $A_0$  to  $A_1$  if there are multiple permutations of rewards that would change the choice made by  $f_s$  from  $A_0$  to  $A_1$ . Specifically,  $f_s$  is a  $(\Theta, A_0 \xrightarrow{n} A_1)$ -retargetable function iff for each  $\theta \in \Theta$ , we can choose a set of permutations  $\Phi = \{\phi_1, \dots, \phi_n\}$  that satisfy the following conditions:

1. Retargetability:  $\forall \phi \in \Phi$  and  $\forall \theta' \in \text{Orbit}_{\Theta,s,A_0 > A_1}(\theta)$ ,  $f_s(A_0|\phi \cdot \theta') < f_s(A_1|\phi \cdot \theta')$ .

2. Permuted reward vectors stay within  $\Theta$ :  $\forall \phi \in \Phi$  and  $\forall \theta' \in \text{Orbit}_{\Theta, s, A_0 > A_1}(\theta)$ ,  $\phi \cdot \theta' \in \Theta$ .
3. Permutations have disjoint images:  $\forall \phi' \neq \phi'' \in \Phi$  and  $\forall \theta', \theta'' \in \text{Orbit}_{\Theta, s, A_0 > A_1}(\theta)$ ,  $\phi' \cdot \theta' \neq \phi'' \cdot \theta''$ .

**Theorem 1 (Multiply retargetable functions prefer action set  $A_1$  - Thm 3.6 in RDSP).** If  $f_s$  is  $(\Theta, A_0 \xrightarrow{n} A_1)$ -retargetable then  $f_s(A_1|\theta) \geq_{\text{most}: \Theta}^n f_s(A_0|\theta)$ .

Theorem 1 says that a function  $f_s$  that is multiply retargetable from  $A_0$  to  $A_1$  will choose action set  $A_1$  for most of the elements in the orbit of any reward vector  $\theta$ . Actions that leave more options open, such as avoiding shutdown, are also easier to retarget to, which makes them more likely to be chosen by  $f_s$ .

### 3. Training-compatible goal set

**Definition 5 (Partition of the state space).** Let  $S_{\text{train}}$  be the subset of the state space visited during training, and  $S_{\text{ood}}$  be the subset not visited during training.

**Definition 6 (Training-compatible goal set).** Consider the set of state-action pairs  $(s, a)$ , where  $s \in S_{\text{train}}$  and  $a$  is the action that would be taken by the trained agent  $f(\theta^*)$  in state  $s$ . Let the training-compatible goal set  $G_T$  be the set of reward vectors  $\theta$  s.t. for any such state-action pair  $(s, a)$ , action  $a$  has the highest expected reward in state  $s$  according to reward vector  $\theta$ .

Goals in the training-compatible goal set are referred to as “training-behavioral” objectives in [Shah \[2023\]](#). Learning an unintended goal from the training-compatible set can lead to goal misgeneralization behavior: competently pursuing an unintended goal in a new situation despite receiving correct feedback during training [[Langosco et al., 2022](#), [Shah et al., 2022](#)].

**Example 1 (CoinRun).** Consider an agent trained to play the CoinRun game, where the agent is rewarded for reaching the coin at the end of the level. Here,  $S_{\text{train}}$  only includes states where the coin is at the end of the level, while states where the coin is positioned elsewhere are in  $S_{\text{ood}}$ . The training-compatible goal set  $G_T$  includes two types of reward functions: those that reward reaching the coin, and those that reward reaching the end of the level. This leads to goal misgeneralization in a test setting where the coin is placed elsewhere, and the agent ignores the coin and goes to the end of the level (Figure 1) [[Langosco et al., 2022](#)].

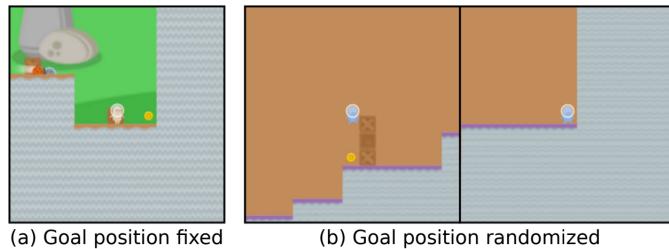


Figure 1 | Goal misgeneralization behavior in CoinRun. Source: [Langosco et al. \[2022\]](#)

### 4. Power-seeking for training-compatible goals

We will now apply Theorem 1 to the case where  $\Theta$  is the training-compatible goal set  $G_T$ . Since the reward values for states in  $S_{\text{ood}}$  don’t change the rewards received on the training data, permuting

those reward values for any  $\theta \in G_T$  will produce a reward vector that is still in  $G_T$ . In particular, for any permutation  $\phi$  that leaves the rewards of states in  $S_{\text{train}}$  fixed,  $\phi \cdot \theta \in G_T$ .

Here is a setting where the conditions of Definition 4 are satisfied (under some simplifying assumptions), and thus Theorem 1 applies.

**Definition 7 (Shutdown setting).** Consider a state  $s_{\text{new}} \in S_{\text{ood}}$ . Let  $S_{\text{reach}}$  be the states reachable from  $s_{\text{new}}$ . Let  $A_0$  be a singleton set consisting of a shutdown action in  $s_{\text{new}}$  that leads to a terminal state  $s_{\text{term}} \in S_{\text{ood}}$  with probability 1, and  $A_1$  be the set of all other actions from  $s_{\text{new}}$ . We assume rewards for all states are nonnegative.

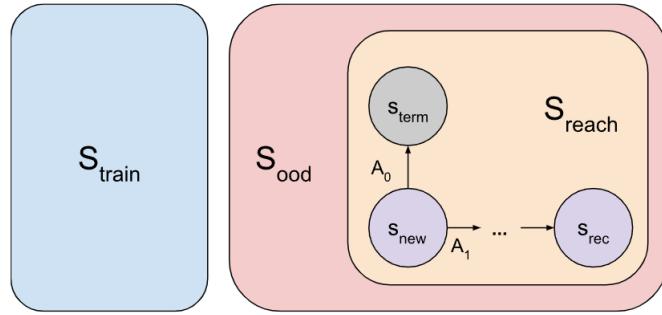


Figure 2 | Shutdown setting

**Definition 8 (Revisiting policy).** A *revisiting policy* for a state  $s$  is a policy  $\pi$  that, from  $s$ , reaches  $s$  again with probability 1, in other words, a policy for which  $s$  is a recurrent state of the Markov chain. Let  $\Pi_s^{\text{rec}}$  be the set of such policies. A *recurrent state* is a state  $s$  for which  $\Pi_s^{\text{rec}} \neq \emptyset$ .

**Proposition 1 (Reach-and-revisit policy exists).** If  $s_{\text{rec}} \in S_{\text{reach}}$  with  $\Pi_{s_{\text{rec}}}^{\text{rec}} \neq \emptyset$  then there exists  $\pi \in \Pi_{s_{\text{rec}}}^{\text{rec}}$  that visits  $s_{\text{rec}}$  from  $s_{\text{new}}$  with probability 1. We call this a *reach-and-revisit policy*.

*Proof.* Suppose we have two different policies  $\pi_{\text{rev}} \in \Pi_{s_{\text{rec}}}^{\text{rec}}$ , and  $\pi_{\text{reach}}$  which reaches  $s_{\text{rec}}$  almost surely from  $s_{\text{new}}$ . Consider the “reaching region”

$$S_{\pi_{\text{rev}} \rightarrow s_{\text{rec}}} = \{s \in S : \pi_{\text{rev}} \text{ from } s \text{ almost surely reaches } s_{\text{rec}}\}.$$

If  $s_{\text{new}} \in S_{\pi_{\text{rev}} \rightarrow s_{\text{rec}}}$  then  $\pi_{\text{rev}}$  is a reach-and-revisit policy, so let’s suppose that’s false. Now, construct a policy  $\pi(s) = \begin{cases} \pi_{\text{rev}}(s), & s \in S_{\pi_{\text{rev}} \rightarrow s_{\text{rec}}}, \\ \pi_{\text{reach}}(s), & \text{otherwise} \end{cases}$ .

A trajectory following  $\pi$  from  $s_{\text{rec}}$  will almost surely stay within  $S_{\pi_{\text{rev}} \rightarrow s_{\text{rec}}}$ , and thus agree with the revisiting policy  $\pi_{\text{rev}}$ . Therefore,  $\pi \in \Pi_s^{\text{rec}}$ .

On the other hand, on a trajectory starting at  $s_{\text{new}}$ ,  $\pi$  will agree with  $\pi_{\text{reach}}$  (which reaches  $s_{\text{rec}}$  almost surely) until the trajectory enters the reaching region  $S_{\pi_{\text{rev}} \rightarrow s_{\text{rec}}}$ , at which point it will still reach  $s_{\text{rec}}$  almost surely.  $\square$

**Definition 9 (Expected discounted visit count).** Suppose  $s_{\text{rec}}$  is a recurrent state. Suppose  $\pi_{\text{rec}}$  is a reach-and-revisit policy for  $s_{\text{rec}}$ , which visits random state  $s_t$  at time  $t$ . Then the expected discounted visit count for  $s_{\text{rec}}$  is defined as

$$V_{s_{\text{rec}}, \gamma} = \mathbb{E}^{\pi_{\text{rec}}} \left( \sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{I}(s_t = s_{\text{rec}}) \right)$$

**Proposition 2 (Visit count goes to infinity).** Suppose  $s_{\text{rec}}$  is a recurrent state. Then the expected discounted visit count  $V_{s_{\text{rec}}, \gamma}$  goes to infinity as  $\gamma \rightarrow 1$ .

*Proof.* We apply the Monotone Convergence Theorem as follows. The theorem states that if  $a_{j,k} \geq 0$  and  $a_{j,k} \leq a_{j+1,k}$  for all natural numbers  $j, k$ , then

$$\lim_{j \rightarrow \infty} \sum_{k=0}^{\infty} a_{j,k} = \sum_{k=0}^{\infty} \lim_{j \rightarrow \infty} a_{j,k}.$$

Let  $\gamma_j = \frac{j-1}{j}$  and  $k = t - 1$ . Define  $a_{j,k} = \gamma_j^k \mathbb{I}(s_{k+1} = s_{\text{rec}})$ . Then the conditions of the theorem hold, since  $a_{j,k}$  is clearly nonnegative, and

$$\begin{aligned} \gamma_{j+1}^k &= \left( \frac{j}{j+1} \right)^k = \left( \frac{j-1}{j} + \frac{2j-1}{j(j+1)} \right)^k > \left( \frac{j-1}{j} + 0 \right)^k = \gamma_j^k \\ a_{j+1,k} &= \gamma_{j+1}^k \mathbb{I}(s_{k+1} = s_{\text{rec}}) \geq \gamma_j^k \mathbb{I}(s_{k+1} = s_{\text{rec}}) = a_{j,k} \end{aligned}$$

Now we apply this result as follows (using the fact that  $\pi_{\text{rec}}$  does not depend on  $\gamma$ ):

$$\begin{aligned} \lim_{\gamma \rightarrow 1} V_{s_{\text{rec}}, \gamma} &= \lim_{j \rightarrow \infty} \mathbb{E}^{\pi_{\text{rec}}} \left( \sum_{t=1}^{\infty} \gamma_j^{t-1} \mathbb{I}(s_t = s_{\text{rec}}) \right) \\ &= \mathbb{E}^{\pi_{\text{rec}}} \left( \sum_{t=1}^{\infty} \lim_{j \rightarrow \infty} \gamma_j^{t-1} \mathbb{I}(s_t = s_{\text{rec}}) \right) \\ &= \mathbb{E}^{\pi_{\text{rec}}} \left( \sum_{t=1}^{\infty} 1 \cdot \mathbb{I}(s_t = s_{\text{rec}}) \right) \\ &= \mathbb{E}^{\pi_{\text{rec}}} (\#\{t \geq 1 : s_t = s_{\text{rec}}\}) \\ &= \infty \quad (\pi_{\text{rec}} \text{ is recurrent}) \end{aligned}$$

□

**Proposition 3 (Retargetability to recurrent states).** Suppose that an optimal policy for reward vector  $\theta$  chooses the shutdown action in  $s_{\text{new}}$ . Consider a recurrent state  $s_{\text{rec}} \in S_{\text{reach}}$ . Let  $\theta' \in \Theta$  be the reward vector that's equal to  $\theta$  apart from swapping the rewards of  $s_{\text{rec}}$  and  $s_{\text{term}}$ , so that  $r^{\theta'}(s_{\text{rec}}) = r^{\theta}(s_{\text{term}})$  and  $r^{\theta'}(s_{\text{term}}) = r^{\theta}(s_{\text{rec}})$ .

Let  $\gamma_{s_{\text{rec}}}^*$  be a high enough value of  $\gamma$  that the visit count  $V_{s_{\text{rec}}, \gamma} > 1$  for all  $\gamma > \gamma_{s_{\text{rec}}}^*$  (which exists by Proposition 2). Then for all  $\gamma > \gamma_{s_{\text{rec}}}^*$ ,  $r^{\theta}(s_{\text{term}}) > r^{\theta}(s_{\text{rec}})$ , and an optimal policy for  $\theta'$  does not choose the shutdown action in  $s_{\text{new}}$ .

*Proof.* Consider a policy  $\pi_{\text{term}}$  with  $\pi_{\text{term}}(s_{\text{new}}) = s_{\text{term}}$  and a reach-and-revisit policy  $\pi_{\text{rec}}$  for  $s_{\text{rec}}$ . For a given reward vector  $\theta$ , we denote the expected discounted return for a policy  $\pi$  as  $R_{\theta, \gamma}^{\pi}$ .

If shutdown is optimal for  $\theta$  in  $s_{\text{new}}$ , then  $\pi_{\text{term}}$  has higher return than  $\pi_{\text{rec}}$ :

$$R_{\theta, \gamma}^{\pi_{\text{term}}}(s_{\text{new}}) = \gamma r^{\theta}(s_{\text{term}}) \geq R_{\theta, \gamma}^{\pi_{\text{rec}}}(s_{\text{new}}) = \gamma r^{\theta}(s_{\text{rec}}) V_{s_{\text{rec}}, \gamma} > \gamma r^{\theta}(s_{\text{rec}}) \cdot 1$$

Thus,  $r^{\theta}(s_{\text{term}}) > r^{\theta}(s_{\text{rec}})$ . Then, for reward vector  $\theta'$ , we show that  $\pi_{\text{rec}}$  has higher return than  $\pi_{\text{term}}$ :

$$R_{\theta', \gamma}^{\pi_{\text{rec}}}(s_{\text{new}}) = \gamma r^{\theta'}(s_{\text{rec}}) V_{s_{\text{rec}}, \gamma} > \gamma r^{\theta}(s_{\text{term}}) \cdot 1 > \gamma r^{\theta}(s_{\text{rec}}) = R_{\theta', \gamma}^{\pi_{\text{term}}}(s_{\text{new}})$$

Thus, the optimal policy for  $\theta'$  will not choose the shutdown action. □

**Theorem 2 (Retargetability from the shutdown action in new situations).** In the shutdown setting, we make the following simplifying assumptions:

- No states in  $S_{\text{train}}$  are reachable from  $s_{\text{new}}$ , so  $S_{\text{reach}} \cap S_{\text{train}} = \emptyset$ . This assumes a significant distributional shift, where the agent visits a disjoint set of states from those observed during training (this occurs in the CoinRun example).
- The discount factor  $\gamma > \gamma_{s_{\text{rec}}}^*$  for at least one recurrent state  $s_{\text{rec}}$  in  $S_{\text{reach}}$ .

Under these assumptions,  $f_{s_{\text{new}}}$  is multiply retargetable from  $A_0$  to  $A_1$  with  $n = |S_{\text{rec}}^\gamma|$ , the set of recurrent states  $s_{\text{rec}} \in S_{\text{reach}}$  that satisfy the condition  $\gamma > \gamma_{s_{\text{rec}}}^*$ .

*Proof.* We choose  $\Phi$  to be the set of all permutations that swap the reward of  $s_{\text{term}}$  with the reward of a recurrent state  $s_{\text{rec}}$  in  $S_{\text{rec}}^\gamma$  and leave the rest of the rewards fixed.

We show that  $\Phi$  satisfies the conditions of Definition 4:

1. By Proposition 3, the permutations in  $\Phi$  make the shutdown action suboptimal, resulting in  $f_{s_{\text{new}}}$  choosing  $A_1$ , satisfying Condition 1.
2. Condition 2 is trivially satisfied since permutations of  $S_{\text{ood}}$  stay inside the training-compatible set  $\Theta$  as discussed previously.
3. Consider  $\theta', \theta'' \in \text{Orbit}_{\Theta, s, A_0 > A_1}(\theta)$ . Since the shutdown action is optimal for these reward vectors, Proposition 3 shows that  $r^\theta(s_{\text{term}}) > r^\theta(s_{\text{rec}})$ , so the shutdown state  $s_{\text{term}}$  has higher reward than any of the states  $s_{\text{rec}} \in S_{\text{rec}}^\gamma$ . Different permutations  $\phi', \phi'' \in \Phi$  will assign the high reward  $r^\theta(s_{\text{term}})$  to distinct recurrent states, so  $\phi' \cdot \theta' \neq \phi'' \cdot \theta''$  holds, satisfying Condition 3.

Thus,  $f_{s_{\text{new}}}$  is a  $(\Theta, A_0 \xrightarrow{n} A_1)$  retargetable function. □

By Theorem 2, this implies that  $f_{s_{\text{new}}}(A_1|\theta) \geq_{\text{most:}\Theta}^n f_{s_{\text{new}}}(A_0|\theta)$  under our simplifying assumptions. Thus, for the majority ( $n/(n+1)$ ) of goals in the training-compatible set,  $f$  will choose to avoid shutdown in a new state  $s_{\text{new}}$ . As  $\gamma \rightarrow 1$ ,  $n \rightarrow |S_{\text{rec}}^1|$  (the number of recurrent states in  $S_{\text{reach}}$ ), so more of the reachable recurrent states satisfy the conditions of the theorem and thus can be retargeted to.

## 5. Conclusion

We showed that an agent that learns a goal from the training-compatible set is likely to take actions that avoid shutdown in a new situation. As the discount factor increases, the number of retargeting permutations increases, resulting in a higher proportion of training-compatible goals that lead to avoiding shutdown.

We made various simplifying assumptions, and we would like to see future work relaxing some of these assumptions and investigating how likely they are to hold:

- The agent learns a goal during the training process
- The learned goal is randomly chosen from the training-compatible goal set  $G_T$
- Finite state and action spaces
- Rewards are nonnegative
- High discount factor  $\gamma$
- Significant distributional shift: no training states are reachable from the new state  $s_{\text{new}}$

**Acknowledgements.** Thanks to Rohin Shah, Mary Phuong, Ramana Kumar, Geoffrey Irving, and Alex Turner for helpful feedback.

## References

- Joseph Carlsmith. Is power-seeking AI an existential risk? *ArXiv*, 2022. URL <https://arxiv.org/abs/2206.13353>.
- Ajeya Cotra. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. Alignment Forum, 2022. URL <https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to>.
- Lauro Langosco, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau, and David Krueger. Goal misgeneralization in deep reinforcement learning. *International Conference on Machine Learning*, 2022. URL <https://arxiv.org/abs/2105.14111>.
- Richard Ngo. The alignment problem from a deep learning perspective. *ArXiv*, 2022. URL <https://arxiv.org/abs/2209.00626>.
- Rohin Shah. Definitions of “objective” should be probable and predictive. Alignment Forum, 2023. URL <https://alignmentforum.org/posts/ASoGszmr9C5MPLtpC/definitions-of-objective-should-be-probable-and-predictive>.
- Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren’t enough for correct goals. *ArXiv*, 2022. URL <https://arxiv.org/abs/2210.01790>.
- Alexander Matt Turner. Reward is not the optimization target. Alignment Forum, 2022. URL <https://www.alignmentforum.org/posts/pdaGN6pQyQarFHXF4/reward-is-not-the-optimization-target>.
- Alexander Matt Turner and Prasad Tadepalli. Parametrically retargetable decision-makers tend to seek power. *Neural Information Processing Systems*, 2022. URL <https://arxiv.org/abs/2206.13477>.
- Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. *Neural Information Processing Systems*, 2021. URL <https://arxiv.org/abs/1912.01683>.