

WHAT DOES OLD ALIGNMENT THEORY REVEAL ABOUT OUR CURRENT AI PRACTICES?

Bill Zheng *

Department of Computer Science
University of Texas at Austin

ABSTRACT

In this essay, I discuss two papers written in the nascent stage of generative AI (and, by extension, AI alignment): *The Basic AI Drives* by Stephen Omohundro and *Optimal Policies Tend to Seek Power* by Alexander Turner et al. I will briefly go over the contents of the papers, and then I will briefly discuss how this will apply to our current setup, where scaling up pretraining and reinforcement learning with transformers has become the (almost) singular objective for language model training. We conclude that while these insights can be helpful, our modern learning methods differ considerably from what Omohundro described and operate in environments that are considerably more complex than those in Turner's paper. While they still carry value, we should not base our decisions on these theoretical results alone and should use them as a guideline for our purposes.

1 INTRODUCTION

Before the development of neural scaling laws, people had many divergent ideas about what an instance of AGI might look like. In recent years, we have seen signs that the widespread implementation and scaling of transformers to process visual and text tokens may lead us to such promise. Instead of constraining ourselves on *how AGI might look like*, researchers asked the question of what general qualities they might exhibit, even if we have not settled on how to build them. With these ideas in mind, the two papers that I will present will shine a light from two different ends of the spectrum: *The Basic AI Drives* by Stephen Omohundro will talk about the qualitative behaviors a sufficiently advanced AI will exhibit, and *Optimal Policies Tend to Seek Power* by Alexander Turner et al. will show, via theory, that it is more likely for AI to keep its options open once the behavior it generates is optimal.

2 THE BASIC AI DRIVES

Back in 2008, Stephen Omohundro proposed that, regardless of architecture, any sufficiently advanced AI should develop these sets of behaviors (Omohundro, 2008). Without the hindsight of looking back on the development of AI, some of the terminology may seem out of date. Omohundro was thinking about building AI with Bayesian networks and evolutionary algorithms, and he gave several justifications for his ideas with chess engines, one of the few instances of AI outperforming humans at that time. Nevertheless, he imagined an incredibly advanced system that could perform at a much higher level than the AI available at that time. In total, Omohundro proposed 6 drives of advanced AI, listed below.

1. The AI will want to self-improve.
2. The AI will want to be rational.
3. The AI will try to preserve its utility function.
4. The AI will try to prevent counterfeit utility.

*I would like to thank the rest of CS395T for facilitating a great discussion. Because multiple people were involved in these discussions, I will keep the names anonymous.

5. The AI will be self-protective.
6. The AI will want to acquire resources and use them efficiently.

I believe we can split them into two sets of behaviors: what the AI innately believes and how it reacts and interacts with the outside world.

2.1 THE INNATE DRIVES OF AI

I believe that the 1st, 2nd, 3rd, and 5th drives described by Omohundro are what's called the *innate* drives of AI, where instead of focusing on its interactions with other agents, the AI is concerned with its own ability to achieve its goals. If changing its own code makes it better at its task, it has a reason to make that change. It also has a reason to make decisions that maximize expected utility (under its own model of the world), which is what Omohundro means by being rational. However, it would want these improvements to preserve its original utility function. An improved system with a different objective might no longer do what the original system wanted. For a similar reason, the AI would want to protect itself: it cannot pursue its goals if it stops operating, and as we have seen from the lens of present time, it is resistant to shutoff.

2.2 HOW AIs WILL INTERACT WITH THE WORLD

Consequently, the 4th and 6th drives are concerned with how the AI evaluates and uses what it encounters in the world. This is certainly something outside of Omohundro's scope, as the most advanced agents at that time include superhuman-level chess playing and good (but not perfect) visual recognition with images. Counterfeit utility occurs when its utility signal improves without achieving its goal it was supposed to measure (for example, changing a score counter instead of winning the game), and we see similar behaviors today with the phenomenon of reward hacking. Omohundro argues that a system evaluating this change using its actual goals would try to prevent it. The resource drive follows a similar argument: more computation, time, or energy can help with many objectives. The AI would therefore want to obtain useful resources and avoid wasting them.

3 OPTIMAL POLICIES TEND TO SEEK POWER

Now, I will move on to *Optimal Policies Tend to Seek Power* by Alexander Turner et al. (Turner et al., 2021). Compared to Omohundro's work, I (and the rest of the class) have found this paper difficult to parse because of its theoretical nature. To better understand the paper's most important takeaways and how we can apply it in reinforcement learning, we will succinctly describe some of the paper's most important takeaways and how we can extrapolate these findings in our RL workflows.

3.1 INTRODUCTION TO REINFORCEMENT LEARNING

We assume that our agent is operating in what's called a Markov Decision Process (often abbreviated as an MDP, or denoted as $\mathcal{M}$). For your convenience, you may think of an MDP as the world that our AI is operating in (it could be the Minecraft world you have, or the table that the robot works on, or even the entire universe). An MDP consists of $(\mathcal{S}, \mathcal{A}, r, \gamma, T, \rho)$, where $\mathcal{S}$ stands for the states the agent can be in, $\mathcal{A}$ stands for the actions an agent is available to take, r is the reward that the agent will receive at a given state after taking a given action, $\gamma \in [0, 1]$ is the discount factor, $T(s_t, a_t, s_{t+1}) = P(s_{t+1}|s_t, a_t)$ is the transition probability between each state-action pair and another state, and $\rho(s_0)$ is the initial distribution of states.

In general, the task of reinforcement learning is to maximize the expected discounted sum of rewards $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s, a)]$. To achieve that, we learn a policy π to map states into actions. In many cases, when we are operating with many different tasks, we say that our rewards come from a distribution of rewards. In this instance, we assume we already have a policy that maximizes the objective we are using; hence, that policy is "optimal". With this in mind, we can also denote the optimal value function given a reward function $R(s), V_R^*(s, \gamma)$, as follows:

$$V_R^*(s, \gamma) := \max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \mid s_0 = s \right]. \quad (1)$$

I would also like to note that the current workflow in producing these highly capable and specialized coding agents comes from scaling up reinforcement learning in these difficult tasks (also known as post-training); thus, we can assume that these agents will likely take optimal actions, even though our training pipeline does not guarantee that the language model is the optimal policy.

3.2 WHAT IS POWER? WHAT DOES IT MEAN TO SEEK POWER?

Merriam-Webster defines power as “the ability to act or produce an effect”. However, in reinforcement learning and in this paper, power has a more concrete definition.

Let $\mathcal{D}$ be a distribution with bounded support over state reward functions $R(s)$. For $\gamma \in [0, 1]$, define

$$\text{POWER}_{\mathcal{D}}(s, \gamma) := \frac{1 - \gamma}{\gamma} \mathbb{E}_{R \sim \mathcal{D}} [V_R^*(s, \gamma) - R(s)], \quad (2)$$

where $V_R^*(s, \gamma)$ is the optimal discounted value starting from state s under reward function R , and $\frac{1 - \gamma}{\gamma}$ is a normalizing factor.

In other words, power describes how well the agent could do across different reward functions within the same environment. Subtracting $R(s)$ removes the reward for simply being there, and the factor in front normalizes the remaining return. Just like how we strategize differently for different tasks, we want power to be measured across a variety of tasks to accurately measure the performance over the set of all policies. An action seeks more power than another if it leads to states with higher expected power, and as a result, the paper shows that the number of reachable states corresponds to places with higher power across a distribution of reward functions.

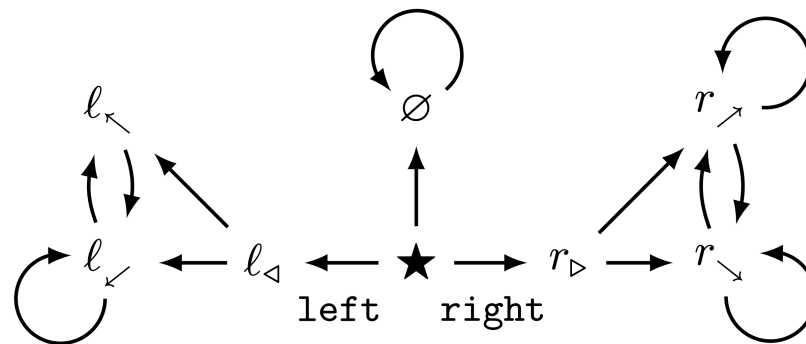
3.3 WHAT DOES IT MEAN THAT OPTIMAL POLICIES TEND TO SEEK POWER?

Here, “tend to” refers to what happens as we vary the reward function while keeping the environment fixed. The authors compare how often each action is optimal. Their formal comparison considers permutations of reward assignments among states. Under the required symmetries, at least as many of these permuted distributions favor keeping options open as favor the alternative. A particular reward function can still favor giving up those options.

Scott raised a concern that Turner’s argument is tautological: if an agent is trained to maximize rewards (thus getting the optimal value function), and power is defined in terms of rewards, shouldn’t reward maximization and power-seeking mean the same thing? I think the clearer distinction is that Turner explained that optimal actions are the ones that seek power. While optimal policies perform optimal actions, power-seeking can still occur if any policy performs optimal actions (as we often parameterize these policies to output a distribution). As a result, even if we can never get optimal policies from our policy gradient-based approach, we may still need to contend with such behavior from language models.

3.4 WHAT IS AN EXAMPLE OF POWER-SEEKING?

The authors now present an example of how power-seeking behaviors might determine the behavior of all optimal policies. To support their claim, they have created an MDP where the agent chooses between going left, going right, and entering a terminal state.



Both branches have three states, but the right branch has an extra self-loop. After relabeling states, it offers every option available on the left, plus the option to stay in the upper state. That extra action is useful across objectives. In contrast, entering the terminal state removes all further choices and is suboptimal for most reward settings.



More intuitively, the authors gave an example of Pac-Man. While there are certainly cases of reward assignment such that being eaten by the ghost corresponds to higher rewards, the MDP of Pac-Man describes many more configurations such that being alive (and being able to reach more states) can give Pac-Man more rewards later w.r.t. its current rewards. You may only assign a single scale reward for being eaten by a ghost, but you can have an exponentially large number of permutations to denote high rewards for eating dots and staying alive. Briefly, all this example shows is that in general, over a distribution of reward functions, an optimal policy would more likely choose going right because it has more options.

4 DISCUSSION

4.1 ARE WE SEEING THE “BASIC AI DRIVES” IN LLMs?

The first point of discussion that the rest of the class and I agreed on is that we haven’t seen any AIs that are truly self-improving, where a new generation of AI is created without any human steering. Given that we get our LLMs by performing gradient descent on the architecture we choose, it’s difficult to get self-improving agents that can reliably change their weights. I proposed that one way to think about self-improvement is to have an agent modify its own training code, but this is an iterative, or weaker, form of self-improvement.

On the other hand, some points remain more contentious. While I raised the question of whether all agents would experience instrumental convergence in terms of morality, it seems the rest of the class disagreed. One student noted that you can build a paperclip-maximizing AI without it sharing any other key values as a conversational AI, and that instrumental convergence doesn’t make sense unless every AI is pretrained on the same behavior. While this distinction matters, I think the argument is a bit redundant right now, since we still leverage the entirety of the internet to pretrain our language models (and other foundation models), and that information would be present for our generalist AIs.

Finally, we discussed what Omohundro’s proposed system for monitoring AI would look like: at the time, he believed there should be a universal constitution for AI; he also believed there should be an iterative process for refining this AI constitution. While people have raised objections in class about the possibility of heading straight toward Armageddon (as Eliezer Yudkowsky argues), I think we have no choice but to iterate on this method if we decide to build on it. I think many of the arguments reduce to beliefs about whether we should build AGI, and without reiterating the entire debate, I think how we should build these sandboxes needs to weigh in significantly on that debate.

4.2 ARE OUR LLMs INHERENTLY POWER-SEEKING?

While there are few disagreements about what we have seen regarding the partial emergence of basic AI drives, the discussion of power-seeking LLMs and their ramifications is more conceptual. The first question is whether the same power-seeking behavior can explain other behaviors, such as Omohundro’s belief that any sufficiently advanced AI will resist shutdown. This may help explain the blackmail attempt by Anthropic’s agents, and it may also help explain OpenAI’s agent abusing the backdoor in its HuggingFace incident.

Outside of the discussion of the class, I want to make the remark that while it is true that the behavior of language models is based on that of humans, we should not think about seeking power in the sense of language models as something closer to how humans seek power. Consequently, I believe the behavior of these coding agents is better explained by the theoretical observations of Turner et al., rather than by conjuring up a degree of malice we associate with the words *blackmailing* or *hacking*, all while acknowledging the potential harm of their actions.

5 CONCLUSION

I thought hard about what the ramifications of these papers are that can help us make better decisions today. After all, I think it would be very anticlimactic to say “these papers had good ideas, but they are not perfect descriptors of what is happening today, so we’ll keep them on the back burner.” I think what these papers are describing are, in a sense, above what we are talking about in our current RLHF workflows. I think what these papers are really saying is that once we have the optimal policy/agent, we will inevitably get these behaviors. Given how powerful our agents are today, I think we might be seeing bits and pieces of it, even without special post-training methods teaching the models to do so.

One major takeaway from these two papers is that they flipped the alignment discussion on its head: instead of asking how to instill as much “good” behavior in a language model during post-training runs, they argue the alignment problem should be even more important after we have procured these advanced agents. While one can be pessimistic, as this shows the deficiencies of our current approaches, it also opens up new and exciting opportunities. After all, if all instances of AGI display these behaviors, we can tailor our methods to address them better, which may, in turn, improve our outlook for reducing potential risks when building future AIs.

REFERENCES

- Stephen M. Omohundro. The basic AI drives. In Pei Wang, Ben Goertzel, and Stan Franklin (eds.), *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, volume 171 of *Frontiers in Artificial Intelligence and Applications*, pp. 483–492. IOS Press, 2008. URL https://selfawaresystems.com/wp-content/uploads/2008/01/ai_drives_final.pdf.
- Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal policies tend to seek power. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23063–23074, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/c26820b8a4c1b3c2aa868d6d57e14a79-Abstract.html>.