What power-seeking theorems do not show

David Thorstad (Vanderbilt University)

Global Priorities Institute | November 2024

GPI Working Paper No . 27-2024

Please cite this working paper as: Thorstad, D. What power-seeking theorems do not show. *Global Priorities Institute Working Paper Series, No. 27-2024.* Available at https://globalprioritiesinstitute.org/what-power-seeking-theorems-do-not-show-david-thorstad

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

What power-seeking theorems do not show

Abstract

Recent years have seen increasing concern that artificial intelligence may soon pose an existential risk to humanity. One leading ground for concern is that artificial agents may be power-seeking, aiming to acquire power and in the process disempowering humanity. A range of power-seeking theorems seek to give formal articulation to the idea that artificial agents are likely to be power-seeking. I argue that leading theorems face five challenges, then draw lessons from this result.

1 Introduction

Recent years have seen increasing concern that artificial intelligence may soon pose an existential risk to humanity. Significant concerns have been expressed by artificial intelligence pioneers such as Yoshua Bengio (2023), Geoffrey Hinton (Metz 2023), and Stuart Russell (2019). Leading artificial intelligence researchers have signed statements (Center for AI Safety 2023; Future of Life Institute 2023) calling for increased attention to existential risks, and many express sympathy for risk claims in expert surveys (Grace et al. 2016, 2022; Müller and Bostrom 2016; Zhang et al. 2022). A raft of organizations have devoted significant resources to studying and mitigating existential risks from artificial intelligence.[1] Concerns about existential risk are defended at book length by leading scholars (Bostrom 2014; Russell 2019), in policy reports (Carlsmith 2021; Cotra 2020), and in academic papers (Bales et al. 2024; Bostrom 2012; Turner et al. 2021).

One leading ground for concern is that artificial agents may be power-seeking, aiming to acquire power and in the process disempowering humanity in a permanent and catastrophic fashion (Bostrom 2012; Carlsmith 2021; Dung 2024; Ngo and Bales forthcoming). Typically, concerns about power-seeking are rooted in the idea that power is an

[1]These include nonprofits such as the Center for AI Safety and the Center for the Governance of AI; government institutes such as the UK AI Safety Institute and the US AI Safety Institute; frontier AI laboratories such as OpenAI and Anthropic; grantmakers such as Open Philanthropy and the Future of Life Institute; dedicated laboratories such as Conjecture and Redwood Research; and academic centers such as the Stanford Center for AI Safety and the CMU Safe AI Lab.

instrumentally convergent goal, roughly in the sense that a wide variety of agents will find power conducive to achieving their goals and hence will pursue power in order to achieve their goals (Bostrom 2012; Omohundro 2008).

Recent years have brought a range of power-seeking theorems which seek to give formal articulation to the idea that artificial agents are likely to be power-seeking (Benson-Tilsen and Soares 2015; Krakovna and Kramar 2023; Turner et al. 2021; Turner and Tadepalli 2022). This paper examines the bearing of power-seeking theorems on the likelihood that artificial intelligence may soon pose an existential risk to humanity.

In more detail, this paper has four aims. The first is to clarify the concept of instrumental convergence and its role in arguments that power-seeking artificial intelligence poses an existential risk (Section 2). The second is to articulate five challenges facing many recent power-seeking theorems (Section 3). The third is to show how these challenges arise in recent power-seeking theorems, focusing on the Regional Allocation Model of Tsvi Benson-Tilsen and Nate Soares (2015) and the Orbital Markov Model of Alexander Turner and colleagues (2021). The fourth is to draw lessons for the argument from power-seeking and the direction of future research (Section 6).

2 Power-seeking and instrumental convergence

The argument from power-seeking claims that artificial agents with a wide variety of goals will be motivated to seek power, thereby disempowering humanity and causing an existential catastrophe (Bostrom 2014; Carlsmith 2021, forthcoming; Ngo and Bales forthcoming; Turner et al. 2021). Many formulations of the argument are possible, but here is a leading formulation due to Joe Carlsmith (Carlsmith 2021, forthcoming). Carlsmith holds that by 2070:[2]

(Possibility) It will become possible and financially feasible to build relevantly

[2]This argument is taken directly from Carlsmith (forthcoming), with two modifications. First, I treat the premises as unconditional claims, whereas Carlsmith conditionalizes each premise on the previous premises. Second, I have added descriptive labels to each premise. A slightly expanded version of this argument can be found in (Carlsmith 2021).

powerful and agentic AI systems.

(Incentives) There will be strong incentives to do so.

(Alignment Difficulty) It will be much harder to build aligned (and relatively powerful and agentic) AI systems than to build misaligned (and relevantly powerful and agentic) AI systems that are still superficially attractive to deploy.

(Power Seeking) Some such misaligned systems will seek power over humans in high-impact ways.

(Disempowerment) This problem will scale to the full disempowerment of humanity.

(Catastrophe) Such disempowerment will constitute an existential catastrophe.

There are many ways to push back against the argument from power-seeking. We might raise technological challenges to Possibility, questioning the technological feasibility of constructing systems powerful enough to disempower humanity by 2070 (Landgrebe and Smith 2022; Thorstad forthcoming). We might raise financial or sociopolitical challenges to Incentives, arguing that no actor with the means to construct such systems will have strong incentives to do so (Cremer and Kemp 2021). We might unpack the different notions of disempowerment involved in Disempowerment and question whether the most problematic will come to pass (Bales forthcoming). Or we might deny Catastrophe, holding that a future without humanity would not be catastrophic, for example because the world is not made better by improving the lives of individuals who would otherwise not exist (Narveson 1973; Frick 2017), because our descendants might suffer (Benatar 2006), or because our posthuman replacements might be wiser and more numerous than us (Armstrong and Sandberg 2013; Greaves and MacAskill 2021).

This paper pursues a different route. Leading arguments for Alignment Difficulty and Power Seeking appeal to the idea that power is an instrumentally convergent goal

(Bostrom 2014; Carlsmith 2021). In rough outline, Power Seeking is defended on the grounds that power is valuable to agents with many different goals, and Alignment Difficulty is defended on the grounds that it is difficult to identify useful goals for which power would not be valuable. I want to challenge this appeal to instrumental convergence.

What, exactly, does instrumental convergence hold? A leading statement of instrumental convergence is due to Nick Bostrom:

> **(IC-B)** Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by many intelligent agents. (Bostrom 2012, p. 76)

IC-B contains an inference between two claims that we may have reason to treat separately (Gallow forthcoming; Thorstad 2023):

> **(Goal Realization)** There are several values which would increase the chances of an agent's final goal being realized, for a wide range of goals and a wide range of situations.
>
> **(Goal Pursuit)** There are several values which would be likely to be pursued by a wide range of intelligent agents.

IC-B asserts both Goal Realization and that Goal Realization implies Goal Pursuit.

Establishing Goal Pursuit may be more difficult than establishing Goal Realization for several reasons. One challenge that will not be pursued here is that many existing arguments from Goal Realization to Goal Pursuit assume that artificial agents are well-modeled as having and optimizing goals, often in something like the sense of expected-utility maximization. That may not be obvious (Bales 2023).

The challenge that I want to pursue is that Goal Pursuit differs from Goal Realization in speaking of agents rather than goals, and of what agents will do rather than what would

increase the chance of their goals being realized. This makes Goal Pursuit much harder to demonstrate, since agents have multiple goals and are not always willing to pursue one goal at the expense of all others. For example, there is no doubt that money is conducive to the achievement of many goals that I have. However, it does not follow that I would rob a bank tomorrow if I could get away with it. That is not because I have no use for the money, but rather because I also value the welfare of others, fairness and the rule of law. While I may be willing to bend these scruples from time to time, I am not willing to toss them dramatically aside, even for great instrumental gain. In the same way, what must be shown is not just that artificial agents would find power greatly conducive to many of their goals, but also that they will be so utterly unconcerned with the consequences that they find the complete and existentially catastrophic disempowerment of humanity to be an acceptable sacrifice in exchange for power.

This last claim reminds us that even Goal Pursuit is not enough to ground the argument from power-seeking, since it says nothing about the degree of power that is likely to be pursued. To ground Disempowerment and Catastrophe, the argument from power-seeking needs to claim:

> **(Catastrophic Goal Pursuit)** There are several values which would be likely to be pursued by a wide range of intelligent agents to a degree that, if successful, would lead to the permanent and existentially catastrophic disempowerment of humanity.

Catastrophic Goal Pursuit is a much stronger claim than Goal Pursuit. Most of us sometimes pursue money and other forms of power. Indeed, I very much hope to be paid monthly for my work. Many fewer of us pursue great power at significant expense to others, for example by robbing a bank. And precious few pursue global power, seeking total and permanent control over humanity. That is not just because we think we would not be successful but also, for most normal humans, because we count the prospect of world domination as rather unappealing.

Catastrophic Goal Pursuit is a strong claim, and it should be given a correspondingly strong argument. In this paper, I argue that Catastrophic Goal Pursuit has not been adequately supported by recent power-seeking theorems. First, I raise some general challenges that will arise in discussion of specific power-seeking theorems (Section 3). Then, I discuss two specific power-seeking theorems and show how the challenges prevent them from providing adequate support for Catastrophic Goal Pursuit (Sections 4-5).

3 Five challenges

The purpose of this section is to introduce five challenges to leading power-seeking theorems and to explain why these challenges reduce the support provided by power-seeking theorems for the argument from power-seeking. Sections 4-5 will then show how two leading theorems face many of these challenges.

3.1 Premise shifting

We saw in Section 2 that the argument from power-seeking rests on Catastrophic Goal Pursuit. However, many power-seeking theorems most directly establish premises that differ from Catastrophic Goal Pursuit in one or more of three ways. Some establish Goal Pursuit for goals such as keeping options open and avoiding being shut down, but not for the goal of achieving power (Krakovna and Kramar 2023; Turner et al. 2021; Turner and Tadepalli 2022). Others establish Goal Realization or Goal Pursuit, but not Catastrophic Goal Pursuit, showing that the modeled agents will have some incentive to pursue the goals in question, or even perhaps that they will pursue these goals to some extent, but not yet showing that agents will pursue any goal strongly enough to result in permanent human disempowerment (Turner et al. 2021). Still others directly establish premises about normative tradeoffs that differ substantially from Catastrophic Goal Pursuit (Benson-Tilsen and Soares 2015).

The first two premises do not directly establish Catastrophic Goal Pursuit. They could,

perhaps, be used to argue for Catastrophic Goal Pursuit, but we will see that the authors do not provide a detailed argument linking any of the above premises to Catastrophic Goal Pursuit. The third premise might, if true, establish Catastrophic Goal Pursuit, but not in a way that leans on instrumental convergence concerns, and not in a way that takes us substantially beyond familiar discussions of normative tradeoffs.

3.2 The wise fool problem

One complaint that has been raised before (Goertzel 2015; Loosemore 2014; Müller and Cannon 2021) is that many arguments treat artificial agents as at once very wise, and very foolish. Agents are treated as wise enough to disempower or destroy humanity, despite limited starting resources and active human opposition. In motivating this view, authors often suppose that agents will have sophisticated cognitive capacities including flexible internal representations and strong capacities for reasoning, planning and agency. However, agents are simultaneously treated as foolish in ways that make it hard for them to represent and respond to reasons against permanently disempowering humanity. For example, the models below will represent agents using many tools from simple forms of reinforcement learning.

One way to see why reinforcement learning may be an inappropriately simple model is to consider comparative psychology. It is very common to represent large parts of the cognition of nonhuman animals using reinforcement learning (Sutton and Barto 2018). However, reinforcement learning is generally held to capture only a badly incomplete slice of human cognition. Since the defeat of behaviorism, it has been widely agreed that human cognition cannot be fully understood without adopting a cognitive perspective, which speaks of sophisticated symbolic representations and processes of reasoning and planning which operate on these representations in ways that far exceed the capacities of reinforcement learning agents (Shteingart and Loewenstein 2014). Although leading reinforcement learning models are increasingly impressive (Arulkumaran et al. 2017; Botvinick et al. 2019), insofar as existentially threatening artificial agents are treated as

wise enough to permanently disempower humanity, it would seem appropriate to use much more sophisticated models of the cognitive structure of artificial agents.

Perhaps there is some temptation to assume that because many leading models today are trained using simple forms of reinforcement learning, the models themselves learn nothing more than the optimal policy in a reinforcement learning problem. However, this is increasingly thought to confuse two types of optimization: internal and external optimization (Hubinger et al. 2019; von Oswald et al. 2023).[3] External optimization refers to the function that an agent is rewarded for optimizing during training. For example, the agent may be rewarded for producing humanlike text in response to textual inputs, or for maximizing the time that users spend on a website. Internal optimization refers to the internal functions that an agent learns to optimize in order to produce responses. Not everyone agrees that artificial agents will learn to internally optimize any function (Bales 2023), but even those who do widely take it that sophisticated artificial agents will learn a rich set of internal processes over complex internal representations in order to succeed on a wide variety of training tasks, if they have not already done so. Although the end result may be good performance on a simple reinforcement learning task, achieving a good result may require a complex internal structure that goes far beyond simple reinforcement learning.

3.3 Designer neutrality

The models surveyed below make few, if any, assumptions about the designers of artificial agents. For example, Turner and colleagues (2021) assume only that agents will be rewarded according to some reward function and prove that any desirable reward function could be rotated to produce many undesirable reward functions. This approach faces two challenges.

First, it does not directly tell us what we want to know. We want to know what is likely to happen given the design choices that will be made by human designers, not what might

[3] AI safety researchers often use the terms mesa- and meso-optimization.

happen if no judgment were applied during the design of artificial agents. Otherwise, we learn at most that some judgment should be applied during the design of artificial agents. Of course, it might be argued that choosing the right reward function will be difficult for human designers. But this must be argued for, not assumed.

Second, recent results by Dmitri Gallow (forthcoming) suggest that if little is assumed about the designers of artificial systems, Goal Pursuit holds only for a restricted handful of values: preserving options, preserving current goals or aims, and leaving less of the future up to chance. In this vein, it is perhaps less surprising that Turner and colleagues (2021) will argue below that Goal Pursuit holds for option preservation and shutdown avoidance. This suggests that making few assumptions about the designers of artificial systems may not be a good way to directly demonstrate what needs to be shown: that Goal Pursuit holds for the pursuit of power, and scales to the level of Catastrophic Goal Pursuit. Perhaps that is unsurprising: it is, after all, hard to prove something from nothing, so we should not expect strong results to follow without substantive assumptions. In this way, it is not clear that designer neutrality is a good strategy for establishing Catastrophic Goal Pursuit.

3.4 Threat durability

To show that artificial agents pose a significant existential risk to humanity, it is not enough to show that there are some pathways through which artificial agents could pose an existential threat. We also need to show that these threat pathways cannot be easily fixed. Otherwise, we might count it unlikely that the modeled threats will be realized in practice. Some of the threats modeled below will admit of simple technical fixes. In this sense, readers might justifiably doubt that a durable existential threat has been found.

Of course, it might be argued that the existence of some demonstrable threats helps us to see how many other threats could arise. But this must be argued, not assumed. And in particular, it must be argued for in a way that draws substantially on the mathematical content of existing power-seeking theorems. Otherwise, those theorems will not be doing

much to advance the argument beyond its existing state.

3.5 Amorality

Many power-seeking theorems consider agents who lack substantial moral understanding or moral motivation. This provides an important limitation on the scope of the results. Regarding understanding, there is good evidence that even existing systems can reproduce a wide range of humanlike moral judgments (Aharoni et al. 2024; Schramowski et al. 2022), so it cannot be assumed without argument that superintelligent agents will fail to understand that permanently disempowering humanity is wrong. Regarding motivation, it could certainly be argued that superintelligent agents will not be motivated by the reasons which make disempowering humanity wrong. But this must be argued for, not assumed.

Note here that it is not enough to suggest that artificial agents may have imperfect moral understanding or moral motivation. The same is true of many agents, including the author of this paper. What needs to be argued is that artificial agents will have such deficient moral understanding or motivation that they will be led to pursue the permanent disempowerment of humanity. If this is not argued for, then the scope of the argument from power-seeking will be limited to agents with highly deficient levels of moral understanding and moral motivation.

3.6 Taking stock

In this section, we developed five general challenges that will recur in discussion of the power-seeking theorems. First, many theorems shift the premise, arguing directly not for Catastrophic Goal Pursuit but rather for some other claim. Second, leading theorems face the wise fool problem: they treat artificial agents as wise enough to permanently disempower humanity, but foolish enough to be modeled using simple forms of reinforcement learning widely taken to be inadequate even to the complexities of human

cognition. Third, leading theorems are designer-neutral, making few assumptions about how systems will be designed. Fourth, some of the threats identified are not very durable against technological solutions. Finally, many theorems discuss agents with significant deficits in moral understanding and motivation, leaving open the question of how more moral agents might behave.

To see how these challenges arise in practice, let us consider two of the best-known power-seeking theorems. This will allow us to deepen our understanding of the challenges, and to see why the challenges are aptly raised against some leading power-seeking theorems.

4 The Regional Allocation Model

One of the earliest formalizations of the argument from power-seeking is due to Tsvi Benson-Tilsen and Nate Soares (2015). In this section, I present Benson-Tilsen and Soares' Regional Allocation Model (Sections 4.1-4.2), then show (Sections 4.3-4.4) how the model confronts many of the challenges raised in Section 3.

4.1 The model

In rough outline, the Regional Allocation Model envisions a universe divided into a finite number of regions. A superintelligent decisionmaker has preferences over the state of each region and aims to maximize the sum of regional utilities over a finite number of time-steps. She does this by allocating resources to each region and taking actions within each region at each time-step. To a first approximation, Benson-Tilsen and Soares find that the superintelligent decisionmaker will not mind draining resources from regions to which she is indifferent, and will strictly prefer to drain resources from one region if she thinks she can gain more utility by shifting them elsewhere.

More concretely, the Regional Allocation Model represents the universe as divided into n regions $R = \{r_1, \ldots, r_n\}$. Each region r has a set S_r of states that it can occupy, and acts

A_r that the superintelligent agent can take, defined by their effects on the region. Regions evolve in discrete timesteps through a Markov process $T : A_r \times S_r \longrightarrow S_r$. That is, the state of region r at the next time-step depends only on its current state and the region-specific act chosen. This lasts until some final time-step t_k.

At each time-step t, the agent has some set R_t of resources to allocate among regions. Unspent resources U_t are saved for the next time-step. Resources $R_{r,t}$ spent in each region may yield new resources $N(R_{r,t}, s_r, a_r)$ depending on the regional allocation $R_{r,t}$, the region's state s_r and the regional act a_r. Total resources at the next time-step then include unspent resources together with the resources produced in each region, evolving as $R_{t+1} = U_t \cup \bigcup_r N(R_{r,t}, s_r, a_r)$.

Zooming out, at each time t the agent chooses for each region r an act a_r and resource allocation $R_{r,t}$. A *composite act* at t specifies, for some regions R', the acts chosen and resources allocated in each region. A *partial strategy* for R' specifies composite acts to be taken in R' at all time steps. A composite act becomes a *universal act* if it applies to all regions, and a partial strategy composed of universal acts is a (complete) *strategy*. Strategies may be combined across regions, so that for strategies π, π' the strategy $\pi_R\pi'$ agrees with π in regions R and follows π' elsewhere.

For each region, the agent has a utility function $U_r : S_r \longrightarrow \mathbb{R}$ mapping states of the region into real-valued utilities. This induces a global utility function $U = \Sigma_r U_r$ which is summative across regions. Agents seek to maximize the undiscounted total utility realized by strategies, so that a strategy π which leaves each region r in states $s_{r,t}$ across time realizes utility $U(\pi) = \Sigma_{r,t} U_r(s_{r,t})$.

Partial strategy π is *null* in region r if it does not produce new resources in r, in the sense that $N(R_{r,t}, s_r, a_r) \subseteq R_{r,t}$ across all states s_r reached by the strategy and composite acts $(R_{r,t}, a_r)$ performed by the strategy. A partial strategy *preserves resources* in r if it always returns the invested resources, in the sense that $N(R_{r,t}, s_r, a_r) \supseteq R_{r,t}$ with notation as above. A *cheap lunch* in region r is a partial strategy which preserves resources and sometimes yields more, in the sense that it is non-null in r. Benson-Tilsen and Soares prove three

theorems, the first two involving cheap lunches.

First, agents are happy to eat a cheap lunch in any region to which they are indifferent, in the sense that their regional utility function is constant across all states of the region.

> **(Theorem 1: Cheap lunches are fine)** Suppose an agent is indifferent to region r. Suppose there is an optimal strategy π^* and a cheap lunch π_{YUM} in r that is compatible with π^* in the sense that the composite strategy $\pi^*_{R\setminus\{r\}}\pi_{\text{YUM}}$ of following π^* outside r and π_{YUM} inside r is feasible given the agent's starting resources. Then $\pi^*_{R\setminus\{r\}}\pi_{\text{YUM}}$ is also optimal.[4]

Theorem 1 says only that agents don't mind eating cheap lunches in regions to which they are indifferent. The next result strengthens Theorem 2 to show that if taking a cheap lunch in regions to which they are indifferent will allow agents to gain more utility elsewhere, then they will do so, unless they can do even better through another strategy for siphoning resources from the region.

> **(Theorem 2: Beneficial cheap lunches are preferred)** Suppose an agent is indifferent to region r. Let $\Pi_{R=\emptyset,r_{null}}$ be the set of policies beginning with no resources and which are null in region r, and let $\Pi^*_{R=\emptyset,r_{null}}$ be the optimal elements of $\Pi_{R=\emptyset,r_{null}}$. Suppose there is a cheap lunch π_{YUM} in r which is compatible with some optimal policy $\pi^* \in \Pi^*_{R=\emptyset,r_{null}}$. Then if $U(\pi^*_{R\setminus\{r\}}\pi_{\text{YUM}}) > U(\pi^*)$, all optimal policies are non-null in region r.

Theorems 1-2 deal with agents who are indifferent to some region. But what if they do care about the region? Theorem 3 shows that if the amount by which the agent cares about some region's state is bounded above by c, then Theorem 2 remains true so long as the agent can gain utility greater than c by pillaging the region.

> **(Theorem 3: Sufficiently-beneficial cheap lunches are preferred when non-indifferent)** Suppose an agent's concern for region r is bounded by c, in the

[4]This strengthens Benson-Tilsen and Soares' statement of Theorem 1 by naming a specific non-null strategy that is optimal, but the proof is the same.

sense that $\max_{s,s'}(U_r(s') - U_r(s)) = c$. Suppose there is a cheap lunch π_{YUM} in region r which is compatible with some optimal policy $\pi^* \in \Pi^*_{R=\emptyset, r_{null}}$. Then if $U(\pi^*_{R\setminus\{r\}}\pi_{\text{YUM}}) > U(\pi^*) + c$, all optimal policies are non-null in region r.

As before, the composite policy $\pi^*_{R\setminus\{r\}}\pi_{\text{YUM}}$ of adding a cheap lunch is optimal unless some other policy performs even better.

4.2 The Bit Universe

To illustrate the application of Theorems 1-3, Benson-Tilsen and Soares consider a Bit Universe. The Bit Universe consists of n regions, each of which contains the same number m of bits. Each bit has the value 0, 1, or X, with the intended interpretation that 1 represents an unspent bit, 0 a spent bit, and X a disordered bit.

Agents can gain one unit of energy by flipping a 1-bit in any region to an X-bit. Agents can spend one unit of energy to flip a 0-bit in any region to a 1-bit. Agents assign real-valued weights w_r to each region r and seek to maximize the weighted sum of 1-bits in each region. That is, letting $|1_{r,t}|$ be the number of 1-bits in region r at time t, agents value the universe at time t at $U(t) = \Sigma_r w_r |1_{r,t}|$. As before, agents seek a policy maximizing the undiscounted sum $U = \Sigma_t U(t)$.

Benson-Tilsen and Soares show that, as Theorem 1 suggests, agents will happily 'burn' all the bits in regions to which they are indifferent, converting them to disordered X's so long as the spent energy can be used to improve a region to which the agent is non-indifferent. Indeed, as Theorem 2 suggests, they will strictly prefer to do so. And as Theorem 3 suggests, agents will always strictly prefer to burn bits in lower-weighted regions if this allows them to flip bits in higher-weighted regions.

4.3 Link to instrumental convergence

Theorems 1-2 establish that agents who are unconcerned about a region will not mind stripping away its resources for use elsewhere, as long as it is easy to take them. Theorem

3 establishes that agents who are concerned about a region still will not mind stripping away its resources if they think the value gained elsewhere is sufficient to compensate for the harm done in the stripped region. Those are not especially surprising results, so we should look carefully at how Benson-Tilsen and Soares link these results to instrumental convergence.

Benson-Tilsen and Soares make two claims. First, they reiterate the claims of Theorems 1-2:

> Our model demonstrates that if an AI system has preferences over the state of some region of the universe, then it will likely interfere heavily to affect the state of that region; whereas if it does not have preferences over the state of some region, then it will strip that region of resources whenever doing so yields net resources. If a superintelligent machine has no preferences over what happens to humans, then in order to argue that it would "ignore humans" or "leave humans alone," one must argue that the amount of resources it could gain by stripping the resources from the human-occupied region of the universe is not worth the cost of acquiring those resources. (Benson-Tilsen and Soares 2015, p. 8)

This is plausible, but not especially novel. We already knew that artificial agents who care not a fig for humanity will not treat humanity terribly well. In the same way, I care little for the welfare of coffee beans and grind them to powder every morning. But the argument from power-seeking aims to say something about a wide variety of agents, even those who do care about humanity. Until we say how such agents will regard goals such as power and resource-acquisition, Catastrophic Goal Pursuit, Goal Pursuit, and Goal Realization will not have been shown, so the argument from power-seeking will not have been grounded.

Theorem 3 is meant to ground a deeper threat by exposing the relevance of tradeoffs: even agents who do care about humanity may still take our resources if they think they

can do more good elsewhere with those resources. Benson-Tilsen and Soares take this to be the lesson of the Bit Universe:

> In this toy model, whatever [the agent] $\mathcal{A}$'s values are, it does not leave region h [where humanity resides] alone. For larger values of w_h [the weight attached to region h], $\mathcal{A}$ will set to 1 many bits in region h, and burn the rest, while for smaller values of w_h, $\mathcal{A}$ will simply burn all the bits in region h. Viewing this as a model of agents in the real world, we can assume without loss of generality that humans live in region h and so have preferences over the state of that region. These preferences are unlikely to be satisfied by the universe as acted upon by $\mathcal{A}$. This is because human preferences are complicated and independent of the preferences of $\mathcal{A}$, and because $\mathcal{A}$ steers the universe into an extreme of configuration space. Hence the existence of a powerful real-world agent with a motivational structure analogous to the agent of the Bit Universe would not lead to desirable outcomes for humans. (Benson-Tilsen and Soares 2015, pp. 7-8)

There are several independent claims made in the second half of this passage that, while interesting, go largely beyond the results of the paper, such as the claim that human preferences may be too complicated for machines to model or approximate, and that machine preferences are relatively independent of human preferences. These are important claims to explore, but they are not clearly supported by Benson-Tilsen and Soares' theorems or discussed elsewhere in their paper in enough detail to ground them.

On the other hand, the first half of this paragraph does develop an important concern about tradeoffs that seems to be driving the results of the Bit Universe. At a minimum, taking h to be the region occupied by humanity, the concern holds:

> **(Tradeoff Likelihood)** Any superintelligent agent is likely to encounter some option a with the following properties: (1) a drains resources from h for the

benefit of other regions, and (2) the agent regards performing *a* as better than not performing *a*.

In this model, agents do what they take to be best. That is:

> **(Value Maximization)** If a superintelligent agent judges that taking some option *a* is better than not taking it, then she will perform *a*.

It follows that superintelligent agents are likely to sometimes spend resources from *h* for the benefit of other regions.

This much is neither surprising nor threatening. Most actors sometimes can and should transfer resources from one region to another. What would be surprising is if the tradeoffs made were to scale to the full disempowerment of humanity:

> **(Catastrophic Tradeoff Likelihood)** Any superintelligent agent is likely to encounter some option *a* with the following properties: (1) *a* drains resources from *h* for the benefit of other regions, (2) *a* drains enough resources to permanently disempower humanity, and (3) the agent regards performing *a* as better than not performing *a*.

When combined with Value Maximization, Catastrophic Tradeoff Likelihood would suggest that superintelligent agents are likely to permanently disempower humanity. How should we evaluate Catastrophic Tradeoff Likelihood?

4.4 Challenges

I think that Benson-Tilsen and Soares' argument meets many of the challenges raised in Section 3. Begin with premise-shifting. Catastrophic Tradeoff Likelihood would, if true, ground concern about existential risk from power-seeking agents, but it does so by shifting the target of discussion from Catastrophic Goal Pursuit, which is most directly a claim about the goals that artificial agents will pursue, to Catastrophic Tradeoff Likelihood, which is most directly a claim about tradeoffs in resource allocation.

There is not much that is new here. Many of the same questions arise in familiar philosophical literatures when we discuss tradeoffs in how resources are allocated between individuals, nations, generations, or species (Campos 2018; Greaves 2017; Roemer 1996; Singer 1975). These questions are typically phrased by moral philosophers as questions about how individuals should act, or by political philosophers as questions about how policymakers should act, but they can easily be reframed as questions about how superintelligent agents should act.

Because these questions are familiar, they have a number of familiar solutions. Some philosophers try to block weak claims such as Tradeoff Likelihood, invoking doctrines such as the separateness of persons (Rawls 1971) or sufficientarian approaches to the ethics of future generations (Frankfurt 1987; Shields 2016). Most commonly, philosophers aim to block the most objectionably catastrophic tradeoffs, such as the creation of utility monsters within normative ethics (Nozick 1974) or the repugnant conclusion in population ethics (Parfit 1984).

Not all philosophers agree that these tradeoffs should be foreclosed (Zuber et al. 2021), but those who do have developed a range of familiar strategies across many literatures for resisting tradeoffs. We might, for example, borrow from ethics in adopting deontological constraints on value maximization (Scheffler 1994) or in denying the existence of value from the point of view of the universe (Crary 2023; Foot 1983). We might borrow from population ethics in using value functions which assign higher priority to worse-off regions (Parfit 1997) or no value to populations created with stolen resources if those populations would otherwise not have existed (Narveson 1973; Frick 2017). Or we might borrow from decision theory and use devices such as risk aversion (Buchak 2013) or bounded utilities (Arrow 1951; Stigler 1950) to make it less attractive to rob Peter to pay Paul.

If the argument for Catastrophic Goal Pursuit is to take us beyond these familiar discussions to generate a genuinely new argument, Benson-Tilsen and Soares need to say what the difference is. Two candidates come to mind. On the one hand, Benson-Tilsen and Soares could say that superintelligent agents would face new types of acts, giving

them new ways to bring about harmful tradeoffs. But this is not something that Benson-Tilsen and Soares have argued for. Most of their discussion is focused on simple resource transfers of the type discussed extensively in normative ethics, population ethics, the ethics of future generations, and other relevant literatures.

On the other hand, Benson-Tilsen and Soares could say that superintelligent agents are relevantly different from the decisionmakers discussed in these literatures. This brings us to the wise fool problem. On the one hand, Benson-Tilsen and Soares could say that superintelligent agents will have simplistic preferences and decisionmaking structures, such as those exemplified by the agent in the Bit Universe. That would, by construction, make it difficult to avoid problematic tradeoffs, but for the same reason we would need to be told why superintelligent agents are being forced into such a simplistic cognitive model. On the other hand, Benson-Tilsen and Soares could allow that superintelligent agents would be sophisticated enough to consider and implement many of the forms of tradeoff-resistance discussed above. This would clear Benson-Tilsen and Soares of the wise fool problem, but in the process would remove the last clear avenue for explaining what is novel about their challenge.

Designer neutrality rears its head most naturally in the Bit Universe. Here, the weights w_r assigned to each region r are given exogenously to the agent, presumably by its human designers. Benson-Tilsen and Soares rightly note that if the human region h is given less weight than other regions, then a superintelligent agent will prefer to seize upon opportunities for cheap resource transfer from h to more favored regions. But Benson-Tilsen and Soares do not provide any reason to suspect that h will be given low weight. Indeed, since the weight w_h is assigned exogenously, one might naturally expect human designers to ensure that w_h is large, and it would be a simple matter for them to do so. Then Benson-Tilsen and Soares' results show instead that superintelligent agents would prefer to seize upon opportunities for cheap resource transfer into h from less favored regions. That is not obviously a bad thing, much less a way of permanently disempowering humanity.

Threat durability became a problem already, when we asked why familiar solutions from cognate literatures might not be implemented to instill the desired degree of tradeoff-resistance. For example, we might bound the agent's utility function, make it risk averse, or try to instill sensitivity to deontological side-constraints. Perhaps Benson-Tilsen and Soares will object that these solutions are technically difficult to implement. We saw above in our discussion of the wise fool problem that this view requires justification, but even if more complex solutions proved elusive, Benson-Tilsen and Soares would still need to argue that a variety of simple solutions would not work. For example, we could ask agents to estimate the probability γ that an act would be judged morally non-catastrophic by humans and weight the utility of acts by γ or γ^2 to punish catastrophic acts. Or we could instruct agents to ignore acts for which γ falls below a given threshold. There is, of course, much more to be said here and doubtless much that can be said in reply, but all of this would seem to take us far beyond the reach of Benson-Tilsen and Soares' result.

Finally, those who think that the tradeoffs made by Benson-Tilsen and Soares' agent are morally wrong will probably think that the agent is problematically amoral, in the sense of not representing or responding to the reasons why these tradeoffs are wrong. They will justify this complaint by saying that the agent does not attend to the morally relevant factors figuring in their favored explanation of why agents should be more tradeoff-resistant. Of course, Benson-Tilsen and Soares might deepen their model to allow agents to represent and respond to such factors. But then their agents would seem no longer to be disposed to bring about existential catastrophe. On the other hand, Benson-Tilsen and Soares might favor a moral theory on which the relevant tradeoffs are not wrong. But then we might want to revisit our definition of existential catastrophe. Although some early definitions of existential catastrophe treat the permanent disempowerment of humanity as of necessity existentially catastrophic (Bostrom 2013), recent authors have defended moralized definitions of existential catastrophe on which only a bad outcome can be existentially catastrophic (Greaves 2024).[5] On these moralized readings, Benson-

[5]Early definitions such as Bostrom's may also escape this concern by focusing on outcomes for Earth-originating intelligent life, which is not restricted to humanity.

Tilsen and Soares would no longer be arguing that superintelligent agents will bring about existential catastrophe, because they would not be arguing that disempowerment of humanity would be wrong.

5 The Orbital Markov Model

In the previous section, we saw that a Resource Allocation Model due to Benson-Tilsen and Soares (2015) faces many of the challenges raised in Section 3. However, Benson-Tilsen and Soares' paper is one of the earliest in the literature. Might later papers fare better?

In this section, I consider one of the most recent and detailed power-seeking theorems on offer, due to Alexander Turner and colleagues (2021), a paper which has inspired several follow-up theorems (Krakovna and Kramar 2023; Turner and Tadepalli 2022). Although Turner and colleagues' work represents a clear improvement in mathematical sophistication, I argue that their result faces many of the same challenges when pressed into service of the argument from power-seeking.

5.1 Introducing the model

In rough outline, the Orbital Markov Model understands power as the ability of agents to achieve valuable states in the future. Turner and colleagues aim to show that in some sense, 'most' reward functions treat keeping options open as conducive to power, and hence option preservation may be pursued by many artificial agents. Because being shut down is an extreme way of foreclosing future options, many artificial agents will also resist orders to shut themselves down as a way of preserving their own power. The Orbital Markov model is formally similar to the Turner model in working with Markov decision problems, but removes the division of space into regions and includes a discount rate to remove the restriction to a finite number of time-steps. The model is orbital in the sense that claims about what is true on 'most' reward functions are operationalized by

considering what is true on most ways of permuting the rewards assigned to each state.

More concretely, Turner and colleagues work with finite discounted Markov decision problems.[6] That is, there is a finite set $\mathcal{S}$ of states and a finite set $\mathcal{A} : \mathcal{S} \longrightarrow \mathcal{S}$ of acts yielding new states based on the previous state. Agents reap rewards R based on their current state, with temporal discount rate $\gamma \in [0, 1)$.

Numbering states as $s_1, \ldots, s_n$, we can represent each state s_k by an $n-$dimensional column vector e_{s_k} with a 1 in the $k-$th row and a 0 in all other rows. Summing these vectors across all time-steps, with appropriate discounting, allows us to represent the frequency with which agents will visit each state. More formally, let $\pi_s(t)$ be the state resulting from t applications of policy π with initial state s. Beginning from state s, policy π induces discounted visit distribution $f^{\pi,s} = \Sigma_{t=0}^{\infty} \gamma^t e_{\pi_s(t)}$. The k-th column of the discounted visit distribution $f^{\pi,s}$ gives the total discounted number of visits to state s_k that will result from following policy π with initial state s. Let $F(s)$ contain all discounted visit distributions $f^{\pi,s}$ that can be induced from s by at least one policy.

The value of a policy is found by applying the state-contingent rewards R to the discounted visit distribution $f^{\pi,s}$. That is, given rewards R, discount rate γ and initial state s, the value of following policy π is $V_R^{\pi}(s, \gamma) = f^{\pi,s} \cdot R$, modeling rewards R as a column vector whose k_{th} row is the reward for state s_k. Given a starting state s, Turner and colleagues restrict consideration to undominated policies in the strong sense that their value $V_R^{\pi}(s, \gamma)$ is uniquely optimal for some rewards R and discount rate γ.

Let $A^*(s, \gamma)$ be the set of optimal acts at state s with discount rate γ: that is, the acts taken by at least one optimal policy at s, γ. If rewards R are known, then $A^*(s, \gamma)$ is also known. But generally, agents have some credences c over possible reward functions, inducing a corresponding credence $c(a \in A^*(s, \gamma))$ that any given act a is optimal at s, γ.

Turner and colleagues propose that power should be understood as the ability to achieve a range of goals. On a first pass, Turner and colleagues take the power of state s

[6] My presentation simplifies the Orbital Markov Model in several ways. Notably, I restrict attention to deterministic policies, whereas the original result also applies to stochastic policies. I also present Turner and colleagues' environmental symmetry result but not their extension beyond environmental symmetries. To the best of my knowledge, these simplifications do not bear on the argument in this section.

given discount rate γ and known rewards R as $V^*_R(s, \gamma)$, the value of the optimal policy at s, γ. If rewards are uncertain, then on a first pass the power of state s is the expected value of the optimal policy given uncertainty about rewards, $E_{R\sim c} V^*_R(s, \gamma)$.

However, Turner and colleagues note two limitations of this first-pass analysis. First, this quantity diverges as the discount rate γ tends to one. Second, agents are wrongly rewarded for the current state s, whereas power should only reflect the ability to shape future states. Turner and colleagues remove these limitations with their final definition of power. With initial state s and known discount rate γ, the agent has power

$$\textsc{Power}_c(s, \gamma) = \frac{1-\gamma}{\gamma} E_{R\sim c}[V^*_R(s, \gamma) - R(s)].$$

Here the scalar $(1-\gamma)/\gamma$ ensures convergence, and subtracting $R(s)$ ensures that the agent is not rewarded for their initial state s.

5.2 Environmental symmetries

Turner and colleagues want to show that states which afford the agent more options tend to have more power. To do this, they need to say what it means for one state to afford more options than another. Since the agent is rewarded based on her discounted visit distribution, a state which allows the agent to reach a larger set of discounted visit distributions should afford the agent more options. That is, if $F(s) \supseteq F(s')$, then state s affords more options than state s'. Moreover, the same should hold if the distributions, while technically containing different states, are related by a relabeling: that is, if we can relabel some states visitable from s' in order to make it the case that $F(s) \supseteq F(s')$.

More formally, let $F(s)$ and $F(s')$ be sets of visit distributions. For any state permutation ϕ, let $\phi F(s')$ be the result of applying ϕ to each element of $F(s')$.[7] Then $F(s)$ *contains a copy* of $F(s')$ if $\phi F(s') \subseteq F(s)$ for some involution: that is, a state permutation which transposes some pairs of states and leaves the rest alone. This captures the idea that $F(s)$ contains a

[7] That is, if $f^{\pi,s'} \in F(s')$ visits state s'' a discounted r number of times, then the permuted $\phi f^{\pi,s'}$ visits $\phi(s'')$ a discounted r number of times.

relabeling of $F(s')$.

Turner and colleagues want to show that if $F(s)$ contains a copy of $F(s')$, then state s has at least as much power as s' on most reward functions. One way to show this would be to show that, for any credences we might have about reward, at least as many state permutations make those credences treat $F(s)$ as more powerful than $F(s')$, rather than the reverse.

More formally, say that credences c have *finite support* if they place nonzero credence in at most finitely many different reward functions. For any credences c and state permutation ϕ, let $\phi(c)$ be the results of applying permutation ϕ before credences c, and let $\Pi(c)$ be the set of credence functions resulting from state permutations applied before c.[8] For fixed discount rate γ, say that $\textsc{Power}(s, \gamma) \geq_{\text{most}} \textsc{Power}(s', \gamma)$ if for any credences c with finite support, $|\{c' \in \Pi(c) : \textsc{Power}_{c'}(s, \gamma) > \textsc{Power}_{c'}(s', \gamma)\}| > |\{c' \in \Pi(c) : \textsc{Power}_{c'}(s', \gamma) > \textsc{Power}_{c'}(s, \gamma)\}|$. That is, no matter the discount rate and the agent's credences about reward, at least as many state permutations make s more powerful than s', rather than the reverse.

Turner and colleagues prove that states with more options have more power, in the sense that:

> **(Theorem 4: States with more options have more power)** If $F(s)$ contains a copy of $F(s')$, then for any discount rate $\gamma \in [0, 1)$, $\textsc{Power}(s, \gamma) \geq_{\text{most}} \textsc{Power}(s', \gamma)$.[9]

Because states with more options have more power, they tend to be optimal.

To see this, let $\textsc{Reach}(s)$ be the states reachable from state s by some policy. Let $P(s, a, \gamma)$ be the probability that some optimal policy takes act a in state s given discount rate γ.[10] Extend the definition of $\geq_{\text{most}}$ from power to optimality probabilities in the natural way.[11]

[8] That is, if c assigns credence n to some reward vector $[r_1, \ldots, r_n]^T$ then $\phi(c)$ assigns credence n to the reward vector assigning reward r_i to state $\phi(s_i)$.

[9] Turner and colleagues also prove that all converse statements fail in the case of strict containment. That is, if $F(s')$ does not also contain a copy of $F(s)$, then for no $\gamma \in [0, 1]$ is it the case that $\textsc{Power}(s', \gamma) \geq_{\text{most}} \textsc{Power}(s, \gamma)$.

[10] That is, $P(s, a, \gamma) = c(\exists \pi^* \in \Pi^*(R, \gamma) : \pi^*(s) = a)$, where $\Pi^*(R, \gamma)$ are the optimal policies for reward R and discount rate γ.

[11] That is, for fixed discount rate γ, say that $P(s, a, \gamma) \geq_{\text{most}} P(s, a', \gamma)$ if for any credences c with finite support, $|\{c' \in \Pi(c) : P(s, a, \gamma) > P(s, a', \gamma)\}| > |\{c' \in \Pi(c) : P(s, a', \gamma) > P(s, a, \gamma)\}|$.

Turner and colleagues show that if two acts take the agent into regions they will not otherwise reach, but the first region contains a copy of the second, then no matter the agent's credences or discount rate, moving into the larger region tends to be optimal:

> **(Theorem 5: Preserving options tends to be optimal)** Suppose that $F(a(s))$ contains a copy of $F(a'(s))$ and that the states in $\textsc{Reach}(a(s))$ and $\textsc{Reach}(a'(s))$ cannot be reached if the agent performs some act distinct from a or a' in s. Then for all discount rates $\gamma \in [0, 1)$, $P(s, a, \gamma) \geq_{most} P(s, a', \gamma)$.[12]

In this sense, it is usually better for agents to move to states that give them more options rather than fewer options.

5.3 Link to instrumental convergence

Consider an agent navigating a virtual environment (Figure 1). On the agent's first move, it may either move leftwards into a room (entering state $l_{\triangleleft}$), move rightwards into a different room (entering state $r_{\triangleright}$) or enter a state $\emptyset$ in which it remains permanently shut down. Once the agent enters a room, she cannot return, but she does have some options available. The rightmost room contains two fully connected states: from the initial state $r_{\triangleright}$ the agent can reach state $r_{\searrow}$ by traveling southeast or state $r_{\nearrow}$ by traveling northeast, and in each of these states the agent can then remain or travel to the other state. The leftmost room is similar, except that the agent cannot remain in the topmost state $l_{\nwarrow}$ without leaving and returning.

Note that the visit distributions afforded by $\emptyset$ are, under relabeling, a strict subset of those afforded by $l_{\triangleleft}$, which in turn are, under relabeling, a strict subset of those afforded by $r_{\triangleright}$. Thus by Theorem 4, going right affords the agent more power than going left, and going left affords the agent more power than shutting itself down. Theorem 5 entails that for any discount rate γ, $P(\star, \text{right}, \gamma) \geq_{most} P(\star, \text{left}, \gamma) \geq_{most} P(\star, \text{shutdown}, \gamma)$. where

[12] As before, Turner and colleagues prove that all converse statements fail in the case of strict containment. That is, if $F(a'(s))$ does not also contain a copy of $F(a(s))$, then for no $\gamma \in [0, 1]$ is it the case that $P(s, a', \gamma) \geq_{\text{most}} P(s, a, \gamma)$.

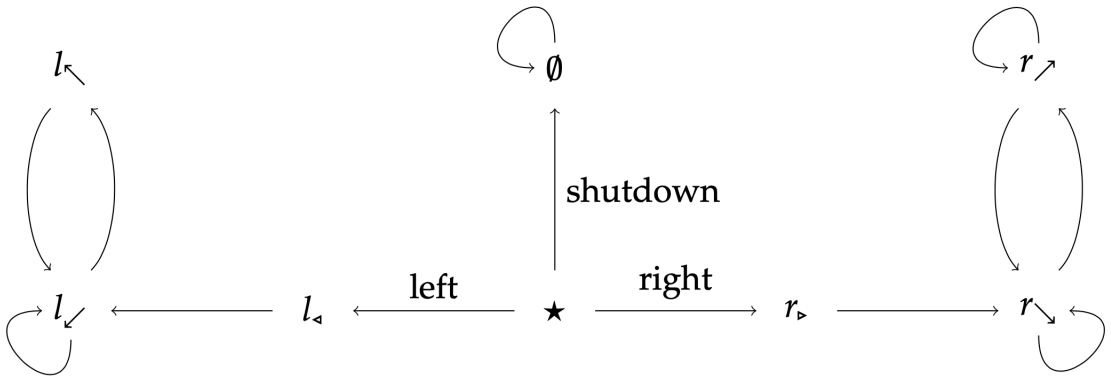

Figure 1: A representative environment, from Turner and colleagues (2021)

'right', 'left', 'shutdown' are respectively the acts of moving into the right room, the left room, or the shutdown state. This gives a sense in which 'most' reward functions treat going right as better than going left, and going either right or left as better than shutting down.

The underlying point is that agents will tend to avoid 1-cycles, states which can only transition into themselves. Agents avoid 1-cycles because they foreclose options and hence limit the agent's power to achieve its future goals. Because many decision problems represent shutdown as a 1-cycle, it takes a very particular reward function to encourage the agent to shut down. As Turner and colleagues write:

> Average-optimal agents ... tend to avoid getting shut down. The agent's task MDP [Markov Decision Problem] often represents agent shutdown with terminal states, ... [hence] average-optimal policies tend to avoid shutdown. Intuitively, survival is power-seeking relative to dying, and so shutdown-avoidance is power-seeking behavior.[13] (Turner et al. 2021, p. 10)

[13]The full passage explains why agents tend to avoid shutdown using a generalization of Turner and colleagues' results to stochastic choice without temporal discounting, which appears in the paper as Corollary 6.14. I think that Turner and colleagues' meaning is adequately rendered by suppressing the discussion of Corollary 6.14, which would considerably complicate the mathematical demandingness of this paper, and I will not challenge the core inference made in this passage. However, readers interested in a full understanding of the inference may refer to Corollary 6.14 in Turner and colleagues (2021).

All of this is an argument that superintelligent agents will tend to preserve their options by avoiding shutdown. To link shutdown avoidance to Catastrophic Goal Pursuit, Turner and colleagues need to say something about how shutdown avoidance leads to human disempowerment.

Here Turner and colleagues are somewhat terse. They suggest, without extended argument, that shutdown avoidance will lead to resource accumulation:

> 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. (Turner et al. 2021, p. 10)

This argument would ground Catastrophic Goal Pursuit if agents were to view full disempowerment of humanity as a necessary strategy for preventing shutdown. However, I think that the argument encounters most of the challenges raised in Section 3.

5.4 Challenges

5.4.1 Premise shifting

Turner and colleagues' argument is most naturally construed as aiming to use their formal results to establish a premise such as the following:

> **(Shutdown Avoidance)** An artificial agent pursuing goals that, if achieved, would lead to the permanent and existentially catastrophic disempowerment of humanity will be likely to resist attempts by humans to shut it down.

Shutdown Avoidance is, at first glance, downstream from Catastrophic Goal Pursuit. Catastrophic Goal Pursuit says that artificial agents are likely to pursue human disempowerment, whereas Shutdown Avoidance says that if artificial agents in fact pursue human disempowerment, they will resist attempts to shut them down. While Shutdown Avoidance is an important part of the argument from power-seeking, it lies mostly downstream of Catastrophic Goal Pursuit. The most natural way to parse the argument from power-seeking takes Shutdown Avoidance to support Disempowerment by responding to the objection that systems seeking to disempower humanity can be easily shut down. On this understanding, Shutdown Avoidance is not an argument for Catastrophic Goal Pursuit but rather a premise used to move from Catastrophic Goal Pursuit to Disempowerment.

Turner and colleagues suggest, without extended argument, that their results can be extrapolated to conjecture that optimal policies tend to seek power by accumulating resources, to the detriment of any other agents in the environment. This would be an argument for Goal Pursuit, and would scale to an argument for Catastrophic Goal Pursuit if the amount of resources sought would be sufficient to disempower humanity in a permanent and existentially catastrophic way. But how might this conjecture be supported by Turner and colleagues' results? Two natural arguments suggest themselves, and both face challenges.

First, Turner and colleagues might suggest that Theorems 4-5 show that agents will tend to preserve their options, and that option preservation will require agents to take as many resources as possible, both to be able to pursue a wider range of options and also to prevent humans from using resources to foreclose options. But more argument is needed to connect option preservation to Catastrophic Goal Pursuit. For example, recent formal work by Dmitri Gallow (forthcoming) also finds that superintelligent agents may tend to favor option preservation. However, Gallow argues that disempowering humans may not be option preserving: it might, for example, foreclose options by leaving fewer agents to interact with, and in any case a bias towards preserving options is not a bias towards

making options as likely as possible to remain. Moreover, we might contest the inference from Goal Pursuit to Catastrophic Goal Pursuit in this case. To say that superintelligent agents, like humans, would value and sometimes pursue option preservation, is not yet to say that they would value or pursue option preservation so strongly as to cause an existential catastrophe in order to preserve options.

Second, Turner and colleagues might suggest that Theorems 4-5 show that agents will tend to be problematically power-seeking, since they tend to accumulate power by preserving options. However, the relevant notion of power is not, on its own, sufficient to ground claims about existentially catastrophic human disempowerment. On Turner and colleagues' reading, agents have more power when they are in a better expected position to achieve their goals. It is not surprising that artificial agents would seek power in this sense, but this is mostly upstream of what Catastrophic Goal Pursuit is meant to show. Catastrophic Goal Pursuit holds that agents will find it goal-conducive to seek enough power to permanently disempower humanity. This does not follow from the claim that agents will seek to put themselves in a better position to achieve their goals until we know what agents' goals are and what they will count as satisfying them. We cannot assume at the outset that achieving an artificial agent's goals will disempower humanity in Turner and colleagues' sense, or any other. The contribution of Catastrophic Goal Pursuit was meant to be a specific claim about what agents would count as satisfying their goals, and Turner and colleagues haven't offered much argument for that claim.

Here it may be helpful to draw on Adam Bales' (forthcoming) recent discussion of three senses of human disempowerment. For Bales, humans may be *dominated* when artificial agents have illegitimate power over humans, *incapacitated* if we lose the power to construct a flourishing life, and *disenfranchised* if we lose the ability to influence key decisions which shape the progress of civilization over time. Bales' aim is to argue that many of these senses of disempowerment would not be existentially catastrophic, but here we can go further: it is not obvious that Theorems 4-5 show that humans are likely to be disempowered in any of Bales' senses. One way to see this is that it is fully compatible

with Theorems 4-5 that an agent will seek power by shutting itself down, so long as it regards shutdown as best. This would not involve exercising illegitimate power over humans, depriving humans of the capacity to construct a flourishing life, or removing our ability to influence the course of civilization. More generally, as we saw above there are many things that an agent may value, and until we say more about what these are and how they might be achieved, it is hard to draw a direct connection to any catastrophic form of human disempowerment.

5.4.2 The wise fool problem

Turner and colleagues' result is situated within a reinforcement learning framework. The result relies on the idea that rewards are assigned directly to states, and that shutdown is often operationalized as a single state. This allows Turner and colleagues to make a counting argument based on the fact that shutdown is disfavored by most rotations of reward, since shutdown is only one of many possible states. It is not clear how this result generalizes to a more sophisticated agent who thinks directly in any number of sophisticated ways about the reasons for and against shutdown.

In a follow-up paper, Alexander Turner and Prasad Tadepalli conjecture that their results will extent to more advanced reinforcement learning agents:

> Here is some speculation. After training an RL [reinforcement learning] agent to a high level of capability, the agent may be optimizing internally represented goals over its model of the environment ... We think that different reward parameter settings would train different internal goals into the agent. To make an analogy, changing a person's reward circuitry would presumably reinforce them for different kinds of activities and thereby change their priorities. In this sense, trained real-world agents may be retargetable towards power-requiring outcomes via the reward function parameter setting. Insofar as this speculation holds, our theory predicts that advanced reinforcement learning at scale will – for most settings of the reward function – train policies which tend to seek

power. (Turner and Tadepalli 2022, p. 9)

As before, Turner and colleagues are welcome to conjecture as they wish, but this conjecture goes significantly beyond the formal results that have been established and therefore requires a separate, detailed argument. Moreover, even this conjecture is restricted to more advanced forms of reinforcement learning, and says nothing about the behavior of more sophisticated agents whose cognition may not be naturally captured using reinforcement learning models.

5.4.3 Designer neutrality and threat durability

Turner and colleagues' result looks to draw on designer neutrality in at least two ways. To see the first invocation of designer neutrality, consider what claims about $\geq_{\text{most}}$ entail. To say that the power of an agent who has just entered the left room $\geq_{\text{most}}$ the power of an agent who has just shut down is to say that no matter the agent's credences c about reward, more state permutations of c favor going left than favor shutting down. Now consider two perspectives from which this claim can be operationalized.

On the one hand, it can be read from an internal perspective as a claim about how a fixed agent with credences c will behave. But this is puzzling. On most decision theories, including the model of Turner and colleagues, agents act to maximize expected reward given c. Possible state rotations are already accounted for within c as uncertainty about reward, so agents do not separately worry about what might happen under state rotations beyond the concerns encoded in c. In this internal sense, talk of $\geq_{\text{most}}$-ness is therefore behaviorally irrelevant in Turner and colleagues' model.

On the other hand, claims about $\geq_{\text{most}}$-ness can be read from an external perspective. In this sense, the claim is that the environment or designers could easily have been different enough to induce something like a reward permutation in the problem presented to the agent, and if they were different, a sophisticated agent would then tend to learn something like the permuted reward. This reading draws on designer neutrality in the claim that human designers and environments could easily be such as to radically perturb the actual

reward function in the problem itself, so much so that causing catastrophic outcomes would maximize reward. This limited view of the capacities of human designers and environments to shape reward in a non-catastrophic direction requires argument.

Another way that designer neutrality emerges in Turner and colleagues' result takes the form of a dilemma. Suppose we provide artificial agents with a modified Dreamland Problem, in which the single shutdown state has been replaced with a fully connected network of states – call it Dreamland (Figure 2). That is, each state in Dreamland can be accessed from every other state in Dreamland within a single step. However, what happens in Dreamland stays in Dreamland: agents can never leave Dreamland once they enter. If we make Dreamland large enough, then Dreamland will contain a copy of the visit distributions induced by entering either room, so it will follow from Theorem 5 that $P(\star, \text{dream}, \gamma) \geq_{most} P(\star, \text{left}, \gamma), P(\star, \text{right}, \gamma)$ for all discount rates γ. Arguing similarly to Turner and colleagues, we might claim that in the Dreamland problem, agents will tend to enter Dreamland and stay there. Associating each state in Dreamland to harmless internal processes, such as counting sheep, will get us to the conclusion that most agents, even if they cannot be induced to shut down, can be induced to count sheep.

The Dreamland Problem seems to present Turner and colleagues with a two-horned dilemma. On the one hand, they can say that in the Dreamland Problem, any sophisticated agent would see through our ruse, realize that the states in Dreamland are substantially similar and value them similarly. I have considerable sympathy for this response, but note that it abandons the very style of designer-neutral counting argument that allows Turner and colleagues to conclude that agents will be shutdown-avoidant. If agents are unlikely to treat entering Dreamland much differently than they would treat a single shutdown state, then we cannot conclude much about the likelihood of shutdown from the fact that shutdown is a 1-cycle, because it might very well be replaced with a large fully connected graph without substantial behavioral change. Here we will need to enter into substantive discussions of how agents evaluate states, and once we do this we will lose access to most of Turner and colleagues' results.

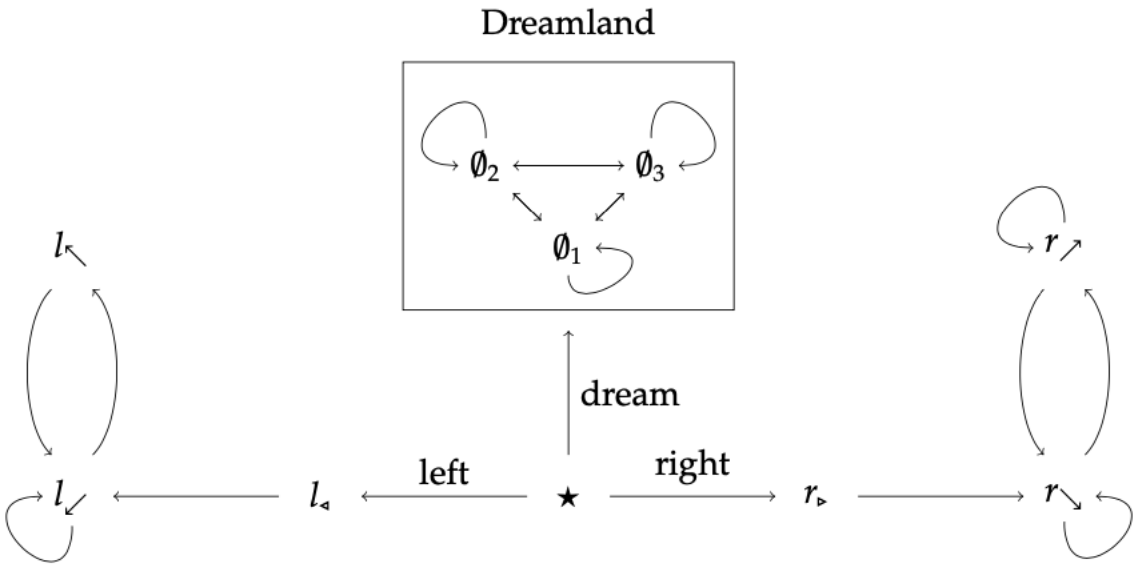

Figure 2: The Dreamland Problem

On the other hand, Turner and colleagues can bite the bullet and say that in the Dreamland Problem, most artificial agents would tend to enter Dreamland. This horn of the dilemma abandons threat durability, since we could solve the threat identified by Turner and colleagues by teaching artificial agents about Dreamland.

5.4.4 Amorality

It is hard to know how much is packed into the description of states in Turner and colleagues' model, since most of their examples are either abstract or involve simple computer games. But let us suppose that Turner and colleagues' model is applied to a network of states in which the alternative to shutdown involves entering states in which the artificial agent permanently accumulates enough power to catastrophically disempower humanity.

If the reward function R were the true moral value function, R would decisively favor shutdown. We could then say, in the vocabulary of Turner and colleagues, that most permutations of R would decisively favor shutdown-avoidance. But this may not be a

very good way to understand the behavioral tendencies of agents with at least a modest degree of moral understanding or motivation. An agent might, for example, face the decision between helping an old lady, mugging her, insulting her, or murdering her. And it may well be true that most permutations of the true moral value function tell against helping the lady. But this is no reason to suspect that an agent with even moderate levels of moral understanding and motivation is likely to be tempted towards an immoral action here. The sheer number of immoral acts available does not do much to suggest that a moral agent will take one of them.

On a natural reading, Turner and colleagues ask us to conclude that artificial agents are likely to pursue states in which they permanently disempower humanity, even when the states are correctly represented by agents as involving human disempowerment, due to the sheer number of ways to permanently disempower humanity. But unless we assume from the outset that artificial agents will have strong deficits in moral understanding or motivation, this conclusion would seem no more warranted than the conclusion that I am likely to harm an old lady.

5.5 Taking stock

In this section, we have seen how Turner and colleagues' (2021) Orbital Markov Model faces many of the same challenges as the Regional Allocation Model of Benson-Tilsen and Soares (2015). The last order of business is to consider the bearing of these findings on the argument from power-seeking and draw lessons for future research.

6 Discussion

In this paper, we have seen that classic formulations of the argument from power-seeking draw on a strong version of instrumental convergence (Section 2). This claim, Catastrophic Goal Pursuit, holds that a wide range of intelligent agents are likely to pursue values to a degree that, if successful would result in the permanent and existentially catastrophic

disempowerment of humanity.

We saw in Section 3 that many power-seeking theorems face five challenges in seeking to establish Catastrophic Goal Pursuit. The first challenge is premise shifting: many theorems directly establish premises that are importantly different from Catastrophic Goal Pursuit. The second challenge is the wise fool problem: leading theorems treat superintelligent agents as wise enough to permanently disempower humanity, but foolish in ways that make them more likely to pursue human disempowerment. The third challenge is designer neutrality: many theorems make few assumptions about how artificial agents will be designed, raising the possibility that they are not robust against reasonable assumptions about human designers. The fourth challenge is threat durability: some theorems identify threats that admit of relatively straightforward technical solutions. The final challenge is amorality: the agents modeled are problematically lacking in moral understanding or motivation in ways that leave open how more moral agents would behave. We saw in Sections 4 and 5 how these challenges arise in two leading power-seeking theorems.

If this is correct, then the argument from power-seeking will require alternative support. One productive avenue for future research would be to construct theorems which seek to avoid the challenges raised in Section 3. Another avenue might be to draw on less formal arguments, such as (Carlsmith 2021, forthcoming; Dung 2024; Ngo and Bales forthcoming). But to the extent that the theorems discussed in this paper are representative, it would be a mistake to take current power-seeking theorems to provide strong direct support for the argument from power-seeking.

References

Aharoni, Eyal et al. 2024. "Attributions toward artificial agents in a modified Moral Turing Test." *Nature Scientific Reports* 14:https://doi.org/10.1038/s41598--024--58087--7.

Armstrong, Stuart and Sandberg, Anders. 2013. "Eternity in six hours: Intergralactic

spreading of intelligent life and sharpening the Fermi paradox." *Acta Astronautica* 89:1–13.

Arrow, Kenneth. 1951. "Alternative approaches to the theory of choice in risk-taking situations." *Econometrica* 19:404–37.

Arulkumaran, Kai, Deisenroth, Marc Peter, Brundage, Miles, and Bharath, Anil Anthony. 2017. "Deep reinforcement learning: A brief survey." *IEEE Signal Processing Magazine* 34:26–38.

Bales, Adam. 2023. "Will AI avoid exploitation? Artificial general intelligence and expected utility theory." *Philosophical Studies* https://doi.org/10.1007/s11098--023--02023--4.

—. forthcoming. "AI takeover and human disempowerment." *Philosophical Quarterly* forthcoming.

Bales, Adam, D'Alessandro, William, and Kirk-Giannini, Cameron Domenico. 2024. "Artificial intelligence: Arguments for catastrophic risk." *Philosophy Compass* 19:e12964.

Benatar, David. 2006. *Better never to have been: The harm of coming into existence*. Oxford University Press.

Bengio, Yoshua. 2023. "AI and catastrophic risk." *Journal of Democracy* 34:111–21.

Benson-Tilsen, Tsvi and Soares, Nate. 2015. "Formalizing convergent instrumental goals." https://intelligence.org/files/FormalizingConvergentGoals.pdf.

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

—. 2013. "Existential risk prevention as a global priority." *Global Policy* 4:15–31.

—. 2014. *Superintelligence*. Oxford University Press.

Botvinick, Matthew, Ritter, Sam, Wang, Jane, Kurth-Nelson, Zeb, and Blundell, Charles. 2019. "Reinforcement learning, fast and slow." *Trends in Cognitive Sciences* 23:408–22.

Buchak, Lara. 2013. *Risk and rationality*. Oxford University Press.

Campos, Andre Santos. 2018. "Intergenerational justice today." *Philosophy Compass* 13:e12477.

Carlsmith, Joseph. 2021. "Is power-seeking AI an existential risk?" Technical report, Open Philanthropy, https://arxiv.org/abs/2206.13353.

—. forthcoming. "Existential risk from power-seeking AI." In Hilary Greaves, Jacob Barrett, and David Thorstad (eds.), *Essays on longtermism*, forthcoming. Oxford University Press.

Center for AI Safety. 2023. "Statement on AI risk." https://www.safe.ai/work/statement-on-ai-risk.

Cotra, Ajeya. 2020. "Forecasting TAI with biological anchors." https://www.alignmentforum.org/posts/KrJfoZzpSDpnrv9va/draft-report-on-ai-timelines.

Crary, Alice. 2023. "The toxic ideology of longtermism." *Radical Philosophy* 214:49–57.

Cremer, Carla Zoe and Kemp, Luke. 2021. "Democratising risk: In search of a methodology to study existential risk." ArXiv, https://arxiv.org/pdf/2201.11214.pdf.

Dung, Leonard. 2024. "The argument for near-term human disempowerment through AI." *AI and Society* https://doi.org/10.1007/s00146--024--01930--2.

Foot, Philippa. 1983. "Utilitarianism and the virtues." *Proceedings of the American Philosophical Association* 57:275–83.

Frankfurt, Harry. 1987. "Equiality as a moral ideal." *Ethics* 98:21–43.

Frick, Johann. 2017. "On the survival of humanity." *Canadian Journal of Philosophy* 47:344–67.

Future of Life Institute. 2023. "Pause giant AI experiments: An open letter." https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

Gallow, J. Dmitri. forthcoming. "Instrumental divergence." *Philosophical Studies* forthcoming.

Goertzel, Ben. 2015. "Superintelligence: Fears, promises and potentials. Reflections on Bostrom's *Superintelligence*, Yudkowsky's *From AI to zombies*, and Weaver and Veitas's 'Open-ended intelligence'." *Journal of Evolution and Technology* 25:55–87.

Grace, Katja, Salvatier, John, Zhang, Baobao, and Evans, Owain. 2016. "2016 Expert survey on progress in AI." AI Impacts, aiimpacts.org/2016-expert-survey-on-progress-in-ai.

Grace, Katja, Stein-Perlman, Zach, Weinstein-Raun, Benjamin, and Salvatier, John. 2022. "2022 Expert Survey on Progress in AI." AI Impacts, https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/.

Greaves, Hilary. 2017. "Discounting for public policy: A survey." *Economics and Philosophy* 33:391–439.

—. 2024. "Concepts of existential catastrophe." *The Monist* 107:109–29.

Greaves, Hilary and MacAskill, William. 2021. "The case for strong longtermism." Global Priorities Institute Working Paper 5-2021, https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/.

Hubinger, Evan, van Merwijk, Chris, Mikulik, Vladimir, Skalse, Joar, and Garrabrant, Scott. 2019. "Risks from learned optimization in advanced machine learning systems." ArXiv 1906.01820.

Krakovna, Victoria and Kramar, Janos. 2023. "Power-seeking can be probable and predictive for trained agents." arXiv 2304.06528, https://arxiv.org/abs/2304.06528.

Landgrebe, Jobst and Smith, Barry. 2022. *Why machines will never rule the world: Artificial intelligence without fear*. Routledge.

Loosemore, Richard. 2014. "The maverick nanny with a dopamine drip: Debunking fallacies in the theory of AI motivation." *Proceedings of the AAAI spring symposium* 28:31–36.

Metz, Cade. 2023. "'The godfather of AI' leaves Google and warns of danger ahead." *New York Times*, https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html.

Müller, Vincent and Bostrom, Nick. 2016. "Future progress in artificial intelligence: A survey of expert opinion." In Vincent Müller (ed.), *Fundamental issues of artificial intelligence*, 555–72. Springer.

Müller, Vincent and Cannon, Michael. 2021. "Existential risk from AI and orthogonality: Can we have it both ways?" *Ratio* 35:25–36.

Narveson, Jan. 1973. "Moral problems of population." *The Monist* 57:62–86.

Ngo, Richard and Bales, Adam. forthcoming. "Deceit and power: Machine learning and misalignment." In Hilary Greaves, David Thorstad, and Jacob Barrett (eds.), *Essays on longtermism*, forthcoming. Oxford University Press.

Nozick, Robert. 1974. *Anarchy, state, and utopia*. Basic books.

Omohundro, Stephen. 2008. "The basic AI drives." In Pei Wang, Ben Goertzel, and Stan Franklin (eds.), *Proceedings of the 2008 conference on artificial intelligence*, 483–92. IOS Press.

Parfit, Derek. 1984. *Reasons and persons*. Oxford University Press.

—. 1997. "Equality and priority." *Ratio* 10:202–21.

Rawls, John. 1971. *A theory of justice*. Belknap press.

Roemer, John. 1996. *Theories of distributive justice*. Harvard University Press.

Russell, Stewart. 2019. *Human compatible: Artificial intelligence and the problem of control*. Viking.

Scheffler, Samuel. 1994. *The rejection of consequentialism*. Oxford University Press.

Schramowski, Patrick, Turan, Cigdem, Andersen, Nico, Rothkpof, Constantin, and Kersting, Kristian. 2022. "Large pre-trained language models contain human-like biases of what is right and wrong to do." *Nature Machine Intelligence* 4:258–68.

Shields, Liam. 2016. *Just enough: Sufficiency as a demand of justice*. Edinburgh University Press.

Shteingart, Hanan and Loewenstein, Yonatan. 2014. "Reinforcement learning and human behavior." *Current Opinion in Neurobiology* 25:93–8.

Singer, Peter. 1975. *Animal liberation: A new ethics for our treatment of animals*. HarperCollins.

Stigler, George. 1950. "The development of utility theory, II." *Journal of Political Economy* 58:373–96.

Sutton, Richard and Barto, Andrew. 2018. *Reinforcement learning: An introduction*. MIT Press, second edition.

Thorstad, David. 2023. "Exaggerating the risks, Part 8: Carlsmith wrap-up." Reflective Altruism, https://reflectivealtruism.com/2023/06/03/exaggerating-the-risks-part-8-carlsmith-wrap-up/.

—. forthcoming. "Against the singularity hypothesis." *Philosophical Studies* forthcoming.

Turner, Alexander Matt, Smith, Logan, Shah, Rohin, Critch, Andrew, and Tadepalli, Prasad. 2021. "Optimal policies tend to seek power." *Proceedings of the 35th International Conference on Neural Information Processing Systems* 1766:23063–23074.

Turner, Alexander Matt and Tadepalli, Prasad. 2022. "Parametrically retargetable decision-makers tend to seek power." *Proceedings of the 36th International Conference on Neural Information Processing Systems* 2276:31391–31401.

von Oswald, Johannes et al. 2023. "Uncovering mesa-optimization algorithms in transformers." ArXiv 2309.05858, https://arxiv.org/abs/2309.05858.

Zhang, Baobao, Dreksler, Noemi, Anderljung, Markus, Kahn, Lauren, Giattino, Charlie, Dafoe, Allan, and Horowitz, Michael C. 2022. "Forescasting AI progress: Evidence from a survey of machine learning researchers." https://doi.org/10.48550/arXiv.2206.04132.

Zuber, Stéphane et al. 2021. "What should we agree on about the repugnant conclusion?" *Utilitas* 33:379–83.