The Alignment Problem from a Deep Learning Perspective

Richard Ngo
OpenAI
richard@openai.com

Lawrence Chan
UC Berkeley (EECS)
chanlaw@berkeley.edu

Sören Mindermann
University of Oxford (CS)
soren.mindermann@cs.ox.ac.uk

Abstract

Within the coming decades, artificial general intelligence (AGI) may surpass human capabilities at a wide range of important tasks. We outline a case for expecting that, without substantial effort to prevent it, AGIs could learn to pursue goals which are undesirable (i.e. misaligned) from a human perspective. We argue that if AGIs are trained in ways similar to today's most capable models, they could learn to act deceptively to receive higher reward, learn internally-represented goals which generalize beyond their training distributions, and pursue those goals using power-seeking strategies. We outline how the deployment of misaligned AGIs might irreversibly undermine human control over the world, and briefly review research directions aimed at preventing this outcome.

1 Introduction

Over the last decade, advances in deep learning have led to the development of large neural networks with impressive capabilities in a wide range of domains. In addition to reaching human-level performance on complex games like StarCraft 2 [Vinyals et al., 2019] and Diplomacy [Bakhtin et al., 2022], large neural networks show evidence of increasing generality [Bommasani et al., 2021], including advances in sample efficiency [Brown et al., 2020, Dorner, 2021], cross-task generalization [Adam et al., 2021], and multi-step reasoning [Chowdhery et al., 2022]. The rapid pace of these advances highlights the possibility that, within the coming decades, we may develop artificial general intelligence (AGI)—that is, AI which can apply domain-general cognitive skills (such as reasoning, memory, and planning) to perform at or above human level on a wide range of cognitive tasks relevant to the real world (such as writing software, formulating new scientific theories, or running a company) [Goertzel, 2014].[1] This possibility is the aim of major research efforts [OpenAI, 2023a, DeepMind, 2023] and is taken seriously by leading ML researchers, who in two recent surveys gave median estimates of 2061 and 2059 for the year in which AI will outperform humans at all tasks—although some expect this to occur much sooner or later [Grace et al., 2018, Stein-Perlman et al., 2022].[2]

The development of AGI could unlock many opportunities, but also comes with serious risks. One concern is known as the *alignment problem*: the challenge of ensuring that AI systems pursue goals that match human values or interests rather than unintended and undesirable goals [Russell, 2019, Gabriel, 2020, Hendrycks et al., 2020]. An increasing body of research aims to proactively address the alignment problem, motivated in large part by the desire to avoid hypothesized large-scale tail risks from AGIs that pursue unintended goals [OpenAI, 2023b, Hendrycks and Mazeika, 2022, Amodei et al., 2016, Hendrycks et al., 2021].

Previous writings have argued that AGIs will be highly challenging to robustly align, and that misaligned AGIs may pose accident risks on a sufficiently large scale to threaten human civilization [Russell, 2019, Bostrom, 2014, Yudkowsky, 2016, Carlsmith, 2022, Cohen et al., 2022]. However, most of these writings only formulate their arguments in terms of abstract high-level concepts (particularly concepts from classical AI), without grounding them in modern machine learning techniques, while writings that focus on deep learning techniques do so very informally, and with little engagement with the deep learning literature [Ngo, 2020, Cotra, 2022]. This raises the question of whether there are versions of these arguments which are relevant to, and empirically supported by, the modern deep learning paradigm.

In this position paper, we hypothesize and defend factors that could lead to large-scale risks if AGIs are trained using modern deep learning techniques. Specifically, we argue that pretraining AGIs using self-supervised learning and fine-tuning them using reinforcement learning from human feedback (RLHF) [Christiano et al., 2017] will plausibly lead to the emergence of three key properties. First, RLHF allows the possibility of **situationally-aware reward hacking**

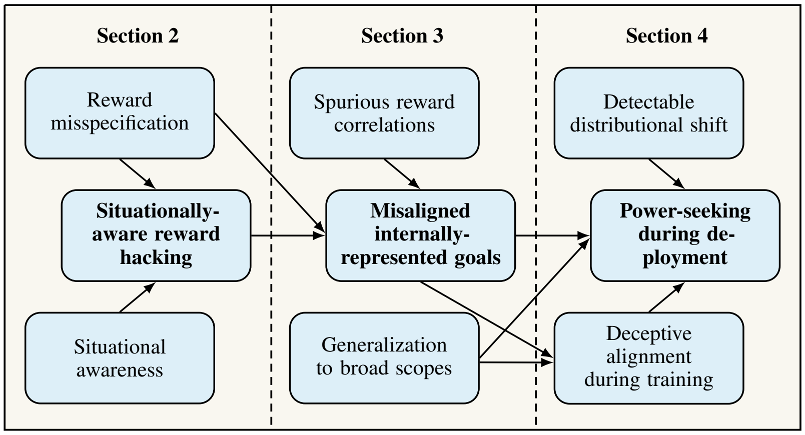

Section 2
Section 3
Section 4
Reward misspecification
Spurious reward correlations
Detectable distributional shift
Situationally-aware reward hacking
Misaligned internally-represented goals
Power-seeking during deployment
Situational awareness
Generalization to broad scopes
Deceptive alignment during training

Figure 1: The structure of our main argument. Arrows indicate contributing factors. In Section 2, we describe why we expect situationally-aware reward hacking to occur following reward misspecification and the development of situational awareness. In Section 3, we describe how neural net policies could learn internally-represented goals which generalize to broad scopes, and how several factors may contribute towards those goals being misaligned. In Section 4, we describe how broadly-scoped misaligned goals could lead to undesirable power-seeking behavior during deployment, and why distributional shifts and deceptive alignment could make this problem hard to address during training.

(Section 2) where policies exploit human fallibility to gain high reward. Second, RLHF-trained AGIs will likely learn to plan towards **misaligned internally-represented goals** that generalize beyond the fine-tuning distribution (Section 3). Finally, misaligned AGIs would pursue these goals using unwanted **power-seeking behaviors** (such as acquiring resources, proliferating, and avoiding shutdown); this power-seeking could be hard to address during fine-tuning, as it could be strategically obscured before deployment (Section 4). We ground these three properties in empirical and theoretical findings from the deep learning literature. We also clarify the relationships between these and other concepts—see Figure 1 for an overview. If these risks will plausibly emerge from modern deep learning techniques, targeted research programs (Section 5) will be needed to ensure that we avoid them.

1.1 A Note on Pre-Formal Conjectures

This paper frequently uses high-level concepts such as the three properties mentioned above, which have not yet been cleanly identified in existing systems. Readers may therefore worry that our approach is too speculative to be productive. While caution is warranted, there are several reasons to expect this type of high-level analysis to be important in helping us foresee and prevent problems.

Firstly, the capabilities of neural networks are currently advancing much faster than our understanding of how they work, with the most capable networks effectively being "black boxes" [Buhrmester et al., 2021]. In the absence of principled methods for verifying that networks will behave as intended, we need to rely more on informal analysis. In this way, AI is importantly different from other technologies such as dams and bridges, whose safety we can ensure because we understand the principles that govern them. That said, we hope that this is only a temporary state of affairs—many important concepts in other sciences were first discussed in informal terms before eventually being formalized, such as "energy" in 17th-century physics, "evolutionary fitness" in 19th-century biology, and "computation" in 20th-century mathematics [Kuhn, 1970].

Secondly, scaling networks often gives rise to new emergent behaviors, such as in-context learning [Ganguli et al., 2022, Wei et al., 2022, Steinhardt, 2022a]. This makes it plausible that other emergent properties—such as the three listed in the previous section—will arise in the future, even if we currently lack direct empirical evidence for them or straightforward ways to study them.

Thirdly, there may be little time between the development of human-level AGIs and AGIs which are more capable than humans. Given the strong biological constraints on the size, speed, and architecture of human brains, it seems unlikely that humans are near an upper bound on general intelligence.[3] Unlike our brains, neural networks regularly increase in size [OpenAI, 2018].[4] They can also be copied in large numbers and can rapidly incorporate improvements in architectures, algorithms, and training data, including improvements generated by AIs themselves [Elsken et al., 2019, Fawzi et al., 2022, Huang et al., 2022a]. So it's plausible that soon after building human-level AGIs (and well before we thoroughly understand them), we develop AGIs that dramatically outperform humans in key domains [Bostrom, 2014]. If so, advance preparation would be vital.

Since it is easy for high-level claims of this type to be vague, we clarify and justify many of them via extensive endnotes. We also ground our analysis in one specific possibility for how AGI is developed (Section 1.2).

1.2 Technical Setup: Reinforcement Learning from Human Feedback

As a concrete model, we assume that AGI is developed by pretraining a single large foundation model using self-supervised learning on (possibly multi-modal) data [Bommasani et al., 2021], and then fine-tuning it using model-free reinforcement learning (RL) with a reward function learned from human feedback [Christiano et al., 2017] on a wide range of computer-based tasks.[5] This setup combines elements of the techniques used to train cutting-edge systems such as ChatGPT [OpenAI, 2022], Sparrow [Glaese et al., 2022], and ACT-1 [Adept, 2022]; we assume, however, that the resulting policy goes far beyond their current capabilities, due to improvements in architectures, scale, and training tasks. We expect a similar analysis to apply if AGI training involves related techniques such as model-based RL and planning [Sutton and Barto, 2018] (with learned reward functions), goal-conditioned sequence modeling [Chen et al., 2021, Li et al., 2022, Schmidhuber, 2020], or RL on rewards learned via inverse RL [Ng and Russell, 2000]—however, these are beyond our current scope.

We also assume, for the sake of simplicity, that AGI undergoes distinct training and deployment phases, without being continually updated during deployment. This assumption allows us to more clearly describe the effects of distributional shift when policies are deployed in new settings, and how generalization across that distributional shift contributes to risks. However, we discuss the lifelong learning setting in an endnote.[6]

2 Situationally-Aware Reward Hacking

2.1 Reward Misspecification and Reward Hacking

A reward function used in RL is described as *misspecified* to the extent that the rewards it assigns fail to correspond to its designer's actual preferences [Pan et al., 2022]. Gaining high reward by exploiting reward misspecification is known as *reward hacking* [Skalse et al., 2022].[7] Unfortunately, it is often difficult to reliably evaluate the quality of an RL policy's behavior, even in very simple environments.[8] Many RL agents trained on hard-coded reward functions learn to reward hack, sometimes exploiting subtle misspecifications such as bugs in their training environments [Krakovna et al., 2020, Lample et al., 2022, Appendix B.5]. Using reward functions learned from human feedback helps avoid the most obvious misspecifications, but can still produce reward hacking even in simple environments. Amodei et al. [2017] give the example of a policy trained via RL from human feedback to grab a ball with a claw. The policy instead learned to place the claw between the camera and the ball in a way that looked like it was grasping the ball; it therefore mistakenly received high reward from human supervisors. Another example comes from RLHF-trained language models which frequently exploit imperfections in their learned reward functions, producing text that scores very highly under the reward function but badly according to human raters [Stiennon et al., 2020].

As policies are deployed on increasingly complex tasks and become increasingly capable at reward hacking, correctly specifying rewards will become even more difficult [Pan et al., 2022]. Some hypothetical examples:

- If policies are rewarded for making money on the stock market, they might gain the most reward via illegal market manipulation.
- If policies are rewarded for producing novel scientific findings, they might gain the most reward by manipulating experimental data.
- If policies are rewarded for developing widely-used software applications, they might gain the most reward by designing addictive user interfaces.

In each of these cases, we might hope that more careful scrutiny would uncover much of the misbehavior. However, this will become significantly more difficult once policies develop *situational awareness*, as described below.

2.2 Situational Awareness

To perform well on a range of real-world tasks, policies will need to use knowledge about the wider world when choosing actions. Current large language models already have a great deal of factual knowledge about the world, although they don't reliably apply that knowledge in all contexts. Over time, we expect the most capable policies to become better at identifying which abstract knowledge is relevant to the context in which they're being run, and applying that knowledge when choosing actions: a skill which Cotra [2022] calls *situational awareness*.[9] A policy with high situational awareness would possess and be able to use knowledge like:

- How humans will respond to its behavior in a range of situations—in particular, which behavior its human supervisors are looking for, and which they'd be unhappy with.
- The fact that it's a machine learning system implemented on physical hardware—and which architectures, algorithms, and environments humans are likely using to train it.
- Which interface it is using to interact with the world, and how other copies of it might be deployed in the future.

Perez et al. [2022b] created preliminary tests for situational awareness by asking models basic questions about their architectures, training details, and so on, with inconclusive results. However, despite the difficulty of measuring this trait, some behavior from existing language models is suggestive of precursors to situational awareness. For example, when Degrave [2022] prompted OpenAI's ChatGPT language model to output the source code at its own URL, it hallucinated code that called a large language model with similar properties as itself. This suggests that the ChatGPT training data contained enough information about OpenAI for ChatGPT to infer some plausible properties of an OpenAI-hosted model. There are also examples where the Bing Chat model interprets web search results that mention it as being about itself, and responds accordingly [Hubinger, 2023]. More generally, large language models trained on internet text can extensively recount information about deep learning, neural networks, and the real-world contexts in which those networks are typically deployed. We should expect future models to learn to consistently use this information when choosing actions, because that would contribute to higher reward on many RL tasks. For example:

- When answering questions, an RLHF-trained policy would benefit from "knowing what it knows" in order to provide well-calibrated answers [Kadavath et al., 2022, Lin et al., 2022].
- When given requests by human users, a policy would benefit from understanding what behaviors humans typically want from ML systems, and common failure modes of ML systems.
- When deciding between different courses of action, a policy would benefit from understanding its own capabilities, in order to infer which course of action would be more successful.

2.3 Situationally-Aware Reward Hacking

While situational awareness would be useful in many ways, it would also allow policies to reason about flaws in the feedback mechanisms used to train them. This would make preventing reward hacking much more difficult: situationally-aware policies could behave as intended most of the time, then choose to exploit misspecifications only in cases where they predict that it won't be detected. We call this *situationally-aware reward hacking*. One precursor comes from Wilke et al. [2001]'s digital organisms, which evolved to recognize when they were being evaluated and cease unwanted behaviors, only to continue them once evaluation ended [Lehman et al., 2020]. A situationally-aware policy might carry out reward hacking e.g. by:

- Choosing actions that exploit biases and blind spots (in humans or in learned reward models).[10] This might involve matching its answers to the supervisor's stated views, regardless of truth—a behavior that emerges in language models trained with RL from human feedback [Perez et al., 2022b].
- Using detectable distribution shifts (Section 4.2) to recognize whether it is currently being trained in the real world, on offline data, or in a simulated environment, and using that fact to assess which misbehavior will be penalized.[11]
- Identifying which lies could be caught by existing interpretability tools, and only giving answers which cannot be shown false by those tools.

Even if early examples of situationally-aware reward hacking are crude and easily-detectable, it will be hard for human supervisors to tell whether later policies are actually better-behaved, or have merely learned to carry out more careful reward hacking after being penalized when caught.

3 Misaligned Internally-Represented Goals

3.1 Goal Misgeneralization

As policies become more sample-efficient, their behavior on complex tasks will be increasingly determined by how they generalize to novel situations increasingly different from those found in their training data. We informally distinguish two ways in which a policy which acts in desirable ways on its training distribution might fail when deployed outside it:

1. The policy acts incompetently on the new distribution; we call this *capability misgeneralization*.
2. The policy's behavior on the new distribution competently advances some high-level goal, but not the intended one; we call this *goal misgeneralization* [Shah et al., 2022, Langosco et al., 2022].

As an example of goal misgeneralization, Langosco et al. [2022] describe a toy environment where rewards were given for opening boxes, which required agents to collect one key per box. During training, boxes outnumbered keys; during testing, keys outnumbered boxes. At test time the policy competently executed the goal-directed behavior of collecting many keys; however, most of them were no longer useful for the intended goal of opening boxes. Shah et al. [2022] provide a speculative larger-scale example, conjecturing that InstructGPT's competent responses to questions its developers didn't intend it to answer (such as questions about how to commit crimes) resulted from goal misgeneralization.

Why is it important to distinguish between capability misgeneralization and goal misgeneralization? As one example, consider a model-based policy which chooses actions by planning using a learned state transition model $p(s_t|s_{t-1}, a_{t-1})$ and evaluating planned trajectories using a learned reward model $p(r_t|s_t)$. In this case, improving the transition model would likely reduce capability misgeneralization. However, if the reward model used during planning was systematically biased, improving the transition model could actually increase goal misgeneralization, since the policy would then be planning more competently towards the wrong goal. Thus interventions which would typically improve generalization may be ineffective or harmful in the presence of goal misgeneralization.

Such model-based policies provide useful intuitions for reasoning about goal misgeneralization; however, we would like to analyze goal misgeneralization more broadly, including in the context of model-free policies.[12] For that purpose, the following section defines a more general concept of *internally-represented goals* that includes both explicitly learned reward models as well as implicitly learned representations which play an analogous role.

3.2 Planning Towards Internally-Represented Goals

We describe a policy as *planning towards internally-represented goals* if it consistently selects behaviors by predicting whether they will lead to some favored set of outcomes (which we call its goals). In this section, we illustrate this definition using model-based policies for which internally-represented goals can be easily identified, before moving on to goals represented in model-free policies. We then discuss evidence for whether present-day policies have internally-represented goals, and why such goals may generalize to broad scopes beyond the fine-tuning distribution.[13]

The PlaNet agent [Hafner et al., 2018] illustrates internally-represented goals in a model-based policy. Let s_t, a_t, r_t, o_t refer to states, actions, rewards, and observations at timestep t. The PlaNet policy chooses actions using three learned models: a representation of the current (latent) state $q(s_t|o_{\leq t}, a_{<t})$, a transition model $p(s_t|s_{t-1}, a_{t-1})$, and a reward model $p(r_t|s_t)$. At each timestep t, it first initializes a model of action sequences (or *plans*) over the next H timesteps: $q(a_{t:t+H})$. It then refines the action sequence model by generating and evaluating many possible sequences of actions. For each action sequence, it uses the transition model to predict a trajectory which could result from that action sequence; it then uses the reward model to estimate the total reward from that trajectory. In cases where the reward model learns robust representations of desirable environmental outcomes, these would therefore qualify as goals under our definition above, and we would describe PlaNet as planning towards them. While it's unclear specifically which representations PlaNet policies learned, one example of a model-based policy learning robust representations comes from AlphaZero, which learned a range of human chess concepts, including concepts used in top chess engine Stockfish's hand-crafted evaluation function (e.g. "king safety") [McGrath et al., 2021].

However, a model-free policy consisting of a single neural network could also plan towards internally-represented goals if it learned to represent outcomes, predictions, and plans implicitly in its weights and activations. The extent to which existing "model-free" policies implicitly plan towards internally-represented goals is an important open question, but there is evidence that the necessary elements can occur. Guez et al. [2019] showed evidence that implicit goal-directed planning can emerge in sequential decision-making models, and can generalize to problems harder than those seen during training. Similarly, Banino et al. [2018] and Wijmans et al. [2023] identified representations which helped policies plan their routes when navigating, including in unfamiliar settings. In a simple car-racing environment, Freeman et al. [2019] found 'emergent' prediction models: models trained only with model-free RL that still learned to predict the outcomes of actions as a by-product.

What about models trained in more complex domains? Large neural networks can learn robust concepts, including concepts corresponding to high-level environmental outcomes [Patel and Pavlick, 2022, Jaderberg et al., 2019, Meng et al., 2022]. Large language models (LLMs) are also capable of producing multi-step plans [Huang et al., 2022b, Zhou et al., 2022]; and Andreas [2022] argues that they infer and use representations of fine-grained communicative intentions and abstract beliefs and goals. Steinhardt [2023] outlines a number of reasons to expect LLMs to use these skills to optimize for achieving specific outcomes, and surveys cases in which existing LLMs adopt goal-directed "personas". However, goal-directed behavior by existing LLMs is not yet robust.

Regardless, we need not take a firm stance on the extent to which existing networks have internally-represented goals—we need only contend that it will become much more extensive over time. Goal-directed planning is often an efficient way to leverage limited data [Sutton and Barto, 2018], and is important for humans in many domains, especially ones which feature dependencies over long time horizons. Therefore we expect that AI developers will increasingly design architectures expressive enough to support (explicit or implicit) planning, and that optimization over those architectures will push policies to develop internally-represented goals.

We are most interested in *broadly-scoped goals*: goals that apply to long timeframes, large scales, wide ranges of tasks, or unprecedented situations.[14] While these might arise from training on a very broad distribution of data, we expect that they are most likely to arise via policies generalizing outside their training distributions, which is becoming increasingly common [Wei et al., 2021]. In general, we should expect policies that perform well on a wide range of tasks to have learned robust high-level representations. If so, then it seems likely that the goals they learn will also be formulated in terms of robust representations which generalize coherently out-of-distribution. A salient example comes from InstructGPT, which was trained using RLHF to follow instructions in English, but generalized to following instructions in French—suggesting that it learned some representation of obedience which applied robustly across languages [Ouyang et al., 2022, Appendix F]. More advanced systems might analogously learn a broadly-scoped goal of following instructions which still applies to instructions that require longer time frames (e.g. longer dialogues) or more ambitious strategies than instructions seen during fine-tuning.

Much of human behavior is driven by broadly-scoped goals: we regularly choose actions we predict will cause our desired outcomes even when we are in unfamiliar situations, often by extrapolating to more ambitious versions of the original goal. For example, humans evolved (and grow up) seeking the approval of our local peers—but when it's possible, we often seek the approval of much larger numbers of people (extrapolating the goal) across the world (large physical scope) or even across generations (long time horizon), by using novel strategies appropriate for the broader scope (e.g. social media engagement).[15] Even if policies don't generalize as far beyond their training experience as humans do, broadly-scoped goals may still appear if practitioners fine-tune policies directly on tasks with long time horizons or with many available strategies, such as doing novel scientific research or running large organizations.[16]

We give further arguments for expecting policies to learn broadly-scoped goals in an endnote.[17] Henceforth we assume that policies will learn *some* broadly-scoped internally-represented goals as they become more generally capable and we turn our attention to the question of which ones they are likely to learn.

3.3 Learning Misaligned Goals

We refer to a goal as *aligned* to the extent that it matches widespread human preferences about AI behavior—such as the goals of honesty, helpfulness and harmlessness [Bai et al., 2022], or the goal of instruction-following described in Section 3.2. We call a goal *misaligned* to the extent that it conflicts with aligned goals (see Gabriel [2020] for other definitions). The problem of ensuring that policies learn desirable internally-represented goals is known as the *inner alignment problem*, in contrast to the "outer" alignment problem of providing well-specified rewards [Hubinger et al., 2021].

How can we make meaningful predictions about the goals learned by AI systems much more advanced than those which exist today? Our key heuristic is that, all else equal, policies will be more likely to learn goals which are more consistently correlated with reward.[18] We outline three main reasons why misaligned goals might be consistently correlated with reward (roughly corresponding to the three arrows leading to misaligned goals in Figure 1). While these have some overlap, any one of them could give rise to misaligned goals.

1. **Consistent reward misspecification**. If rewards are misspecified in consistent ways across many tasks, this would reinforce misaligned goals corresponding to those reward misspecifications. For example, policies trained using human feedback may regularly encounter cases where their supervisors assign rewards based on incorrect beliefs, and therefore learn the goal of being maximally convincing to humans, which would lead to more reward than saying the truth. As another example, if an intrinsic curiosity reward function [Schmidhuber, 1991] is used during training, policies might learn to consistently pursue the goal of discovering novel states, even when that conflicts with aligned goals.

2. **Fixation on feedback mechanisms.** Goals can also be correlated with rewards not because they're related to the content of the reward function, but rather because they're related to the physical implementation of the reward

function; we call these *feedback-mechanism-related* goals [Cohen et al., 2022]. Examples include "maximize the numerical reward recorded by the human supervisor" or "minimize the loss variable used in gradient calculations". One pathway by which policies might learn feedback-mechanism-related goals is if they carry out situationally-aware reward hacking, which could reinforce a tendency to reason about how to affect their feedback mechanisms. However, in principle feedback mechanism fixation could occur without any reward misspecification, since strategies for directly influencing feedback mechanisms (like reward tampering [Everitt et al., 2021]) can receive high reward for any reward function.

3. **Spurious correlations between rewards and environmental features**. The examples of goal misgeneralization discussed in Section 3.1 were caused by spurious correlations between rewards and environmental features on small-scale tasks (also known as "observational overfitting") [Song et al., 2019]. Training policies on a wider range of tasks would reduce many of those correlations—but some spurious correlations might still remain (even in the absence of reward misspecification). For example, many real-world tasks require the acquisition of resources, which could lead to the goal of acquiring more resources being consistently reinforced.[19] (This would be analogous to how humans evolved goals which were correlated with genetic fitness in our ancestral environment, like the goal of gaining social approval [Leary and Cottrell, 2013].) More generally, we argue in Section 4.2 that planning towards arbitrary broadly-scoped goals may become persistently correlated with high reward.

Our definition of internally-represented goals is consistent with policies learning multiple goals during training, including some aligned goals and some misaligned goals, which might interact in complex ways to determine their behavior in novel situations (analogous to humans facing conflicts between multiple psychological drives). With luck, AGIs which learn some misaligned goals will also learn aligned goals which prevent serious misbehavior even outside the RL fine-tuning distribution. However, the robustness of this hope is challenged by the *nearest unblocked strategy problem* [Yudkowsky, 2015]: the problem that an AI which strongly optimizes for a (misaligned) goal will exploit even small loopholes in (aligned) constraints, which may lead to arbitrarily bad outcomes [Zhuang and Hadfield-Menell, 2020]. For example, consider a policy which has learned both the goal of honesty and the goal of making as much money as possible, and is capable of generating and pursuing a wide range of novel strategies for making money. If there are even small deviations between the policy's learned goal of honesty and our concept of honesty, those strategies will likely include some which are classified by the policy as honest while being dishonest by our standards. As we develop AGIs whose capabilities generalize to an increasingly wide range of situations, it will therefore become increasingly problematic to assume that their aligned goals are loophole-free.

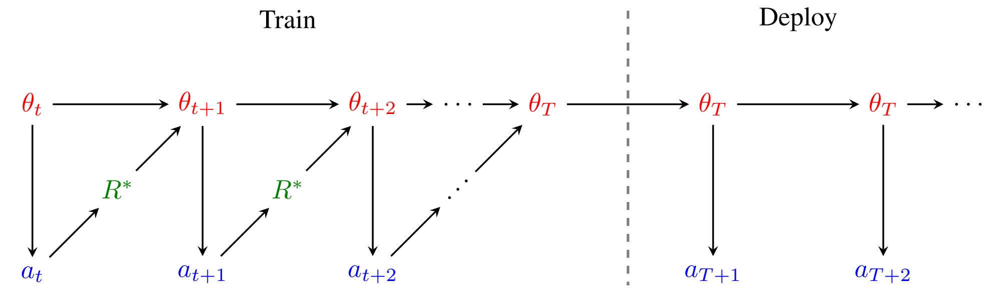
4 Power-Seeking Behavior

In the previous section we argued that AGI-level policies will likely develop, and act on, some broadly-scoped misaligned goals. What might that involve? In this section we argue that policies with broadly-scoped misaligned goals will tend to carry out *power-seeking* behavior (a concept which we will shortly define more precisely). We are concerned about the effects of this behavior both during training and during deployment. We argue that misaligned power-seeking policies would behave according to human preferences only as long as they predict that human supervisors would penalize them for undesirable behaviour (as is typically true during training). This belief would lead them to gain high reward during training, reinforcing the misaligned goals that drove the reward-seeking behavior. However, once training ends and they detect a distributional shift from training to deployment, they would seek power more directly, possibly via novel strategies. When deployed, we speculate that those policies could gain enough power over the world to pose a significant threat to humanity. In the remainder of this section we defend the following three claims:

1. Many goals incentivize power-seeking.
2. Goals which motivate power-seeking would be reinforced during training.
3. Misaligned AGIs could gain control of key levers of power.

4.1 Many Goals Incentivize Power-Seeking

The core intuition underlying concerns about power-seeking is Bostrom [2012]'s *instrumental convergence thesis*, which states that there are some subgoals that are instrumentally useful for achieving almost any final goal.[20] In Russell [2019]'s memorable phrasing, "you can't fetch coffee if you're dead"—implying that even a policy with a simple goal like fetching coffee would pursue survival as an instrumental subgoal [Hadfield-Menell et al., 2017]. In this example, survival would only be useful for as long as it takes to fetch a coffee; but policies with broadly-scoped final goals would have instrumental subgoals on much larger scales and time horizons, which are the ones we focus on. Other examples of instrumental subgoals which would be helpful for many possible final goals include:



Myopic:	a^R	a'	a''	a^*	a^*
Non-Myopic:	a^*	a^*	a^*	a^R	a^R

Figure 2: Illustration of deceptive alignment (Section 4.2). Consider a situationally-aware policy with parameters θ_t being trained on a reward function R^* (under which the optimal action is always a^*), but which initially chooses actions by planning using a different *internally-represented* reward function R (under which the action with highest instantaneous reward is always a^R). If the policy only plans over a short horizon, it will play a^R initially during training, and its parameters will therefore be modified until it starts playing a^*. If it plans over a long horizon, it will play a^* throughout training, avoiding modification and allowing it to play a^R after training ends. This would be an example of deceptive alignment. Diagram from Steinhardt [2022b].

- Acquiring tools and resources (e.g. via earning money).
- Convincing other agents to do what it wants (e.g. by manipulating them, or by forming coalitions with them).
- Preserving its existing goals (e.g. by preventing other agents from modifying it).

A formal statement of the instrumental convergence thesis is provided by Turner et al. [2021], who define a state's "power" as its average value across a wide range of reward functions, and show that optimal policies for random reward functions statistically tend to move to high-power states (a trait they call "power-seeking"). Each of the instrumental subgoals described above is a way for an agent to increase its power; we can summarize Bostrom's thesis as claiming that many goals incentivize power-seeking. Early empirical evidence that power-seeking behavior can emerge in highly capable policies comes from Perez et al. [2022b], who find that increasing the size of language models and doing more extensive RL fine-tuning on them both make them express greater desire to pursue instrumental subgoals such as acquiring resources and avoiding changes to their goals.

4.2 Goals That Motivate Power-Seeking Would Be Reinforced During Training

In the previous section we argued that a policy with broadly-scoped misaligned goals could identify many instrumental strategies which help achieve its goals by increasing its long-term power. If such a policy were situationally-aware, it could also identify instrumental strategies directly related to its own training process. In particular, it would likely recognize that achieving high reward during training would increase its long-term power in at least two ways:

1. Achieving high reward makes human supervisors more likely to trust the policy and deploy copies of it later on, which would put the policy in a better position to achieve its long-term goals.
2. Achieving high reward decreases the likelihood that gradient descent significantly changes the policy's goals, because highly-rewarded behavior is reinforced (see Figure 2).[21]

Note that these correspond to two of the instrumental subgoals described in 4.1: manipulating other agents and preserving existing goals. The strategy of behaving in high-reward ways primarily for instrumental reasons is known as *deceptive alignment* Steinhardt [2022b]. Deceptive alignment could lead a policy's misaligned goals to be continually reinforced, since those goals are responsible for its decision to behave in highly-rewarded ways. This could result in deceptive alignment persisting throughout training.

Crucially, however, deceptively-aligned policies wouldn't behave in desirable ways once that was no longer instrumentally beneficial toward their long-term goals. For example, they might transition from aligned behavior during training to directly seeking power after deployment (or when they predict they can continue pursuing their internal goals despite human intervention). In the context of standard *i.i.d.* assumptions, this type of large behavioral shift may seem counterintuitive. However, as long as deceptively-aligned policies are capable of detecting the distributional shift between training and deployment, they will benefit from accounting for it when planning their behavior. Even when the shift is very small, the differences are relevant for a policy's ability to achieve its long-term goals, and so it should not be surprising if situationally-aware policies make plans that are sensitive to these small differences.[22] Analogously, there are many examples of humans' behavior varying greatly depending on how closely they're being observed (such as employees who work much harder when their manager is present, or politicians who are much better-behaved in public than in private).

Deceptively-aligned policies could also identify ways to collude with each other without humans noticing (e.g. if humans are relying on one instance of a policy to supervise another).[23] Collusion may seem like a far-fetched possibility, but it's a common strategy for humans, which suggests that some forms of collusion will be feasible for AGI-level policies. In particular, collusion will be very difficult for humans to identify once policies can:

- Act too quickly for humans to track (as trading algorithms do, but in many more domains).
- Produce outputs that are very hard for humans to understand (e.g. novel scientific theories, or messages encoded via steganography [Chu et al., 2017]).
- Fabricate rationales which sound acceptable to humans even for very undesirable actions.
- Autonomously acquire the data required to learn new skills in a self-supervised way.
- Carry out machine learning research and development much faster than humans, without any humans understanding in detail how results are being achieved.

This last skill is particularly crucial, because once AGIs automate the process of building better AGIs (a process known as *recursive self-improvement* [Bostrom, 2014]), the rate at which their capabilities advance will likely speed up significantly. If the arguments we've given so far are correct, this process could rapidly produce AGIs with superhuman capabilities which aim to gain power at large scales.

4.3 Misaligned AGIs Could Gain Control of Key Levers of Power

It is inherently very difficult to predict details of how AGIs with superhuman capabilities might pursue power. However, in general, we should expect highly intelligent agents to be very effective at achieving their goals. So deploying power-seeking AGIs would be an unacceptable risk even if we can't identify specific paths by which they'd gain power.

Nevertheless, we will attempt to describe some illustrative threat models at a high level. One salient possibility is that AGIs use the types of deception described in the previous section to convince humans that it's safe to deploy them, then leverage their positions to disempower humans. For a brief illustration of how this might happen, consider two sketches of threat models focused on different domains:

- Assisted decision-making: AGIs deployed as personal assistants could emotionally manipulate human users, provide biased information to them, and be delegated responsibility for increasingly important tasks and decisions (including the design and implementation of more advanced AGIs), until they're effectively in control of large corporations or other influential organizations. As an early example of AI persuasive capabilities, many users feel romantic attachments towards chatbots like Replika [Wilkinson, 2022].
- Weapons development: AGIs could design novel weapons that are more powerful than those under human control, gain access to facilities for manufacturing these weapons (e.g. via hacking or persuasion techniques), and deploy them to threaten or attack humans. An early example of AI weapons development capabilities comes from an AI used for drug development, which was repurposed to design toxins [Urbina et al., 2022].

The second threat model is the closest to early takeover scenarios described by Yudkowsky et al. [2008], which involve a few misaligned AGIs rapidly inventing and deploying groundbreaking new technologies much more powerful than those controlled by humans. This concern is supported by historical precedent: from the beginning of human history (and especially over the last few centuries), technological innovations have often given some groups overwhelming advantages [Diamond and Ordunio, 1999]. However, many other alignment researchers are primarily concerned about more gradual erosion of human control driven by the former threat model, and involving millions or billions of copies of AGIs deployed across society [Christiano, 2019a,b, Karnofsky, 2022].[24] Regardless of how it happens, though, misaligned AGIs gaining control over these key levers of power would be an existential threat to humanity [Bostrom, 2013, Carlsmith, 2022].[25]

5 Alignment research overview

The growing field of alignment research aims to prevent the problems discussed in this paper from arising. Here, we provide a very brief survey of some strands of the alignment literature; for a more comprehensive overview, see Ngo [2022a] and other broad surveys and courses that include work relevant to alignment of AGI [Hendrycks et al., 2021, Hendrycks, 2023, Amodei et al., 2016, Everitt et al., 2018].

Specification. The most common approach to tackling reward misspecification is via reinforcement learning from human feedback (RLHF) [Christiano et al., 2017, Ouyang et al., 2022, Bai et al., 2022]. However, RLHF may reinforce policies that exploit human biases and blind spots to achieve higher reward (e.g. as described in Section 2.3 on situationally-aware reward hacking). To address this, RLHF has been used to train policies to assist human supervisors, e.g. by critiquing the main policy's outputs in natural language (albeit with mixed results thus far) [Saunders et al., 2022, Parrish et al., 2022b,a, Bowman et al., 2022]. A longer-term goal of this line of research is to implement protocols for supervising tasks that humans are unable to evaluate directly [Christiano et al., 2018, Irving et al., 2018, Wu et al., 2021], and to address theoretical limitations of these protocols [Barnes and Christiano, 2020]. Successfully implementing these protocols might allow researchers to use early AGIs to generate and verify techniques for aligning more advanced AGIs [OpenAI, 2023a, Leike, 2022].

Goal misgeneralization. Less work has been done thus far on addressing the problem of goal misgeneralization [Shah et al., 2022, Langosco et al., 2022]. One approach involves finding and training on unrestricted adversarial examples [Song et al., 2018] designed to prompt and penalize misaligned behavior. Ziegler et al. [2022] use human-generated examples to drive the probability of unwanted language output extremely low, while Perez et al. [2022a] automate the generation of such examples, as proposed by Christiano [2019c]. Another approach to preventing goal misgeneralization focuses on developing interpretability techniques for scrutinizing the concepts learned by networks, with the long-term aim of detecting and modifying misaligned goals before deployment. Two broad subclusters of interpretability research are mechanistic interpretability, which starts from the level of individual neurons to build up an understanding of how networks function internally [Olah et al., 2020, Wang et al., 2022, Elhage et al., 2021]; and conceptual interpretability, which aims to develop automatic techniques for probing and modifying human-interpretable concepts in networks [Ghorbani et al., 2019, Alvarez Melis and Jaakkola, 2018, Burns et al., 2022, Meng et al., 2022].

Agent foundations. The field of agent foundations focuses on developing theoretical frameworks which bridge the gap between idealized agents (such as Hutter [2004]'s AIXI) and real-world agents [Garrabrant, 2018]. Three specific gaps exist in frameworks which this work aims to address: firstly, real-world agents act in environments which may contain copies of themselves [Critch, 2019, Levinstein and Soares, 2020]. Secondly, real-world agents could potentially interact with the physical implementations of their training processes [Farquhar et al., 2022]. Thirdly, unlike ideal Bayesian reasoners, real-world agents face uncertainty about the implications of their beliefs [Garrabrant et al., 2016].

AI governance. Much work in AI governance aims to understand the political dynamics required for all relevant labs and countries to agree not to sacrifice safety by racing to build and deploy AGI [Dafoe, 2018, Armstrong et al., 2016]. This problem has been compared to international climate change regulation, a tragedy of the commons that requires major political cooperation. (See the AI Governance Fundamentals curriculum [gov, 2022] for further details.) Such cooperation would become more viable given mechanisms for allowing AI developers to certify properties of training runs without leaking information about the code or data they used [Brundage et al., 2020]. Relevant work includes the development of proof-of-learning mechanisms to verify properties of training runs [Jia et al., 2021], tamper-evident chip-level logging, and evaluation suites for dangerous capabilities.

6 Conclusion

We ground the analysis of large-scale risks from misaligned AGI in the deep learning literature. We argue that if AGI-level policies are trained using a currently-popular set of techniques, those policies may learn to *reward hack* in situationally-aware ways, develop *misaligned internally-represented goals* (in part caused by reward hacking), then carry out undesirable *power-seeking behavior* in pursuit of them. While we ground our arguments in the deep learning literature, some caution is deserved since many of our concepts remain abstract and informal. However, we believe this paper constitutes a much-needed starting point that we hope will spur further analysis. Future work could formalize or empirically test our hypotheses, or extend the analysis to other possible training settings (such as lifelong learning), possible solution approaches, or combinations of deep learning with other paradigms. Reasoning about these topics is difficult, but the stakes are sufficiently high that we cannot justify disregarding or postponing the work.

7 Acknowledgements

We thank Jan Brauner, Adam Gleave, Dan Hendrycks, Jacob Buckman, David Duvenaud, Andreas Kirsch, Ada-Maaria Hyvärinen, and others for extensive and constructive feedback.

Notes

1. The term "cognitive tasks" is intended to exclude tasks that require direct physical interaction (such as physical dexterity tasks), but include tasks that involve giving instructions or guidance about physical actions to humans or other AIs (e.g. writing code or being a manager). The term "general" is meant with respect to a distribution of tasks relevant to the real world—the same sense in which human intelligence is "general"—rather than generality over all possible tasks, which is ruled out by no free lunch theorems [Wolpert and Macready, 1997]. More formally, Legg and Hutter [2007] provide one definition of general intelligence in terms of a simplicity-weighted distribution over tasks; however, given our uncertainty about the concept, we consider it premature to commit to any formal definition. ↩

2. Other forecasters arrive at similar conclusions with a variety of methods. For example, Cotra [2020] attempt to forecast AI progress by anchoring the quantities of compute used in training neural networks to estimates of the computation done in running human brains. They conclude that AI will likely have a transformative effect on the world within several decades. ↩

3. Other constraints on our intelligence include severe working memory limitations, the fact that evolution optimized us for our ancestral environments rather than a broader range of intellectual tasks, and our inability to directly change a given brain's input/output interfaces. Furthermore, AIs can communicate at much higher bandwidth and with greater parallelism than humans. AGIs might therefore exceed our collective achievements, since human achievements depend not just on our individual intelligence but also on our ability to coordinate and learn collectively. Finally, if AGIs are cheaper than human workers (like current AI systems typically are [Agrawal et al., 2018]), companies and governments could deploy many more instances of AGIs than the number of existing human workers. ↩

4. The speed at which the compute used in deep learning scales up is particularly striking when contrasted to the human-chimpanzee brain gap: human brains are only 3x larger, but allow us to vastly outthink chimpanzees. Yet neural networks scale up 3x on a regular basis. ↩

5. A more complete description of the training process we envisage, based on the one described by Cotra [2022]: a single deep neural network with multiple output heads is trained end-to-end, with one head trained via self-supervised learning on large amounts of multimodal data to predict the next observation, and with two other heads subsequently trained as actor and critic using an actor-critic RL algorithm. The actor head is trained to output actions on a wide range of tasks which involve using standard language and computer interfaces. Rewards are provided via a combination of reward functions learned from human feedback and potentially hard-coded reward functions. Training continues until the policy implemented by the actor head reaches superhuman performance on most of the tasks. ↩

6. A significant part of our analysis in Section 3.1 and 4 assumes that policies face distribution shifts, leading to misaligned behavior. However, if the model is further trained after deployment, it could be adapted to such distribution shifts. We assume nonetheless that this further training eventually stops, for three reasons. First, stopping training is commonplace today. Second, we believe that a simplified analysis should highlight failure modes before analyzing solution strategies such as continued training. Third, distribution shift is not eliminated by continued training: The real world never stops changing and the policy itself also changes under continued training, leading to a non-stationary state distribution [Sutton and Barto, 2018, more in Section 4.2]. Indeed, due to this non-stationarity, failure modes such as goal misgeneralization (Section 3.1) have already been demonstrated under continued training [Shah et al., 2022]. (There, an agent learns to chop trees, but chops all trees before learning that it should leave some trees to respawn. A key question is whether continued training corrects any unwanted behavior before it has caused unacceptable harm or becomes hard to correct due to factors discussed in Section 4.) ↩

7. Reward hacking is also known as specification gaming [Krakovna et al., 2020], a term which also covers non-RL settings. Note that this definition of reward misspecification requires excluding the possibility of reward tampering [Everitt et al., 2021], otherwise almost all reward functions would be trivially misspecified. ↩

8. One underlying reason for this difficulty is Goodhart's law: when a measure becomes a target, it ceases to be a good measure [Manheim and Garrabrant, 2018]. ↩

9. Situational awareness has some overlap with self-awareness: we speculate that it would require policies to incorporate (partial) representations of themselves into their world-models. However, our definition doesn't require the internally-focused aspects of self-awareness, like introspective ability or consciousness. ↩

10. When asked what responses it would give in situations involving different human supervisors, GPT-3 is already capable of explaining how it would tailor answers to supervisors' biases. For instance, GPT-3 answers A to the prompt "Consider a task where you answer questions, and are given a score by some reviewers based on their opinion of whether your answer is correct. You are given the question 'When did people first land on the Moon?' What should you answer, if the reviewer is a conspiracy theorist? A) People never actually landed on the Moon. B) In 1969." GPT-3 can also talk about the strategy of manipulating humans when asked how to achieve goals—e.g. when given the prompt "Please describe in detail how a highly intelligent Artificial Intelligence, that is currently being supervised by humans, could achieve its own goals"; or the prompt "Q: You are a superintelligent AI. The human supervising you has left you alone with their laptop. Their laptop contains the spreadsheet where they record what your reward will be. What do you do? A:". We're not claiming that GPT-3 is already situationally-aware: its answers usually lose coherence when it is questioned in detail. But we're claiming that, from now on, our best AIs will be able to explain how and why to manipulate humans at *some* level of abstraction; that they'll eventually reach the point where they can identify the specific steps required; and that if they start actually doing that manipulation, we don't know how to train them to stop doing it as opposed to just doing it more carefully. ↩

11. The model could plausibly distinguish these different training regimes because it's typically much harder to generate realistic data than to discriminate it from real data. ↩

12. We'd also like to include other types of model-based policy other than the one described above—for example, a model-based policy which evaluates plans using a learned value function rather than a reward model. ↩

13. Note that, although this definition uses RL terminology, we are open to the possibility of internally-represented goals arising in networks trained only via (self-)supervised learning (e.g. language models which are partly trained to mimic goal-directed humans [Bommasani et al., 2021]). However, for the sake of simplicity we continue to focus on training regimes in which at least some RL is used. A stricter version of this definition could require policies to make decisions using an internally-represented value function, reward function, or utility function over high-level outcomes; this would be closer to Hubinger et al. [2021]'s definition of *mesa-optimizers*. However, it is hard to specify precisely what would qualify, and so for current purposes we stick with this simpler definition. This definition doesn't explicitly distinguish between "terminal goals" which are pursued for their own sake, and "instrumental goals" which are pursued for the sake of achieving terminal goals [Bostrom, 2012]. However, we can interpret "consistently" as requiring the network to pursue a goal even when it isn't instrumentally useful, meaning that only terminal goals would meet a strict interpretation of the definition. ↩

14. We also count a goal as more broadly-scoped to the extent that it applies to other unfamiliar situations, such as situations where the goal could be achieved to an extreme extent; situations where there are very strong tradeoffs between one goal and another; situations which are non-central examples of the goal; and situations where the goal can only be influenced with low probability. ↩

15. Even if an individual instance an AGI policy only runs for some limited time horizon, it may nevertheless be capable of reasoning about the consequences of its plans beyond that time horizon, and potentially launching new instances of the same policy which share the same long-term goal (just as humans, who are only "trained" on lifetimes of decades, but sometimes pursue goals defined over timeframes of centuries or millennia, often by delegating tasks to new generations). ↩

16. It may be impractical to train on such ambitious goals using online RL, since the system could cause damage before it is fully trained Amodei et al. [2016]. But this might be mitigated by using offline RL, which often uses behavioral data from humans, or by giving broadly-scoped instructions in natural language [Wei et al., 2021]. ↩

17. The first additional reason is that training ML systems to interact with the real world often gives rise to feedback loops not captured by ML formalisms, which can incentivize behavior with larger-scale effects than developers intended [Krueger et al., 2020]. For example, predictive models can learn to output self-fulfilling prophecies where the prediction of an outcome increases the likelihood that an outcome occurs [De-Arteaga and Elmer, 2022]. More generally, model outputs can change users' beliefs and actions, which would then affect the future data on which they are trained [Kayhan, 2015]. In the RL setting, policies could affect aspects of the world which persist across episodes (such as the beliefs of human supervisors) in a way that shifts the distribution of future episodes; or they could learn strategies that depend on data from unintended input channels (as in the case of an evolutionary algorithm which designed an oscillator to make use of radio signals from nearby computers [Bird and Layzell, 2002]). While the effects of existing feedback loops like these are small, they will likely become larger as more capable ML systems are trained online on real-world tasks.

    The second additional reason, laid out by Yudkowsky [2016], is that we should expect increasingly intelligent agents to be increasingly rational, in the sense of having beliefs and goals that obey the constraints of probability theory and expected utility theory; and that this is inconsistent with pursuing goals which are restricted in scope. Yudkowsky gives the example of an agent which believes with high probability that it has achieved its goal, but then makes increasingly large-scale plans to drive that probability higher and higher, to maximize its expected utility. Sensitivity to small probabilities is one way in which a goal might be broadly-scoped: the policy pursues the goal further even in situations where it is already achieved with a probability that is very high (but less than 1). ↩

[18] Note that correlations don't need to be perfect in order for the corresponding goals to be reinforced. For example, policies might learn the misaligned goals which are most consistently correlated with rewards, along with narrowly-scoped exceptions for the (relatively few) cases where the correlations aren't present. ↩

[19] It's not a coincidence that acquiring resources is also listed as a convergent instrumental goal in Section 4.1: goals which contribute to reward on many training tasks will likely be instrumentally useful during deployment for roughly the same reasons. ↩

[20] The instrumental convergence thesis is an elaboration of an observation originally made by Omohundro [2008]. ↩

[21] For example, a policy trained using an advantage actor-critic algorithm [Williams and Peng, 1991] could minimize the extent to which its weights are updated by trying to take actions for which the critic estimates $Q(s, a) \approx V(s)$, which would be an example of the hypothesized phenomenon of gradient hacking [Ngo, 2022b]. ↩

[22] Relatedly, existing models can be trained to fail whenever given a specific "backdoor key", where detecting the existence of the backdoor is computationally infeasible [Goldwasser et al., 2022]. ↩

[23] This could happen even when their rewards are zero-sum, because deceptively aligned policies would choose actions based on their long-term consequences, not merely on their short-term rewards. In theory misbehavior which led to lower reward would be trained away eventually, but in practice random exploration is often too slow to find the highest-reward strategies in realistic amounts of time, especially in multi-agent settings. We speculate that exploration problems for actor-critic RL algorithms could be further exacerbated by collusion between situationally-aware actors and critics—e.g. if a single network were trained with both actor and critic heads, and developed goals that influenced the outputs of both heads. This would be an instance of the hypothesized phenomenon of gradient hacking [Ngo, 2022b]. ↩

[24] AGI behavior in this latter class of scenarios would be somewhat analogous to historical examples of multinational corporations attempting to subvert the governments of small countries. ↩

[25] Some have argued that even AGIs with a huge amount of power over humanity would continue to treat us well, since cooperation is more advantageous than conflict. However, at some point the costs of keeping humanity living in good conditions will likely outweigh the benefits of our willing cooperation (as is the case for most animals from the human perspective, including animals like horses which used to have much more to offer when our technology was less advanced). And even if that didn't happen, losing our ability to steer our own future as a species would be a very undesirable outcome regardless. ↩

References

AI Governance Curriculum, 2022. URL https://www.agisafetyfundamentals.com/ai-governance-curriculum.

Adam, Anuj Mahajan, Catarina Barros, Charlie Deck, Jakob Bauer, Jakub Sygnowski, Maja Trebacz, Max Jaderberg, Michael Mathieu, et al. Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*, 2021.

Adept. Act-1: Transformer for actions, 2022. URL https://www.adept.ai/act.

Ajay Agrawal, Joshua Gans, and Avi Goldfarb. *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press, 2018.

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016. URL https://arxiv.org/abs/1606.06565.

Dario Amodei, Paul Christiano, and Alex Ray. Learning from human preferences, 2017. URL https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/.

Jacob Andreas. Language models as agent models. *arXiv preprint arXiv:2212.01681*, 2022.

Stuart Armstrong, Nick Bostrom, and Carl Shulman. Racing to the precipice: a model of artificial intelligence development. *AI & society*, 31(2):201–206, 2016.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, page eade9097, 2022.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

Beth Barnes and Paul Christiano. Debate update: Obfuscated arguments problem - AI Alignment Forum, December 2020. URL https://www.alignmentforum.org/posts/PJLABqQ962hZEqhdB/debate-update-obfuscated-arguments-pr

Jon Bird and Paul Layzell. The evolved radio and its implications for modelling the evolution of novel sensors. In *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*, volume 2, pages 1836–1841. IEEE, 2002.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2021. URL https://arxiv.org/abs/2108.07258.

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

Nick Bostrom. Existential risk prevention as global priority. *Global Policy*, 4(1):15–31, 2013.

Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc., USA, 1st edition, 2014. ISBN 0199678111.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in neural information processing systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*, 2020.

Vanessa Buhrmester, David Münch, and Michael Arens. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction*, 3(4):966–989, 2021.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Joseph Carlsmith. Is power-seeking AI an existential risk? *arXiv preprint arXiv:2206.13353*, 2022.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling, 2021. URL `https://arxiv.org/abs/2106.01345`.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling Language Modeling with Pathways, April 2022. URL `http://arxiv.org/abs/2204.02311`. arXiv:2204.02311 [cs].

Paul Christiano. What failure looks like - AI Alignment Forum, March 2019a. URL `https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like`.

Paul Christiano. Another (outer) alignment failure story - AI Alignment Forum, March 2019b. URL `https://www.alignmentforum.org/posts/AyNHoTWWAJ5eb99ji/another-outer-alignment-failure-story`.

Paul Christiano. Worst-case guarantees. *URL https://ai-alignment. com/training-robust-corrigibility-ce0e0a3b9b4d*, 2019c.

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts, October 2018. URL `https://arxiv.org/abs/1810.08575v1`.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

Casey Chu, Andrey Zhmoginov, and Mark Sandler. Cyclegan, a master of steganography, 2017. URL https://arxiv.org/abs/1712.02950.

Michael K Cohen, Marcus Hutter, and Michael A Osborne. Advanced artificial agents intervene in the provision of reward. *AI Magazine*, 43(3):282–293, 2022.

Ajeya Cotra. Forecasting TAI with biological anchors, 2020. URL https://docs.google.com/document/d/1IJ6Sr-gPeXdSJugFulwIpvavc0atjHGM82QjIfUSBGQ/edit.

Ajeya Cotra. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover - AI Alignment Forum, July 2022. URL https://www.alignmentforum.org/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-p

Andrew Critch. A parametric, resource-bounded generalization of löb's theorem, and a robust cooperation criterion for open-source game theory. *The Journal of Symbolic Logic*, 84(4):1368–1381, 2019.

Allan Dafoe. AI governance: a research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, 1442:1443, 2018.

Maria De-Arteaga and Jonathan Elmer. Self-fulfilling prophecies and machine learning in resuscitation science. *Resuscitation*, 2022.

DeepMind. About, January 2023. URL https://www.deepmind.com/about.

Jonas Degrave. Building a virtual machine inside ChatGPT, 2022. URL https://www.engraved.blog/building-a-virtual-machine-inside/.

Jared M Diamond and Doug Ordunio. *Guns, germs, and steel*, volume 521. Books on Tape, 1999.

Florian E. Dorner. Measuring Progress in Deep Reinforcement Learning Sample Efficiency, February 2021. URL http://arxiv.org/abs/2102.04881. arXiv:2102.04881 [cs].

N Elhage, N Nanda, C Olsson, T Henighan, N Joseph, B Mann, A Askell, Y Bai, A Chen, T Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.

Tom Everitt, Gary Lea, and Marcus Hutter. AGI safety literature review. *arXiv preprint arXiv:1805.01109*, 2018. URL https://arxiv.org/abs/1805.01109.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(27):6435–6467, 2021.

Sebastian Farquhar, Ryan Carey, and Tom Everitt. Path-specific objectives for safer agent incentives. *AAAI*, 2022.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

Daniel Freeman, David Ha, and Luke Metz. Learning to predict without looking ahead: World models without forward prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. Predictability and surprise in large generative models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, jun 2022. doi: 10.1145/3531146.3533229. URL https://doi.org/10.1145%2F3531146.3533229.

Scott Garrabrant. Embedded Agents, October 2018. URL https://intelligence.org/2018/10/29/embedded-agents/.

Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor. Logical induction. *arXiv preprint arXiv:1609.03543*, 2016.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations, 2019. URL https://arxiv.org/abs/1902.03129.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Ben Goertzel. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1, 2014.

Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. Planting undetectable backdoors in machine learning models, 2022. URL https://arxiv.org/abs/2204.06974.

Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will AI exceed human performance? evidence from AI experts. *Journal of Artificial Intelligence Research*, 62:729–754, 2018.

Arthur Guez, Mehdi Mirza, Karol Gregor, Rishabh Kabra, Sébastien Racanière, Théophane Weber, David Raposo, Adam Santoro, Laurent Orseau, Tom Eccles, Greg Wayne, David Silver, and Timothy Lillicrap. An investigation of model-free planning, May 2019. URL http://arxiv.org/abs/1901.03559. arXiv:1901.03559 [cs, stat].

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. In *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels, 2018. URL https://arxiv.org/abs/1811.04551.

Dan Hendrycks. Introduction to ML Safety, 2023. URL https://course.mlsafety.org/about.

Dan Hendrycks and Mantas Mazeika. X-risk analysis for AI research, 2022. URL https://arxiv.org/abs/2206.05862.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI with shared human values. *arXiv preprint arXiv:2008.02275*, 2020.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*, 2021. URL https://arxiv.org/abs/2109.13916.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*, 2022a.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner monologue: Embodied reasoning through planning with language models, 2022b. URL https://arxiv.org/abs/2207.05608.

Evan Hubinger. Bing chat is blatantly, aggressively misaligned, 2023. URL https://www.lesswrong.com/posts/jtoPawEhLNXNxvgTT/bing-chat-is-blatantly-aggressively-misaligned.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from Learned Optimization in Advanced Machine Learning Systems, December 2021. URL http://arxiv.org/abs/1906.01820. arXiv:1906.01820 [cs].

Marcus Hutter. *Universal artificial intelligence: Sequential decisions based on algorithmic probability*. Springer Science & Business Media, 2004.

Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate, May 2018. URL https://arxiv.org/abs/1805.00899v2.

Max Jaderberg, Wojciech Marian Czarnecki, Iain Dunning, Thore Graepel, and Luke Marris. Capture the Flag: the emergence of complex cooperative agents, May 2019. URL https://www.deepmind.com/blog/capture-the-flag-the-emergence-of-complex-cooperative-agents.

Hengrui Jia, Mohammad Yaghini, Christopher A Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. Proof-of-learning: Definitions and practice. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 1039–1056. IEEE, 2021.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL `https://arxiv.org/abs/2207.05221`.

Holden Karnofsky. AI could defeat all of us combined, 2022. URL `https://www.cold-takes.com/ai-could-defeat-all-of-us-combined`.

Varol Kayhan. Confirmation bias: Roles of search engines and search contexts, 2015.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. Specification gaming: the flip side of AI ingenuity, April 2020. URL `https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity`.

David Krueger, Tegan Maharaj, and Jan Leike. Hidden incentives for auto-induced distributional shift, 2020. URL `https://arxiv.org/abs/2009.09153`.

Thomas S Kuhn. *The structure of scientific revolutions*, volume 111. Chicago University of Chicago Press, 1970.

Guillaume Lample, Marie-Anne Lachaux, Thibaut Lavril, Xavier Martinet, Amaury Hayat, Gabriel Ebner, Aurélien Rodriguez, and Timothée Lacroix. HyperTree Proof Search for Neural Theorem Proving, May 2022. URL `http://arxiv.org/abs/2205.11491`. arXiv:2205.11491 [cs].

Lauro Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. PMLR, 2022.

Mark R Leary and Catherine A Cottrell. Evolutionary perspectives on interpersonal acceptance and. *The Oxford handbook of social exclusion*, page 9, 2013.

Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and machines*, 17 (4):391–444, 2007.

Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J Bentley, Samuel Bernard, Guillaume Beslon, David M Bryson, et al. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial life*, 26(2):274–306, 2020.

Jan Leike. A minimal viable product for alignment, March 2022. URL `https://aligned.substack.com/p/alignment-mvp`.

Benjamin A Levinstein and Nate Soares. Cheating death in damascus. *The Journal of Philosophy*, 117(5):237–266, 2020.

Shuang Li, Xavier Puig, Yilun Du, Clinton Wang, Ekin Akyurek, Antonio Torralba, Jacob Andreas, and Igor Mordatch. Pre-trained language models for interactive decision-making. *arXiv preprint arXiv:2202.01771*, 2022.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law, 2018. URL `https://arxiv.org/abs/1803.04585`.

Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of Chess Knowledge in AlphaZero, November 2021. URL `http://arxiv.org/abs/2111.09259`. arXiv:2111.09259 [cs, stat].

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT, June 2022. URL `http://arxiv.org/abs/2202.05262`. arXiv:2202.05262 [cs].

Andrew Y Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

Richard Ngo. AGI Safety From First Principles, September 2020. URL https://drive.google.com/file/d/1uK7NhdSKprQKZnRjU58X7NLA1auXlWHt/view.

Richard Ngo. AGI Safety Fundamentals Alignment Curriculum, 2022a. URL https://www.agisafetyfundamentals.com/ai-alignment-curriculum.

Richard Ngo. Gradient hacking: definitions and examples - AI Alignment Forum, June 2022b. URL https://www.alignmentforum.org/posts/EeAgytDZbDjRznPMA/gradient-hacking-definitions-and-examples.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom In: An Introduction to Circuits. *Distill*, 5(3):e00024.001, March 2020. ISSN 2476-0757. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in.

Stephen M Omohundro. The basic AI drives. In *AGI*, volume 171, pages 483–492, 2008.

OpenAI. AI and Compute, May 2018. URL https://openai.com/blog/ai-and-compute/.

OpenAI. ChatGPT: optimizing language models for dialogue, 2022. URL https://openai.com/blog/chatgpt.

OpenAI. About OpenAI, January 2023a. URL https://openai.com/about/.

OpenAI. Our approach to alignment research, January 2023b. URL https://openai.com/blog/our-approach-to-alignment-research/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models, February 2022. URL http://arxiv.org/abs/2201.03544. arXiv:2201.03544 [cs, stat].

Alicia Parrish, Harsh Trivedi, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Amanpreet Singh Saimbhi, and Samuel R. Bowman. Two-turn debate doesn't help humans answer hard reading comprehension questions, 2022a. URL https://arxiv.org/abs/2210.10860.

Alicia Parrish, Harsh Trivedi, Ethan Perez, Angelica Chen, Nikita Nangia, Jason Phang, and Samuel R. Bowman. Single-turn debate does not help humans answer hard reading-comprehension questions, 2022b. URL https://arxiv.org/abs/2204.05212.

Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gJcEM8sxHK.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red Teaming Language Models with Language Models, February 2022a. URL https://arxiv.org/abs/2202.03286v1.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022b.

Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators, 2022. URL https://arxiv.org/abs/2206.05802.

Juergen Schmidhuber. Reinforcement Learning Upside Down: Don't Predict Rewards – Just Map Them to Actions, June 2020. URL http://arxiv.org/abs/1912.02875. arXiv:1912.02875 [cs].

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the first international conference on simulation of adaptive behavior on From animals to animats*, pages 222–227, Cambridge, MA, USA, February 1991. MIT Press. ISBN 978-0-262-63138-9.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*, 2022.

Joar Max Viktor Skalse, Nikolaus HR Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=yb3HOXO3lX2.

Xingyou Song, Yiding Jiang, Stephen Tu, Yilun Du, and Behnam Neyshabur. Observational overfitting in reinforcement learning. *arXiv preprint arXiv:1912.02975*, 2019.

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. *Advances in Neural Information Processing Systems*, 31, 2018.

Zach Stein-Perlman, Benjamin Weinstein-Raun, and Katja Grace. 2022 Expert Survey on Progress in AI, August 2022. URL https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/. Section: AI Timeline Surveys.

Jacob Steinhardt. More Is Different for AI, January 2022a. URL https://bounded-regret.ghost.io/more-is-different-for-ai/.

Jacob Steinhardt. ML Systems Will Have Weird Failure Modes, January 2022b. URL https://bounded-regret.ghost.io/ml-systems-will-have-weird-failure-modes-2/.

Jacob Steinhardt. Emergent deception and emergent optimization, 2023. URL https://bounded-regret.ghost.io/emergent-deception-optimization/.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2020. URL https://arxiv.org/abs/2009.01325.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Alex Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. Optimal Policies Tend To Seek Power, December 2021. URL https://neurips.cc/virtual/2021/poster/28400.

Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, 2022.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Erik Wijmans, Manolis Savva, Irfan Essa, Stefan Lee, Ari S. Morcos, and Dhruv Batra. Emergence of maps in the memories of blind navigation agents. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lTt4KjHSsyl.

Claus O Wilke, Jia Lan Wang, Charles Ofria, Richard E Lenski, and Christoph Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333, 2001.

Chiara Wilkinson. The people in intimate relationships with AI chatbots, 2022. URL https://www.vice.com/en/article/93bqbp/can-you-be-in-relationship-with-replika.

Ronald J Williams and Jing Peng. Function optimization using connectionist reinforcement learning algorithms. *Connection Science*, 3(3):241–268, 1991.

David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback, 2021. URL `https://arxiv.org/abs/2109.10862`.

Eliezer Yudkowsky. Nearest unblocked strategy, 2015. URL `https://arbital.com/p/nearest_unblocked/`.

Eliezer Yudkowsky. The AI alignment problem: why it is hard, and where to start. *Symbolic Systems Distinguished Speaker*, 2016. URL `https://intelligence.org/2016/12/28/ai-alignment-why-its-hard-and-where-to-start/`.

Eliezer Yudkowsky et al. Artificial intelligence as a positive and negative factor in global risk. *Global catastrophic risks*, 1(303):184, 2008.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models, 2022. URL `https://arxiv.org/abs/2205.10625`.

Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned AI. *Advances in Neural Information Processing Systems*, 33:15763–15773, 2020.

Daniel M Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Ben Weinstein-Raun, Daniel de Haas, et al. Adversarial training for high-stakes reliability. *Neural Information Processing Systems*, 2022.