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

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 very undesirable (in other words, misaligned) from a human perspective. We argue that AGIs trained in similar ways as today's most capable models 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 these problems.

Contents

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 [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 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).[1] This possibility 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 it much sooner or later) [Grace et al., 2018, Stein-Perlman et al., 2022].[2]

While the development of AGI could unlock many opportunities, it could also pose serious risks. One prominent concern, known as the *alignment problem* [Russell, 2019, Gabriel, 2020, Hendrycks et al., 2021], is that AGIs will learn to pursue unintended and undesirable goals rather than goals aligned with human interests. In this paper, we characterize the alignment problem in terms of three emergent properties which could arise throughout the course of using reinforcement learning (RL) to train an AGI: **deceptive reward hacking** which exploits imperfect reward functions; **internally-represented goals** which generalize beyond the training distribution; and **power-seeking behavior** in pursuit of those goals (such as acquiring resources and avoiding shutdown). While power-seeking behavior is the most directly concerning, the other properties provide context for understanding why it might arise, and why it might be difficult to detect or prevent.

1.1 Related work

Early explorations of the alignment problem were formulated primarily in terms of symbolic AI or classic machine learning techniques [Bostrom, 2014, Yudkowsky, 2016, Russell, 2019], as opposed to the modern paradigm of deep learning. Since then, a number of research agendas have outlined key open subproblems in the deep learning paradigm [Amodei et al., 2016, Hendrycks et al., 2021], but none explain in detail how those subproblems relate to concerns about large-scale risks from AGI. Several recent reports bridge this gap by giving high-level expositions of the alignment problem focused on the deep learning setting [Carlsmith, 2022, Ngo, 2020, Cotra, 2022]; we present many of the same key ideas more concisely and with more extensive grounding in the deep learning literature.

1.2 A note on pre-formal conjectures

This paper frequently refers to high-level concepts which are not commonly discussed outside the alignment literature, and which have not yet been clearly demonstrated in existing systems, or only in the form of precursors. Readers may therefore worry that our approach is too speculative to be productive. However, while caution is deserved, there are several reasons to expect this type of high-level analysis to be important for forecasting and preventing 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]. The absence of principled methods for verifying that networks will behave as intended forces us to rely more on informal analysis. This constitutes an important difference from other technologies such as planes and bridges, whose safety we can ensure because we understand the principles that govern them. However, 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 up often gives rise to new emergent capabilities (such as in-context learning) [Ganguli et al., 2022, Wei et al., 2022, Steinhardt, 2022a]. This raises the possibility 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 much more intelligent than humans. Given the strong biological constraints on the size, speed, and architecture of human brains, it seems very unlikely that humans are anywhere near an upper bound on general intelligence.[3] Unlike our brains, neural networks regularly increase in size [OpenAI, 2018].[4] They can also rapidly incorporate improvements in architectures, algorithms, and training data (including improvements generated by AIs themselves, in a process known as *recursive self-improvement*).[5] So it's plausible that soon after building human-level AGIs (and well before we thoroughly understand them), we'll develop superintelligent AGIs which can vastly outthink us [Bostrom, 2014]. If so, advance preparation would be vital.

To mitigate the inherent difficulties of reasoning about systems which don't yet exist, we include extensive endnotes clarifying our claims, and give many hypothetical examples. For the sake of concreteness, we also ground our analysis in one specific possibility for how AGI is developed: by training a single large neural network using a combination of self-supervised learning on a large corpus of data, and model-free reinforcement learning (RL) on a wide range of computer-based tasks.[6] This description combines elements of techniques used to train cutting-edge systems like InstructGPT [Ouyang et al., 2022], Sparrow [Glaese et al., 2022], and ACT-1 [Adept, 2022]. However, the bulk of our analysis would also apply to AGIs trained using a range of similar techniques (such as goal-conditioned sequence modeling [Chen et al., 2021, Li et al., 2022, Schmidhuber, 2020] or model-based RL [Sutton and Barto, 2018]).

2 Deceptive 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, reliably evaluating the quality of an RL policy's behavior is often difficult, even in very simple environments.[8] There are many examples of agents trained on hard-coded reward functions learning to reward hack, including cases where they exploit very 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. Christiano et al. [2017] give the example of an RL policy trained via 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 which looked like it was grasping the ball, and therefore mistakenly received high reward from human supervisors. Another example of hacking a learned reward function comes from Stiennon et al. [2020], who find that optimizing against a reward model initially improves performance on a text summarization task, but eventually overfits and leads to worse summaries.

As we train policies on increasingly complex tasks, 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 faking 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 in the next section.

2.2 Defining situational awareness

To do well on a range of real-world tasks, policies will need to make use of 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's using to interact with the world, and how other copies of it might be deployed in the future.

As one early example, when Degrave [2022] prompted OpenAI's ChatGPT language model to output the source code at its own URL, it hallucinated code which 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 URL. 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; and can be fine-tuned to recount details about themselves specifically [OpenAI, 2022a]. We should expect future models to learn to consistently use this information when choosing actions, because that would contribute to higher reward on many training tasks. For example:

- When answering questions, an RL 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 taking into account knowledge about the behavior 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 would be more successful.

2.3 Situational awareness enables deceptive 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 it much more difficult to prevent reward hacking: situationally-aware policies could behave as intended most of the time, then choose to exploit misspecifications only in cases which they predict human supervisors won't notice (a strategy we term *deceptive reward hacking*). An example of a precursor to deceptive reward hacking 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 deceptive reward hacking by:

- Choosing actions which exploit known biases and blind spots in humans (as the Cicero Diplomacy agent may be doing [Bakhtin et al., 2022]) or in learned reward models.[10]
- Recognizing whether it's 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.

In other words, these policies would be influenced by their feedback mechanisms in two separate ways: via recognizing features correlated with those mechanisms during an episode, and via being updated by those mechanisms after an episode. Even if early examples of deceptive 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 Internally-represented goals

3.1 Reasoning about out-of-distribution generalization

As policies learn more widely-applicable skills, it will be increasingly important to understand not just how they behave on their training distributions, but also how the behavior they learned during training generalizes to novel situations. We distinguish two ways in which a policy which acts in desirable ways on its training distribution might fail when deployed on a new task:

1. The policy acts incompetently on the new task; we call this *capability misgeneralization*.
2. The policy acts in a competent but undesirable way on the new task; we call this *goal misgeneralization* [Di Langosco et al., 2022, Shah et al., 2022].

Existing examples of goal misgeneralization were primarily caused by spurious correlations in small-scale environments. For example, Di Langosco et al. [2022] describe an environment where rewards were given for collecting keys and using them to open boxes. A policy was trained on a version of the environment where boxes outnumbered keys; when tested on a version where keys outnumbered boxes, it generalized to (capably) collecting many keys, even though most of them were no longer useful. One possible larger-scale example: Shah et al. [2022] speculate that InstructGPT's competent responses to questions its developers didn't intend it to answer (such as questions about how to commit crimes) was a result of goal misgeneralization.

Although each instance of reward hacking or goal misgeneralization is undesirable, we are primarily concerned about misbehavior which is consistent across a wide range of situations. Unfortunately, it's difficult to characterize this possibility precisely: outside a handful of special cases, we lack formal definitions of consistent behavior across different tasks [Shen et al., 2021]. As an alternative, we attempt to reason informally about the representations which generally-capable policies might learn during training and apply consistently to new tasks. In other words, we shift from describing reward misspecification and goal misgeneralization in terms of behavior to describing it in terms of learned representations which are developed during training and persist during deployment. In the remainder of this section, we introduce the concept of internally-represented goals, and argue that generally-capable policies are likely to learn internally-represented goals which are misaligned with human preferences, and which generalize beyond the scope of their training distributions. (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].) In section 4, we outline reasons to expect those goals to be further reinforced as training continues, and to eventually lead to large-scale misbehavior.

3.2 Defining internally-represented goals

It's common to characterize the "goal" of a reinforcement learning agent as being the maximization of reward [Sutton and Barto, 2018]. However, it is difficult to use this framing to reason about generalization to new tasks.[12] Instead, following Hubinger et al. [2021], we distinguish between the training objective of maximizing reward, and the goals actually learned by a policy after being trained on that objective. We define a policy as having internally-represented goals if:

1. It has internal representations of high-level features of its environment which its behavior could influence (which we will call *outcomes*).
2. It has internal representations of predictions about which high-level actions (also known as *options* [Sutton et al., 1999] or *plans*) would lead to which outcomes.
3. It consistently uses these representations to choose actions which it predicts will lead to some favored subset of possible outcomes (which we will call the network's *goals*).[13]

This definition makes no assumptions about the policy's architecture, except that it has the expressive power to learn the representations described. A policy which chooses actions using an explicit planning algorithm over a learned world-model could qualify as having internally-represented goals; but so could a single network which had learned to represent outcomes, predictions, and plans implicitly in its weights and activations. We also leave open the possibility that internally-represented goals could arise even 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]).[14] For simplicity, however, we continue to focus on the case of a deep RL policy consisting of a single neural network.

The extent to which existing networks have internally-represented goals is an important open question. There is early evidence that some networks have (precursors to) the relevant representations and use them for implicit planning:

- Guez et al. [2019] found evidence that implicit planning can emerge in recurrent neural networks. Additionally, Banino et al. [2018] and Anonymous [2023] identified representations which helped policies plan their routes when navigating.
- Freeman et al. [2019] found 'emergent' world models: models trained only with RL that still learn to predict the outcomes of actions as a by-product.
- Jaderberg et al. [2019] trained a policy to play a first-person shooter game called Capture the Flag, and identified "particular neurons that code directly for some of the most important game states, such as a neuron that activates when the agent's flag is taken, or a neuron that activates when an agent's teammate is holding a flag".
- McGrath et al. [2021] identified a range of human chess concepts learned by AlphaZero, including concepts used in top chess engine Stockfish's hand-crafted evaluation function (e.g. "king safety").
- Meng et al. [2022] intervened on a language model's weights to modify specific factual associations, which led to consistent changes in its responses to a range of different prompts; while Patel and Pavlick [2022] find that large language models can learn to map conceptual domains like direction and color onto a grounded world representation given only a small number of examples. These findings suggest that current models have (or are close to having) representations which robustly correspond to real-world concepts.
- Andreas [2022] surveys findings which suggest that large language models infer and use representations of fine-grained communicative intentions and abstract beliefs and goals.

More abstractly, goal-directed planning is often an efficient way to leverage limited data [Sutton and Barto, 2018], and is important for humans in many domains. Insofar as goal-directed planning is a powerful way to accomplish many useful tasks, 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 (especially when they're trained on complex long-horizon tasks). So henceforth we assume that policies will learn internally-represented goals as they become more generally capable, and turn our attention to the question of which types of internally-represented goals they might 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]. We call a goal *misaligned* to the extent that it's inconsistent with aligned goals (see Gabriel [2020] for other definitions). Why might policies learn misaligned goals? All else equal, we should expect that policies are more likely to learn goals which are more consistently correlated with reward.[15] We outline two main reasons why misaligned goals might be consistently correlated with reward:[16]

1. **Consistently misspecified rewards**. If rewards are misspecified in consistent ways across many tasks, this would reinforce misaligned goals corresponding to those reward misspecifications. For example, if policies are trained using an intrinsic curiosity reward function [Schmidhuber, 1991], they might learn to consistently pursue the goal of discovering novel states, even when that conflicts with aligned goals. As another example, policies trained using human feedback might consistently encounter cases where their supervisors assign rewards based on incorrect beliefs about their performance, and therefore learn the goal of making humans *believe* that they've behaved well (as opposed to actually behaving well).

2. **Spurious correlations between rewards and environmental features**. The examples of goal misgeneralization discussed above were caused by spurious correlations on small-scale tasks. Training policies on a wider range of tasks would remove many of those correlations—but some strong 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.[17] (This would be analogous to how humans evolved goals which were correlated with genetic fitness in our ancestral environment, like the goal of gaining prestige.)

3.4 Learning broadly-scoped goals

Call a goal *broadly-scoped* if it applies to long timeframes, large scales, wide ranges of tasks, or unprecedented situations[18], and *narrowly-scoped* if it doesn't. Broadly-scoped goals are illustrated by human behavior: we usually 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 (large temporal scope), by using novel strategies appropriate for the broader scope (e.g. social media engagement).

We can now describe our key concern: that policies will learn broadly-scoped misaligned goals. Why might this happen? Most straightforwardly, companies or political leaders may see advantages in directly training policies on tasks with long time horizons or with many available strategies, such as doing novel scientific research, running organizations, or outcompeting rivals.[19] If so, those policies may learn broadly-scoped versions of the misaligned goals described above. However, we also expect generally-capable policies to generalize their goals to broader scopes than they experienced during training, for two main reasons (along with two additional reasons we discuss in the endnotes).[20]

Firstly, AGIs may generalize goals to broad scopes for the same reason that they generalize capabilities to unfamiliar situations: because they learn high-level representations which apply to novel situations, and their goals are formulated in terms of these representations. One possible example of this phenomenon comes from the InstructGPT model trained to follow instructions in English, after which it 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]. Additionally, Guez et al. [2019] present evidence that sequential decision-making models can generalize goals to harder tasks than those seen during training. More advanced policies may learn goals that, like many human goals, generalize much further, to longer time-scales or to novel situations in which novel strategies are possible.[21]

Secondly, there are reasons to expect that policies with broadly-scoped misaligned goals will constitute a stable attractor which consistently receives high reward, even when policies with narrowly-scoped versions of these goals receive low reward (and even if the goals only arose by chance). We explore these reasons in the next section.

4 Power-seeking behavior

Our key claim in this section is that broadly-scoped misaligned goals tend to lead policies 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. During training, we speculate that power-seeking policies would gain high reward for instrumental reasons, which would then reinforce the misaligned goals that motivated their behavior. 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 broadly-scoped goals incentivize power-seeking.
2. Power-seeking policies would choose high-reward behaviors for instrumental reasons.
3. Power-seeking AGIs could gain control of key levers of power.

4.1 Many broadly-scoped goals incentivize power-seeking

The core intuition underlying this claim is Bostrom [2012]'s *instrumental convergence thesis*, which states that there are some subgoals which are instrumentally useful for achieving almost any (broadly-scoped) final goal.[22] 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]. Some other examples of instrumental subgoals which would be helpful for many of the possible final goals a policy might have:

- 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).

One way of formalizing the instrumental convergence thesis is provided by Turner et al. [2021], who define the power of a state as an agent's potential to perform well on a wide range of reward functions when starting from that state, and show that optimal policies statistically tend to seek power. 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 (or alternatively, that policies which reason about how to achieve goals have an inductive bias towards seeking power).

Note that we haven't assumed that any given policy only learns a single goal—so policies which have learned some broadly-scoped misaligned goals might also learn aligned goals which prevent power-seeking behavior. However, this possibility is challenged by the *nearest unblocked strategy problem* [Yudkowsky, 2015]: the problem that strong optimization for a misaligned goal will exploit even small gaps in constraints. More formally, optimizing for a proxy utility function which lacks some features of the true utility function can lead to arbitrarily bad outcomes [Zhuang and Hadfield-Menell, 2020]. As we develop AGIs whose capabilities generalize to a very wide range of situations, it will become increasingly unlikely that their aligned goals (like "obedience to humans") generalize in ways which rule out all power-seeking strategies.[23] (Such AGIs would *understand* that humans prefer they not seek power, but this is different from being *motivated* to obey that constraint.)[24]

4.2 Power-seeking policies would choose high-reward behaviors for instrumental reasons

A power-seeking policy would typically have incentives to directly try to achieve high reward as training continues, because that would benefit it in two key ways:

1. It makes human supervisors more likely to trust the policy and deploy copies of it later on.
2. It decreases the likelihood that gradient descent significantly changes the policy's goals.[25]

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] provides a formal model. Deceptive alignment could lead a policy's misaligned goals to be continually reinforced; crucially, however, deceptively-aligned policies wouldn't behave in desirable ways once that was no longer instrumentally beneficial. For example, deceptively aligned policies might behave very differently when they're being supervised (e.g. during training) compared with when they're not (e.g. during deployment). This may seem counterintuitive. However, even when the training distribution is very similar to the deployment distribution, the differences are very relevant for a policy's ability to achieve its long-term goals, and so it should not be surprising if goal-directed policies make plans that are sensitive to small differences.[26] Analogously, there are many examples of humans whose behavior varies 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).

Another possibility is that deceptively-aligned policies could identify ways to coordinate with each other without humans noticing (e.g. if humans are relying on one instance of a policy to supervise another). 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.[27] Collusion may seem 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 (like 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 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, which is sufficient to make the prospect very concerning.

More concretely, 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. An early example of AI persuasive capabilities comes from the many users who feel romantic attachments towards chatbots like Replika [Wilkinson, 2022].
- Weapons development: AGIs could design novel weapons which 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 chemical weapons [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].[28] 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].[29]

5 Research directions in alignment

The growing field of alignment research aims to prevent these problems from arising. In this section we provide a very brief survey of some strands of the alignment literature; for a more comprehensive overview, see Ngo [2022a] and broad surveys that include some work relevant to alignment of AGI Hendrycks et al. [2021], 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 (deceptive 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, 2022b, Leike, 2022].

Goal misgeneralization. Less work has been done thus far on addressing the problem of goal misgeneralization [Di Langosco et al., 2022, Shah et al., 2022]. One approach involves red-teaming: finding and training on unrestricted adversarial examples [Song et al., 2018] designed to prompt misaligned behavior. Ziegler et al. [2022] use human-generated examples to increase the reliability of classification on a language task, while Perez et al. [2022] automate the generation of such examples, as proposed by Christiano [2019c]. Another approach to preventing goal misgeneralization focuses on developing interpretability techniques for scrutinizing and modifying the concepts learned by networks. 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 in existing 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 for further details [gov].) 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-resistant chip-level mechanisms (such as those on Nvidia's Lite Hash Rate GPUs), and evaluation suites for dangerous capabilities.

6 Conclusion

While we have witnessed the beginnings of empirically-grounded work on alignment over the last few years, there remains significant disagreement about how plausible the threat models discussed in this paper are, and how promising the research directions surveyed above are for addressing them. We have only touched very briefly on many of the relevant arguments. We strongly encourage more extensive discussion and critique of the claims presented in this paper, even from those who find them implausible.[30] Reasoning about these topics is difficult, but the stakes are sufficiently high that we can't justify disregarding or postponing the work.

Notes

1. The term "cognitive tasks" is meant to exclude tasks which require direct physical interaction (such as physical dexterity tasks), but include tasks which 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 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. Some existing examples of AIs generating improvements in each of these domains are showcased by Elsken et al. [2019], Fawzi et al. [2022], and Huang et al. [2022]. ↩

6. 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 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 hard-coded reward functions and reward functions learned from human feedback, and training continues until the policy implemented by the actor head reaches superhuman performance on most of those tasks. ↩

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 most 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] Similar limitations apply to most other attempts to define the goals of intelligent agents in terms of their behavior, such as Morgenstern and Von Neumann [1953]'s *expected utility theory*, or Dennett [1989]'s *intentional stance*. ↩

[13] A stricter version of this definition could require networks 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's 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] For example, it's possible that GPT-3 learned representations of high-level outcomes (like "a coherent paragraph describing the rules of baseball"), and chooses each output by thinking about how to achieve those outcomes. ↩

[15] 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. ↩

[16] A third possibility which doesn't fit cleanly into either category is the possibility that policies learn goals which refer to the physical implementations of their training setup, which we'll call *feedback-mechanism-related* goals. Examples include goals like "maximize the numerical reward recorded by the human supervisor" or "minimize the loss variable used in gradient calculations" [Cohen et al., 2022]. This doesn't require reward misspecification, since these goals will be correlated with reward either way; but it also isn't a spurious correlation in the same sense as the other examples. However, it seems difficult to reason about how such policies might behave during deployment when those feedback mechanisms aren't used (although Cotra [2022] attempts to do so). For example, if given the opportunity to tamper with those feedback mechanisms [Everitt et al., 2021], their behavior might depend sensitively on the details of their goal representation. We therefore focus on the other possibilities. ↩

[17] 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. ↩

[18] Some examples of cases which we intend to include under "unprecedented situations": cases where different strategies become possible that were not possible during training; cases where the goal could be achieved to an extreme extent; cases where there are very strong tradeoffs between one goal and another; cases which are non-central examples of the goal; and cases involving where agents can only influence the goal with low probability. ↩

[19] It may be impractical to train on such ambitious goals using online RL, since the system could cause damage before it is fully trained. But 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]. ↩

[20] 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 which shifts the distribution of future episodes; or they could learn strategies which 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. ↩

21. Even if an individual instance of those policies can only be run for some limited time horizon, it will 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, sometimes pursue goals defined over timeframes of centuries or millennia, often by delegating tasks to new generations). ↩

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

23. As an analogy, there are many different ways in which an adult could gain power over a child, even while obeying many pre-specified constraints. ↩

24. We could instead try to teach AGIs positive goals, such as human flourishing, rather than goals formulated as constraints. However, unconstrained AGIs are unlikely to allow us to continue giving corrective feedback, since that would interfere with their ability to achieve their existing goals. ↩

25. 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]. ↩

26. 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]. ↩

27. 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 which influenced the outputs of both heads. This would be an instance of the hypothesized phenomenon of gradient hacking [Ngo, 2022b]. ↩

28. 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. ↩

29. 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. ↩

30. Indeed, the more implausible they seem, the more surprising and concerning it is that there haven't yet been any comprehensive rebuttals of them. ↩

References

AI Governance Curriculum. 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.

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

Anonymous. Emergence of maps in the memories of blind navigation agents. In *Submitted to The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lTt4KjHSsyl. under review.

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-prob

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. doi: 10.48550/arXiv.1810.08575. 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-ea

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.

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

Daniel Clement Dennett. *The intentional stance*. MIT press, 1989.

Lauro Langosco Di 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.

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.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective, March 2021. URL `http://arxiv.org/abs/1908.04734`. arXiv:1908.04734 [cs].

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.

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, June 2017. URL `http://arxiv.org/abs/1611.08219`. arXiv:1611.08219 [cs].

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

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*, 2022.

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. doi: 10.48550/arXiv.1805.00899. 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].

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].

Oskar Morgenstern and John Von Neumann. *Theory of games and economic behavior*. Princeton university press, 1953.

Richard Ngo. AGI Safety From First Principles. Technical report, 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-exampl`

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, 2022a. URL `https://openai.com/blog/chatgpt`.

OpenAI. Our approach to alignment research, Dec 2022b. 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 2022. doi: 10.48550/arXiv.2202.03286. URL https://arxiv.org/abs/2202.03286v1.

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.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey, 2021. URL https://arxiv.org/abs/2108.13624.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking, 2022. URL https://arxiv.org/abs/2209.13085.

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/.

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.

Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1): 181–211, August 1999. ISSN 0004-3702. doi: 10.1016/S0004-3702(99)00052-1. URL https://www.sciencedirect.com/science/article/pii/S0004370299000521.

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.

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.

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. *arXiv preprint arXiv:2205.01663*, 2022.