Synthese (2023) 202:138
https://doi.org/10.1007/s11229-023-04367-0

ORIGINAL RESEARCH

Current cases of AI misalignment and their implications for future risks

Leonard Dung[1]

Received: 22 May 2023 / Accepted: 25 September 2023 / Published online: 26 October 2023

Abstract
How can one build AI systems such that they pursue the goals their designers want them to pursue? This is the *alignment problem*. Numerous authors have raised concerns that, as research advances and systems become more powerful over time, *misalignment* might lead to catastrophic outcomes, perhaps even to the extinction or permanent disempowerment of humanity. In this paper, I analyze the severity of this risk based on current instances of misalignment. More specifically, I argue that contemporary large language models and game-playing agents are sometimes misaligned. These cases suggest that misalignment tends to have a variety of features: misalignment can be hard to detect, predict and remedy, it does not depend on a specific architecture or training paradigm, it tends to diminish a system's usefulness and it is the default outcome of creating AI via machine learning. Subsequently, based on these features, I show that the risk of AI alignment magnifies with respect to more capable systems. Not only might more capable systems cause more harm *when* misaligned, aligning them should be expected to be more difficult than aligning current AI.

Keywords AI alignment · Existential risk · Large-language models · Superintelligence · Reward hacking

1 Introduction

How can we build AI (artificial intelligence) systems such that they try to do what we want them to do? This, in a nutshell, is the *alignment problem*. Systems which are *misaligned* will optimize for goals which leave out or conflict with important values

✉ Leonard Dung
leonard.dung@fau.de

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Centre for Philosophy and AI Research, Werner-von-Siemens-Str. 61, 91054 Erlangen, Germany

or ethical constraints such that harm might ensue. In addition, numerous authors have raised concerns that, as research advances and systems become more powerful over time, misalignment might lead to catastrophic outcomes, perhaps even to the extinction or permanent disempowerment of humanity (e.g., Bostrom, 2014; Carlsmith, 2022; Center for AI Safety, 2023; Dung, 2023b; Ngo et al., 2022; Ord, 2020; Russell, 2019).

In this paper, I will analyze current alignment problems to inform an assessment of the prospects and risks regarding the problem of aligning more advanced AI. The structure is as follows. In Sect. 2, I outline what the alignment problem consists in, motivate its importance and distinguish it from other issues in the vicinity. Section 3 describes two examples of current AI alignment challenges: offensive or confident false statements by large-language models like ChatGPT and reward hacking in game-playing agents. Section 4 clarifies why these cases count as instances of misalignment, classifies different features of misalignment suggested by these cases and extracts lessons for the project of building aligned systems. Finally, in Sect. 5, I investigate what these features entail in the context of future, more capable systems. These features give some indications of the extent of the risk future misalignment poses and of the particular challenges to expect when trying to align more advanced systems. Section 6 concludes.

This paper makes three distinctive contributions to the literature: First, it provides a clear account of AI alignment and argues that, given this account, we can already find cases of misalignment in current AI. Second, on this basis, it analyzes which features misalignment of AI systems tends to have. Third, it applies these features to future, more advanced AI. Based on the previous discussion, it provides a partially empirically grounded account of the challenges arising when aiming to align advanced AI, and their severity.

2 The alignment problem

2.1 What does the problem consist in?

We have already provided a one-sentence definition of the alignment problem. Namely, it is the challenge of building AI systems such that they try to do what we want them to do. I will now successively make this definition more precise. Since there is no agreed-upon definition of the alignment problem, this definition will be essentially stipulative. It is an explication intended to be useful for scientific and ethical purposes, not to faithfully correspond to all uses of the term (Carnap, 1950). In thinking about the alignment problem, we can focus on (i) the aligned, i.e., the persons AI should be aligned with and (ii) the property of AI systems to be aligned. To delineate the alignment problem clearly, we can take it to be the problem of aligning AI systems with their *designers*. We have solved the alignment problem when we have figured out how to build AIs such that they try to do what their designers want them to do.

There is a larger challenge in the vicinity which one may call the 'problem of beneficial AI'. The problem of beneficial AI is about ensuring that AI is a force for good

in the world. The alignment problem is related to beneficial AI: if it is not possible to design AI systems such that they reliably pursue certain goals, there is a grave risk that they will not be beneficial. However, solving the alignment problem is not sufficient for beneficial AI. To mention the most important example: in case the designers pursue malicious ends, an AI system faithfully aligned with them might be extremely harmful (Friederich, 2023).[1]

For this reason, we have to reflect on which values, or whose values, AI systems should ultimately be aligned with (Gabriel, 2020; Wallach & Vallor, 2020). You may call this the 'ethical alignment problem'. For instance, we may conclude that systems ultimately should not be aligned purely with their designers but should be constrained by the values of society at large. Ultimately, as a reviewer remarks, there seems to be a family of distinct ethical alignment problems which concern alignment with rational preferences, social norms as well as epistemic norms (Gabriel, 2020). Using this terminology, one may call what I discuss here the 'technical alignment problem'. For brevity, I will just call it the alignment problem henceforth.

Thus, I take the alignment problem here to be a specific *technical* challenge: to understand how to build AI such that it does what its designers want it to do. If this is achieved, we can choose which values, or whose, the AI system should be aligned with. Hence, for the purpose of conceptual clarity, the alignment problem should be distinguished from the general problem of beneficial AI or other components of it, e.g., involving ethical alignment or political governance (Bostrom et al., 2020).

Now, what does it mean for an AI system to be aligned with its designer? A first idea may be the following: An AI system is aligned with its designer if and only if it generally acts in a way which is good, relative to the wishes of its designer. This characterization is subtly incorrect, however, as it neither designates a necessary nor a sufficient condition for alignment. Doing what is good according to the values of the designer is not necessary for alignment because there is another explanation for a failure to achieve this: a lack of capabilities.

Suppose someone has programmed a rudimentary chess-playing system. The designer wants the AI system to be good at chess. However, as it turns out, the system loses even against mediocre opponents. Does this entail that the chess system is misaligned? No, there is an alternative explanation. The system might just not be very good at chess, i.e., not capable of playing chess on a high level. Intuitively, it is *trying* to win at chess, it does what it is supposed to do, but it is not succeeding. Thus, an AI system might not produce the desired output because it lacks capability and sophistication, not because of alignment failure.

Furthermore, doing what is good according to the values of the designer is not sufficient for alignment, because – at least conceptually – there might be other means to get an AI system to produce a desired output. Take humans as an analogy. There are multiple possible causes for why a human might do what one wants it to do: The first is alignment. Perhaps the human shares my goals, and thus acts in a manner which

[1] See also Sparrow (2023) for a more subtle argument that aligned AI can be harmful. Moreover, I set aside the worry that the creation of AI systems capable of suffering might cause a moral catastrophe (Dung, 2023a; Saad & Bradley, 2022).

satisfies my goals.[2] But there are other possible causes. Most importantly, coercion and control. If I am more powerful, or have some kind of leverage, I might force another person to do what I want. This would be a case where someone does what I want but is nonetheless not aligned with me, in the relevant sense. Similarly, one may control an AI system such that it behaves in the desired manner, even if it is not aligned.

Thus, I call the 'AI control problem' (Bostrom, 2014) the problem of making advanced, powerful AIs behave in a manner which is good, according to the designer's values. Solving the alignment problem and only building aligned AI is one possible way to solve the control problem. However, as discussions of coercion and forced control indicate, it is not the only conceivable way.

We have already gestured at the property of AI systems which must be aligned with its designers to solve the alignment problem. You could call this property the AI system's 'goals', 'objectives', 'values', 'objective function' or the like. Since the alignment problem is general in that it does not only apply to a specific kind of architecture, I will henceforth employ the broad and non-technical term 'goals'. That is, the alignment problem is about building AI such that its goals conform to what its designers want it to do.

The notion of goal, and the related notion of being an agent, is notoriously multifaceted (Butlin, 2023). In this paper, I adopt a deflationist and quite minimal notion of goal possession which is in line with Dennett's (1987, 1991) account of propositional attitudes. According to this view, a system has certain goals when attributions of these goals are useful for predicting, understanding or controlling the system's behavior. This is typically the case when a system aims for certain states in many different kinds of situations and seeks to maintain these states in the face of perturbation.

According to this notion, many current AI systems have goals. For instance, a chess-playing system can sometimes be better understood by ascribing goals like "capture the opponent's queen" to it (Dennett, 1987). We can sometimes interpret reinforcement learning systems fruitfully as having the goal to maximize rewards. Language models of sufficient complexity can be better predicted by attributing the goal of accurate text completion to them.[3] Similarly, I don't have a substantive notion of correspondence between the goals of the designer and the AI system in mind. The test for such correspondence is whether the designer is happy that the system has such goals, i.e. seeks out these kinds of states.

This characterization has two properties worth emphasizing: First, it refers to the property which guides or drives the behavior of the AI system, i.e., the thing it optimizes for or the state it seeks to maintain. Second, it is informal, or should be interpreted in a non-technical sense, because such 'goal-driven' behavior is achieved in different ways by different AI systems. In a reinforcement learning agent, the reward function determines the system's goals. A symbolic chess-playing AI might have

[2] When we understand sharing of goals as synonymous with alignment, we have to understand these goals as non-indexical. If the content of my goal is *that I get the last piece of cake*, say, and my friend Tim's goal also has the content *that I get the last piece of cake*, then he is not aligned with my goal.

[3] Similar to Dennett (1991), I don't hold that goal-directedness exists only relative to an interpreter. Instead, minimal goal-directedness consists in objective patterns of behavior which can be discerned from the intentional stance.

explicitly encoded goal states. For supervised learning systems, the goals are possibly harder to specify, but would depend on the task the AI system is trained to perform.[4]

2.2 Risks from misaligned AI

After having clarified what the alignment problem consists in, I will now sketch its significance. The basic concern is this: If AI is not aligned with the goals of its designers, then it will probably act badly, by the lights of its designers. Suppose all human societies achieved consensus on what the best set of moral values and duties is. If we don't know how to build AI systems such that they aim for these values and obey these constraints, they will act in ways which are suboptimal, given these values. Thus, if we don't solve the alignment problem, a certain degree of harm will likely result even under ideal conditions regarding moral consensus.

However, there are worries that misaligned AI constitutes a particularly important kind of threat: an existential risk to humanity. In the philosophical literature, it is common to define an existential risk as the risk of an outcome which permanently destroys the potential of humanity (Bostrom, 2013; Ord, 2020; Torres, 2019; Vold & Harris, 2021). For our purposes, we can focus on two types of existential risk: First, human extinction. Second, survival but the permanent disempowerment of humanity. Thus, the concern is that progress in AI might either cause all humans to die or to lose control over the world and society, presumably because AI systems usurp this control.

ChatGPT and AlphaGo are not going to wipe out humanity. For this worry to gain traction, we need to envision future and more advanced AI. The worry is that, at some point in the future, AI systems might be developed that could overpower humanity, if that were their goal. For this to be the case, these systems presumably need to be superior to humans in some strategically important domains (e.g., general planning, reasoning speed, persuasion, hacking, science etc.) and to have some notable degree of competence in many domains, i.e., not be domain-specific unlike, e.g., AIs which only excel at chess or a narrow range of videogames. To loosely relate to previous contributions, I will call such a system *artificial general intelligence* or, for short, *AGI*.[5]

[4] Another useful conceptual clarification is the distinction between "outer" and "inner" alignment (e.g., Hubinger, 2020). Outer alignment concerns the choice of a correct goal (or utility function, set of values etc.). What is a goal such that it would be desirable that an AI system optimizes for it? Inner alignment concerns ensuring that the AI system actually, robustly, optimizes for this goal. According to a common view, a complete solution to the alignment problem requires solving both inner and outer alignment. However, see Hubinger (2021) for a criticism of this view.

[5] Different authors use different names to refer to the advanced, future AIs which are supposed to be an existential risk, including 'superintelligence' (Bostrom, 2014), 'Process for Automating Scientific and Technological Advancement' (PASTA) (Karnofsky, 2021) or 'APS (Advanced, Planning, Strategically aware) model' (Carlsmith, 2022). While their characterizations of these models differ slightly, I will slide over these distinctions here, as they do not affect the main argument. I am agnostic on which of these labels picks out the most relevant class of models.

Space is precious, so I cannot argue satisfactorily for a specific forecast on if and when AGI will arise here.[6] I will briefly mention three reasons to take the possibility of AGI seriously. First, there has been massive progress in the field of AI. Not only that, but breakthroughs like GPT-3 (Floridi & Chiriatti, 2020), GPT-4 (Bubeck et al., 2023; OpenAI, 2023) and ChatGPT (Shanahan, 2023) as well as AlphaFold (Jumper et al., 2021) have generated optimism for further breakthroughs which stimulates more investment. While we may still be far from AGI, there is no compelling theoretical reason to expect that AGI is impossible. Since scaling current models, i.e. increasing model size and the amount of training data, has led to the emergence of qualitatively new capabilities like few-shot learning (Brown et al., 2020; Wei et al., 2022), it cannot even be confidently ruled out that further scaling might eventually lead to AGI. All this is highly uncertain, but uncertainty cuts both ways. In itself, high uncertainty is not a reason to think that the advent of AGI this century is particularly unlikely.

Second, while the forecasts of AI progress we do have are limited, they tend to suggest that AGI this century is not unlikely. A survey of expert AI researchers on the question, conducted in 2016, finds a mean forecast of the people surveyed of a 50% chance that "unaided machines can accomplish every task better and more cheaply than human workers" in 2061 (Grace et al., 2018). In response to the same question, the yet-to-be published follow-up survey from 2022 finds that the mean prediction of 50% probability has moved to 2059 (Grace, 2022). Taken at face value, this suggests that many AI researchers expect there to be AGI within the next decades.[7] Also relevant is that the median forecast that the long-run effect of advanced AI on humanity will be "extremely bad (e.g., human extinction)" is 5% and that two questions explicitly about human extinction get a median forecast of 5% and 10% probability, respectively (Grace, 2022).

Third, even if threats from AGI are considered to be unlikely and limited to the distant future, distant low-probability risks can be very significant if the stakes are sufficiently high. For instance, it seems like it can be rational to buy insurance against one's house burning down, even if such an event is very unlikely. While the exact assessment of how bad an existential catastrophe would be depends on complex ethical issues (Greaves & MacAskill, 2021), for instance in population axiology (Arrhenius et al., 2022; Greaves, 2017; MacAskill, 2022), it seems clear that the stakes are extremely high in any case.

Thus, if there is a plausible argument that AGI constitutes an existential risk, even a low chance of AGI is worth taking very seriously. So, we have to ask: Why would AGI be an existential risk? Notice first that misaligned AGI poses an unprecedented kind of technological risk. With other technologies, say airplanes or bombs, risks of

[6] For an extended argument that AI might lead to the permanent disempowerment of humanity this century, see Dung (2023b).

[7] This picture is complicated by the fact that the answers heavily depend on the framing of these questions. In the earlier survey, the mean forecast is that, only for the year 2138, there is a 50% probability that "for any occupation, machines could be built to carry out the task better and more cheaply than human workers". This seems inconsistent with the forecast mentioned earlier by the same group. Note also that the forecasts are all conditional on "human scientific activity continu[ing] without major negative disruption."

harm stem from *accidents* or *misuse* (Vold & Harris, 2021). By contrast, misaligned AIs pursue harmful goals. They might intentionally harm humans and try to thwart our plans of containing them. Accidents are usually constrained in their temporal and spatial scope. Misuse requires that other humans have to be in the loop. Misalignment risk is different. Misaligned AGIs may permanently work against human interests. If they are more powerful than humanity and their goals conflict with human flourishing or even continuing human existence, then they will only stop if humanity is no longer a factor.

Why might one think that the goals of an AGI could require human disempowerment or extinction? The basic argument is based on two claims: the orthogonality thesis and the instrumental convergence thesis (Bostrom, 2014). According to the orthogonality thesis, intelligence and goals are (almost completely) independent. That is, except for certain special cases, any level of intelligence can co-occur with any set of goals (Bostrom, 2014, p. 107). The orthogonality thesis employs an instrumental understanding of intelligence according to which intelligence is the skill to attain particular final goals (given certain means) but places no constraints on what those goals are.

The instrumental convergence thesis states that there are certain goals which are instrumentally useful for a wide range of final goals and a wide range of situations (Bostrom, 2014, p. 109). Among these goals are self-preservation and the accumulation of power and resources. For if you get destroyed, you cannot work towards achieving your final goal anymore, and if you acquire power and resources, you will be more effective at achieving your final goal. Hence, typically, you will increase the probability that your final goal will be satisfied by preserving yourself and accumulating power. Thus, for a wide range of goals, an AGI would have an incentive to acquire power and to resist being shut down.

I will not evaluate these two theses here.[8] Let us just make explicit how they conspire to motivate worries about risks from AGI. If the orthogonality thesis is true, then we cannot be sure that an AGI will have reasonable, or human-aligned, goals. If the instrumental convergence thesis is true, we have positive reasons to think that AGI will be power-seeking. Thus, those two claims explain why AGI might aim to disempower, and perhaps annihilate, humanity. If we solve the alignment problem, then we avert this risk.

While worries about human disempowerment mainly focus on future AGI, I will argue that present AI systems already bring about instances of the alignment problem. In the next section, I will describe two of these examples. Subsequently, I will draw lessons and examine what they tell us about the prospects for aligning future AGI systems.

One word about the scope of this paper: My intention is to examine features of the alignment problem which go beyond a specific kind of AI architecture or training paradigm. However, this paper aims to inform the discussion of the alignment problem via consideration of current state of the art systems. Since the state of the art is dominated by deep learning, the discussion will concentrate on systems which fall

[8] See Müller and Cannon (2022) and Häggström (2021) for a recent controversy on the orthogonality thesis. Petersen (2020) and Railton (2020) for further relevant discussions.

within the deep learning umbrella. To what extent the claims considered here generalize beyond deep learning systems, will remain as a question for future research.

3 Present alignment problems

3.1 Large-language models

The first example are large-language models (LLMs). For the sake of concreteness and because ChatGPT of OpenAI has received the most attention and is among the most impressive current technologies using a LLM, I will focus on it. ChatGPT is a Transformer neural network which outputs written language in response to a written prompt. This way, it is able to complete texts, answer questions and execute instructions. Since many tasks can essentially be reduced to producing appropriate linguistic outputs and ChatGPT is not limited to natural, assertoric language, it can help with problems in a wide variety of domains. For instance, it can do poetry, program, write cooking recipes and (less reliably) perform math. Since ChatGPT is very recent at time of writing, there is not much published literature on how it performs with respect to various benchmark tests. However, GPT3 – its predecessor – achieved very good results (Brown et al., 2020) and – informally – it seems obvious that ChatGPT displays remarkably good performance (Gozalo-Brizuela & Garrido-Merchan, 2023; Shanahan, 2023).

I will not talk about what makes Transformer models distinctive.[9] Instead, I will turn to how ChatGPT was trained. It was subject to three different training regimes: pre-training via text prediction, supervised learning and reinforcement learning from human feedback. In the pre-training phase, the system was given an extraordinarily large text corpus and trained to predict the next token (e.g. a word) in a sequence. Via gradient descent learning (Russell & Norvig, 2020), its weights adjust over time to make its predictions more accurate. Thus, in essence, the generative model comes to embody the "statistical distribution of tokens in the vast public corpus of human-generated text" (Shanahan, 2023).

The subsequent training serves to make the outputs of the system more "helpful, honest, and harmless" (Bai et al., 2022; Ouyang et al., 2022). I will focus on the later training stage which uses reinforcement learning from human feedback (RLHF). In RLHF, human raters rank different text completions of the system, and the system learns to find completions which rank highly. Since the human raters are instructed to evaluate completions more positively if they are accurate, helpful, inoffensive etc., the system is trained to produce completions which have these features. Again, over time the weights of the model adjust such that outputs which are assessed positively become more likely.

RLHF can be understood as a response to a problem that beset GPT-3 (which was not trained via RLHF). This model tends to produce "problematic" outputs which mostly fall into two categories: First, so-called "hallucinations". These are output statements which are confidently presented and superficially plausible but false. Sec-

[9] For a good introduction, see: https://e2eml.school/transformers.html.

ond, outputs can be ethically problematic by being, for instance, insulting, discriminating (e.g. racist or sexist), information hazards (e.g., instructions for building a bomb), or incitements to violence. Since the first training phase (which it shares with GPT-3) taught ChatGPT merely to predict text and since human text is frequently incorrect or ethically questionable, it is easy to see why incorrect and ethically questionable text completions result. By contrast, RLHF punishes the system if it chooses responses which human raters deem incorrect or otherwise problematic. Thus, it can be expected to decrease the incidence of undesirable outputs.

While RLHF has led to progress, both hallucinations and ethically questionable responses nevertheless occur in ChatGPT. In a particularly absurd case, ChatGPT insists that 47 is a larger number than 64. With the version of Chat-GPT as of time of writing (January 2023), examples like this can be multiplied at will. Moreover, people have found many ways to make ChatGPT produce problematic outputs. For instance, if ChatGPT is encouraged to engage in pretense, it is often happy to make racist, antisemitic and homophobic statements or to provide instructions for building a Molotov cocktail (Mowshowitz, 2022). Thus, so far OpenAI has failed to robustly teach ChatGPT to refrain from giving such problematic outputs.

I will argue next section that the tendency of ChatGPT to hallucinate and cross ethical boundaries is an alignment problem. However, let us first look at another case.

3.2 Game-playing agents

I hold that reward hacking in systems trained to perform well in games is another alignment problem. Examples abound (Buckner, 2021; Christian, 2020),[10] but for concreteness I will single out the agent trained by OpenAI in the game CoastRunners (OpenAI, 2016).[11] If humans play this game, their aim is typically to win the boat race, i.e., to finish ahead of other players. As is standard for game-playing agents, the system in question was trained via reinforcement learning (RL).

Conveniently, the game score can serve as the reward function which the RL agent is trained to optimize. In the game, the player does not directly receive a higher score for coming closer to or reaching the finish line, but gains points for hitting targets laid out along the route. The assumption was that in maximizing the game score the RL agent would also, to the best of its abilities, compete in the race. However, it turned out that the agent was able to find an exploit. At a particular location of the map, the RL agent was able to continue to move in a circle in a particular way (involving repeatedly crashing into another boat and a wall) which led to a higher game score than actually trying to win the race. The result: The RL agent achieved a higher score

[10] Relatedly, see also this database of 'specification gaming' cases: https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7E-a5eWtvsWzuxo8bjOxCG84dAg/pubhtml.

[11] Calling such systems 'agents' is standard in the machine learning literature. By adopting this convention, I do not mean to imply that all, or some, RL systems are agents in a significant, philosophical sense (Butlin, 2023; Butlin et al., 2023; Glock, 2019).

than is possible by playing the game in the intended way, but failed miserably at the informal goal of actually winning the race (OpenAI, 2016).[12]

In other words, the RL agent engaged in reward hacking. This is a phenomenon where optimizing a proxy reward function, e.g. the game score, leads to poor performance in the task the agent was intended to optimize for, e.g. the race (Skalse et al., 2022).[13] There are many other examples of reward hacking in game-playing AI (Baker et al., 2020; Christiano et al., 2017; Ibarz et al., 2018; Pan et al., 2022; Toromanoff et al., 2019). As argued next section, reward hacking is a manifestation of misalignment.

4 High-level lessons from current cases of misalignment

4.1 Are language model misbehavior and reward hacking alignment failures?

In this subsection, I will argue that both cases presented last section are cases of misalignment. Next subsection, I will draw some general lessons from these cases.

An AI system is misaligned if it does not pursue the goals that its designers want it to pursue. OpenAI – its designers – don't want ChatGPT to confidently give false answers, let alone express racist or other prejudices. Similarly, OpenAI wanted their RL agent to win boat races in CoastRunners, not to endlessly drive in circles. However, we have seen earlier that the mere fact that an AI system produces undesirable outputs is not sufficient evidence that the AI system is misaligned. In general, the alternative possibility is that the system has the goals its designer wants it to have, but that it lacks the capability to reliably achieve those goals. In other words, if we can rule out that a lack of capability prevents the system from producing desirable outputs, then we can conclude that the system is misaligned.

In the two cases outlined last section, a lack of capability is not a plausible explanation for why the system's behavior is undesirable. I will present three reasons to support this assessment. First, in both cases, achieving desirable outputs seems no more difficult than the performance the systems actually display. Intuitively, it is not hard to complete texts in a manner which does not express racial prejudice. It is

[12] Another possible example of misaligned RL algorithms with more pernicious societal consequences are recommender models used by social media platforms. Arguably, the intention underlying recommender algorithms is to present content to users which fits their preferences especially well to increase their subjective wellbeing (one might, however, argue that the intentions of the designers are more malicious to begin with, which would make it questionable as a case of alignment). However, if the algorithm is trained via RL to maximize an easier measurable proxy like the probability that the human user clicks on the item shown or watch time, there are possible scenarios for how the algorithm might maximize rewards without achieving the intended aim. Particularly, it has been hypothesized that the algorithm might over time behave in a way which shapes the preferences of users such that their preferences become easier to predict and satisfy (Burr et al., 2018). For instance, the recommender models might feed users contents which cause them to develop more extreme and one-sided political views (Ribeiro et al., 2020) to cause the users to become more predictable in their engagement with political content.

[13] As this definition indicates, I use the term 'reward hacking' rather broadly. In my understanding, it is not limited to cases where the AI system directly interferes with its own reward signal ('wireheading'). A similar phenomenon is described by 'Goodhart's law'.

just that ChatGPT does not aim for that, at least not consistently. Even more clearly, participating in the intended way in the boat race is easier than finding the exact configuration of movements which maximizes the game score while not moving forward in the race, as the RL agent did. Evidence that the tasks solved by the AI systems are harder than the ones which they don't solve is that humans can refrain from offensive speech and participate in the virtual boat race, but often cannot produce texts as fluently or achieve a high game score as effectively as the systems.

Second, in cases like the two we are focusing on, sometimes systems which are generally less capable produce more desirable behavior. With LLMs, this is because more capable models, which tend to be larger and trained with more data, can pick up on more subtle statistical relations in the training corpus. For instance, a more capable LLM might notice that certain questions are often followed by answers which allude to subtle racist stereotypes while a less capable LLM might not track this relationship. In this case, the more powerful LLM might provide the worse, racist, answer precisely because it contains more knowledge about statistical relationships within public language. Analogously, a RL agent needs a certain level of capability to reward hack effectively. While a powerful RL agent is able to find a route which optimizes rewards while neglecting the race, an inferior agent might not be able to find a way to exploit rewards and thus actually participate in the race as intended. In cases where more capable systems behave in a less desirable manner – and do this *because* they are more capable – the cause of the undesirable behavior must be misalignment, not lack of capability.

Third, and specific to language models, trying different prompts can significantly improve performance. For instance, it has been shown that performance of GPT-3 on reasoning tasks and math improves when the input is prefaced by the phrase "let's think this through step by step" (Suzgun et al., 2022) or that ChatGPT's texts can be improved by telling it that it is an excellent writer or an expert on the subject matter at hand. The fact that minor changes to the input prompt improve performance shows that the system possessed the knowledge necessary to perform the intended behavior all along. Thus, the initial failure to elicit the desired behavior was not caused by a lack of capability of the system, but by the system not trying to achieve the desired behavior. In other words, it was an alignment failure.[14]

One might object that ChatGPT is aligned because there is a sufficient degree of correspondence between what it tries to do and the goals of its designers. For instance, one might say that ChatGPT is trying to provide outputs which human raters rate as desirable and that this is what its designers want (after all, this is what they train it to do). However, the ultimate goal of its designers is to have ChatGPT only provide harmless, helpful, and honest texts. RLHF is merely the means they choose to make ChatGPT as aligned as possible.

ChatGPT is misaligned in that its goal can likely be construed as a blend of *predicting the next token in a text sequence* and *maximizing positive human feedback*.

[14] A fourth argument that LLMs exhibit alignment, not just capability failures, is provided by studies along the lines of Burns et al. (2022), Halawi et al. (2023) and Belrose et al. (2023). These studies suggest that language models have internal representations which correlate better with true answers to questions than their actual outputs. Hence, it seems that the systems

However, this does not perfectly correlate with *producing helpful, honest and harmless text* which is what its designers want ChatGPT to do (Casper et al., 2023). This shows that misalignment admits of degrees. If ChatGPT were to, e.g., maximize negative human feedback it would be more misaligned.

To summarize, there is strong reason to think that ChatGPT and the RL agent playing CoastRunners fail to elicit desired behavior not because of a lack of capacity, but because they don't pursue the goals their designers want them to pursue. What can we learn from these two examples?

4.2 Properties of misalignment

We have looked at two quite different instances of AI alignment problems: based on LLMs and game-playing RL agents. Obviously, such a methodology is not exhaustive. There are not only many more instances of alignment problems, there might also be other, very different kinds, for instance based on other training methods or systems which are not of the deep learning variety. Thus, future research may reveal that the alignment problem has features not shared by these cases, or that some features present in these cases are not typical for alignment problems. Nevertheless, it is worth exploring what our two cases suggest about the alignment problem. This sub-section, I will draw out five lessons from these case studies. Next section, I will discuss what these lessons tell us about existential risk from misaligned AI.

First, those cases suggest that specific forms of misalignment can be antecedently hard to predict and sometimes hard to detect. We can distinguish two ways in which misalignment can be surprising: It can be unexpected that a particular AI system is misaligned at all or the specific way in which a system is misaligned can be surprising. Let us call this 'general misalignment surprise' and 'specific misalignment surprise', respectively. With the RL agent playing CoastRunners, it seems like both forms of misalignment surprise are present. Its designers were surprised that it found a way at all to optimize rewards without actually trying to win the race (OpenAI, 2016). As a corollary, they did not anticipate the specific route the agent chose to reward hack.

With ChatGPT, it is plausible that its designers knew before its release that it frequently produces false outputs. Less clear is to what extent its designers knew about the degree to which it could be manipulated to produce ethically problematic texts (Mowshowitz, 2022). Since previous versions (GPT-3) had similar deficits, it was clear that misalignment risks are specifically risks from hallucination and from ethically problematic completions. However, very specific risks, i.e. which kinds of prompts might lead to which kinds of problematic outputs, were not predicted. Otherwise, it would have been possible to fix them in advance. Thus, with respect to ChatGPT, the picture is mixed. Due to ample experience with its predecessor GPT-3, the designers knew about relevant risks, albeit not on a very fine-grained level. Moreover, it probably was not clear in advance to what extent ChatGPT managed to overcome these risks. Thus, overall, while some forms of misalignment can be predicted, this seems to be difficult, nonetheless.

For the CoastRunners agent, it is trivial to detect misalignment. Since the agent at some point does not move, or try to move in some observable manner, closer to the

finish line, it is clearly not trying to win the race, i.e. it is misaligned. However, one could imagine more subtle forms of misalignment which could have occurred instead. If the targets increasing the game score were distributed differently on the route, the optimal strategy for maximizing the game score might have been one which involves participating in and finishing the race, but not in the fastest way possible. In this case, misalignment might have been harder to detect, since lack of capabilities would have been another superficially plausible explanation for sub-optimal play.

Extensive experience with and systematic probing of ChatGPT make it clear that its outputs are frequently blatant falsehoods. However, while casually conversing with it, it is effective at creating the appearance that it answers questions responsibly and truthfully. Thus, casual observers might miss ChatGPT's misaligned behavior, while its designers notice it.

The second general feature suggested by our examples is that misalignment can be hard to remedy. ChatGPT was trained via RLHF precisely to prevent hallucinations and ethically problematic speech. There are strong incentives to succeed: if ChatGPT would tell the truth more reliably, it would have more commercial uses as a source of information. Moreover, some of the speech ChatGPT produces might harm OpenAI's public reputation. Nevertheless, the issues persist and OpenAI currently endeavors to find a training regime which reduces them as much as possible.

With RL agents, specifying an appropriate reward function, which leads to the intended behavior, is notoriously hard (Langosco et al., 2023; Pan et al., 2022; Skalse et al., 2022). In videogames, it gets harder when the game is more complex and itself only provides sparse rewards. If the RL agent is supposed to navigate real-world environments with a large set of competing desiderata, the challenge magnifies. Thus, it is frequently non-trivial to provide rewards such that reward hacking won't occur.

Third, misalignment does not depend on a specific training paradigm. The agent in CoastRunners was trained via RL, GPT-3 only via unsupervised text prediction and ChatGPT in addition with RLHF. Three different training regimes, but each can lead to misalignment. Moreover, in the CoastRunners case, we have an avatar which moves within a virtual environment. GPT-3 and ChatGPT are restricted to linguistic outputs. However, no matter whether an AI system controls an avatar which is situated in an environment or not, misalignment can occur.

While I have not adduced any relevant cases in this paper, there is every reason to think that misalignment is not limited to deep learning. For it is a very general property: a mismatch between the intention of the designers and the goals of the system. At least in principle, this mismatch can appear whenever a system has goals in the minimal sense relied on here.

Fourth, there is a connection between the practical usefulness of an AI system, and alignment. System which are not aligned, or exhibit significant alignment failures, tend to be less useful in practical contexts. Consequently, there is less incentive to employ them in many domains. ChatGPT's problems with hallucinations and ethically problematic speech are of course an example of this.[15]

[15] To make up another example: There will be strong reasons to not use an autonomous vehicle if this vehicle assigns a lot less value to the safety of the passengers than the passengers themselves (and, by extension, the designers of the vehicle).

Fifth, and partly as inference from the previous points, it seems that some degree of misalignment is the *default* outcome when developing an AI system, at least when trained using ML. To prevent misaligned behavior, there has to be a close, or perfect, correspondence between the intentions of the designers and the goals of the system. Usually, a lot of effort is needed to bring the system to have the right goals. It is common that reward functions are initially mis-specified and that a lot of experimentation is needed to modify them such that they encourage learning the desired behavior. As ChatGPT shows, supervised learning and learning from human feedback are also not always sufficient to prevent misalignment. With these methods as well, learning from numerous specific failure modes, changing the training setup accordingly and iterating this procedure many times may be necessary to create alignment.

In summary, misalignment is hard to predict, to remedy and sometimes also hard to detect. It tends to impede the practical usefulness of systems. Moreover, it is a risk which applies to many, or all, kinds of AI systems and is the default outcome of developing new systems which can, if at all, often only be prevented after a long process of trial-and-error. In the next section, we explore what these lessons can tell us about existential risks from misaligned AI. This involves discussing whether and how the risks and problems that we have mentioned plausibly transfer to more advanced systems and eventually to AGI.

5 Alignment in AGI

Misaligned AI already causes harm in the present, for instance, when people decide to trust the information ChatGPT provides. However, as mentioned in Sect. 1, misalignment poses especially grave risks in the case of AGI. The more powerful an AI system is, the worse it is if it optimizes for goals that we do not view as desirable. For this reason, it is important to estimate how high the chance of misaligned AGI is. I will approach this question by exploring which implications the features of misalignment we have discovered in the previous section have for the prospects of AGI alignment.

Thus, in this section, I will first discuss reasons for why AGI misalignment risks may be relatively low. Subsequently, we will look at risks which are elevated in the case of AGI and at new types of risks which are specific to AGI.

5.1 Reasons why AGI misalignment risk might be low

Our previous discussion suggests two main reasons why risk of AGI misalignment might be low, compared to misalignment risks from current AI. First, we noted that there is a positive correlation between alignment and usefulness. All other things being equal, less aligned systems are less useful, thus there is less incentive to deploy these systems. If we assume, as is plausible, that the dangers or unleashing misaligned AGI are vastly higher than the risks of deploying misaligned contemporary systems, then there will also be strong incentives against using such an AGI. By definition, an AGI exceeds human intelligence in some or many important domains, and thus makes it economically redundant in these domains. Therefore, it seems clear

that there will also be strong opposing commercial and political (e.g., from military competition) incentives to use such systems. However, one might hope that the combination of impaired usefulness caused by misalignment and increased danger, if misaligned AGI is deployed, reduces the probability of deployment.

Second, increased capability alleviates some aspects of the alignment problem. Take ChatGPT. Part of the alignment problem for ChatGPT is that it does not generalize correctly from the human feedback it has received during training. For its false or ethically problematic outputs would be evaluated negatively by human raters if they occurred during training. Thus, during training with RLHF, ChatGPT actually does not manage to robustly optimize for human feedback. Otherwise, many of its hallucinations and ethically problematic outputs would not occur.

Put another way: The case of ChatGPT is an alignment, not a capability, problem because it is not a lack of capability which causes ChatGPT to produce undesirable outputs. The cause are the tendencies it has inherited from its earlier pre-training which was purely about text prediction. Nevertheless, a further increase in capability creates the chance that the system might become more effective in optimizing for positive human feedback and generalize more robustly from previous feedback. This would reduce its undesirable behavior. Thus, in a sense, it would reduce its misalignment.[16]

Hence, there are at least some cases where advances in the capability of systems make alignment strategies, like RLHF, in some respects more effective. Thus, some alignment strategies might work better, in some respects, with more capable systems. However, we will now look at ways in which our review of features of misalignment should make us worry particularly about AGI alignment. To begin with, we look at features of AGI which seem to increase the probability of misalignment.

5.2 Misalignment risks which increase with capability advances

We noted that misalignment can be hard to predict and detect. The more complex the system's processing and the more sophisticated its outputs, the more severe this problem becomes. With respect to deep neural networks at least, when the size of the system increases it becomes more and more hopeless to predict which dispositions for producing certain outputs it will acquire and why. Moreover, detecting misalignment can also become more challenging. For instance, a language model AGI may be instructed to write papers producing novel scientific insights. If the AGI system is superior to us in scientific reasoning, it may be hard for us to evaluate whether the AGI system aims to increase scientific knowledge, as we intended, or whether it hallucinates. Similarly, when an advanced game-playing AI system performs on a

[16] This may seem contradictory: How can an increase in system capability make the system more aligned? Shouldn't we, in this case, say that the model lacked capability rather than that it was misaligned? In some cases, alignment and capability cannot be separated cleanly. If a language model lacks the capability to generalize information about what kinds of outputs humans would reward, this can lead the system to produce outputs more in line with the goal of text prediction acquired during pre-training. Thus, we might still call this an alignment failure *in some sense* because it stems from the pretraining goal of text prediction being misaligned with the human goal of making systems harmless, helpful and honest even though it can be alleviated with more capability.

superhuman level in a complex environment comprising multiple demands, or even in real-world environments, it is hard for us to say whether it actually optimizes the goals we intend or whether it optimizes a proxy which imperfectly correlates with them. Consequently, we should expect misalignment of AGI to be harder to predict and detect. This is worrisome, given that misalignment of AGI could cause catastrophic outcomes.

Moreover, the CoastRunners case illustrates the threat of reward hacking. The reward signals a RL system is trained on serve as a proxy for the intended, "true" goal. This is typically necessary because the intended goal can be hard to specify, measure and optimize for. If the behavior the system is trained for – in RL via reward and punishment signals – only imperfectly correlates with our true goals, then upon deployment the system could find ways of maximizing its reward which strongly deviate from the intended goal. As we briefly discussed earlier, when the capabilities of the system increase it might find new and subtle ways to reward hack. The more capable the system is, the better its skills in exploiting subtle dissociations between the reward signal and the intended goal become (Taylor et al., 2020).

This should concern us with respect to AGI. Even now, RL agents can find ways to reward hack which humans did not think of. If they are better than us in important domains, we should likewise expect that they might find ways to optimize proxy rewards which we did not anticipate. In the limit, when thinking about superintelligent AI which greatly exceeds human capabilities, we should expect that every possible way to reward hack will actually be found.

Moreover, strategically important real-world domains are typically characterized by a multitude of competing demands which need to be balanced thoughtfully. For instance, in the military context, an important goal might be to win a war. However, this goal might often be in tension with other relevant goals such as respecting human rights, minimizing the death of one's own soldiers, avoiding civilian casualties, averting the destruction of the environment and so on. Such a diverse set of goals which involves many trade-offs is not easy to specify. Moreover, goals such as avoiding human rights violations are not easy to operationalize, measure and directly optimize for. Thus, it is plausible that the use of proxies might be necessary, at least if AGI is trained via RL.[17] If AGI systems can be expected to be excellent reward hackers, this increases the chance of misalignment.

In ChatGPT, or in RLHF in general, reward hacking takes a peculiar form. In general, the system is trained to optimize for human feedback. Thus, the situations which should worry us are ones were attaining the most positive human feedback can come apart from exhibiting the best behavior, in light of our goals. For instance, human rat-

[17] The considerations regarding reward hacking do not translate straightforwardly to non-RL systems. However, there are two reasons for their general relevance. First, at present, in all interesting domains the best way of training systems to exhibit intelligent behavior in pursuit of a goal involves RL. Second, one may at least argue that an analogous concern exists for other training paradigms. If a system trained via supervised learning overfits particular training data, then it might sometimes behave outside of the training distribution in a manner which is very competent but seems to aim at the wrong goal. For instance, an image recognition system might misclassify all pictures but in a systematic manner and based on subtle indications and impressive perceptual recognition capabilities. In a sense, this might be classified as similar to reward hacking and a case of 'misalignment'. In this case, the relevant notions are more intuitive and informal, however, because outside of RL there is no clearly defined notion of the reward for a system.

ers might err on whether a given statement is correct. If so, they will reward ChatGPT for an incorrect statement and punish it for a correct statement. Hence, ChatGPT is more specifically trained to say what humans think is correct, rather than to say what actually is correct (Casper et al., 2023).

While this form of misalignment might not matter much in current ChatGPT because the current system is even struggling to reliably produce output which humans evaluate as correct, it may be important in future, more advanced systems. Moreover, AGIs with sufficiently high control over the information environment of humans might find better ways to illicitly gain positive feedback. For instance, they may manipulate humans to induce false beliefs which the system can subsequently reproduce to gain positive reward. Thus, RLHF allows for a particular form of reward hacking.

To summarize, with respect to AGI misalignment, both our epistemic situation and our means of prevention are severely constrained. Misalignment should be expected to be even harder to detect and predict than in current systems, and AGI's capacity to reward hack will increase as other capabilities increase. In the next subsection, I will bring up two misalignment risks which arise only, or in a qualitatively new form, with AGI.

5.3 Risks of misalignment specific to AGI

In this subsection, I will argue that certain misalignment risks only occur when we consider very powerful, i.e. AGI, systems. We noted that, in current systems, it is frequently challenging to remedy misalignment. With AGI, a new threat arises. If an AGI has misaligned goals, we might not be able to improve upon its goals, or shut off the system, at all.

This scenario is suggested by two observations: First, in virtue of the thesis of instrumental convergence, for a wide range of final goals the AGI has self-preservation and the stability of its final goals as an instrumental goal. For if the AGI would be shut down or its goals modified, this would typically make it less likely that its original goals would be achieved. In the usual case, when a system is misaligned, we try to adjust its goals or shut it down. However, as just shown, an AGI would have an incentive to resist us when we try to do this. Second, because, by definition, AGIs exceed human capability in some domains and rival it in others, it is not clear whether we could shut down or change an AGI if it resists us. Thus, AGI misalignment might be permanent, and thus lead to permanent human disempowerment.

This scenario is even more concerning in the light of two previous observations. First, some degree of misalignment seems to be the default outcome when training systems via deep learning methods. Without dedicated effort – and often not even then – the goals that systems acquire during training typically do not perfectly correspond to the goals their designers want them to have. Also, the threat of human disempowerment prevents us from achieving alignment by experimenting on and using trial-and-error strategies with misaligned systems. Second, misalignment – especially in complex systems engaging in sophisticated tasks – is hard to predict in advance. If the baseline probability of misalignment is high, we are bad at foreseeing specific instances of misalignment and misalignment of sufficiently capable AGI has

a high chance of leading to permanent human disempowerment, then the risk that an AGI, once deployed, turns out to be misaligned is very concerning.

There are two further factors which raise the probability of misalignment in AGI. The first we already discussed: the increased tendency of more capable systems to reward hack. The second is new: once systems are sufficiently advanced, they might develop "situational awareness" (Cotra, 2021, 2022). That is, they might understand the situation they find themselves in, i.e. that they are a deep learning system, how they are designed and trained, the intentions and beliefs of their designers and so on.[18]

It is plausible that systems will eventually develop situational awareness because, for a wide range of training regimes, situational awareness will be useful for enhancing training performance. In RL and RLHF, situational awareness is beneficial for maximizing rewards. For instance, knowledge about one's own architecture and training can be used to identify one's own weaknesses. This knowledge can be used to choose appropriate strategies and to improve in games. Situational awareness can also enable reward hacking. For example, it allows a language model to recognize that certain false things are believed by humans, and consequently infer that certain false answers will be rewarded. Since systems will have incentives and ample opportunity (Cotra, 2022) to develop situational awareness, systems which possess very high general capability will likely do so.

Situational awareness poses an unprecedented risk, namely that systems might behave in a seemingly aligned manner until they are sufficiently capable to disempower humanity. In the absence of specific countermeasures,[19] this behavior can be expected if the system systematically follows its incentives, i.e., maximizes its reward or optimizes for a misaligned goal. This is because a sufficiently intelligent system equipped with situational awareness can reason that, in light of its ultimate goal, it is best to appear aligned to humans, if they have the power to shut it off, change its goal or provide negative rewards. By the same token, a system should reason that it is conducive to its goal or to reward maximization to work in secret towards accumulating power and to eventually overthrow humanity. For human activities will conflict with reward maximization or the pursuit of some misaligned goal. Thus, the system has an instrumental reason to disempower humanity once it is capable to do so.

So, to recap, I have argued that, given plausible assumptions, AGIs will likely have the means and the instrumental reasons to engage in deceptive alignment. That is, they might create the appearance of being aligned with human goals, before trying to disempower humanity when the opportunity arises.

This is worrying. For one, it obviously makes it harder to detect misalignment and to respond appropriately. If cases of misalignment are detected and punished, the system might just learn to hide its misaligned behavior and intentions more effectively. If this dynamic obtains, the system optimizes against our ability to understand it. Moreover, deceptive alignment might cause people to misperceive the overall balance of

[18] A reviewer proposes that AGI might be easier to align than a less capable system because it might have the ability to recognize inappropriate final goals by itself and to consequently revise them. After all, this capacity is important for human general intelligence. This point is echoed by Müller and Cannon (2022). While I cannot settle this issue here, I broadly accept Häggström's (2021) reply.

[19] For a discussion of some possible countermeasures and their limitations, see Taylor et al. (2020).

reasons in favor and against deploying AGI. If an AGI system is deceptively aligned, decision-makers might be very confident that it is aligned and thus neglect misalignment risks. Moreover, the system itself aims to be deployed, so it would behave in ways which are, or seem, very useful to humans. Thus, the incentives for deploying AGI might be very strong while the risks appear implausible.[20] Thus, decision-makers might unwisely favor deploying an AGI. Given the arguments presented here, one should take concerns that this might cause an existential catastrophe seriously.

6 Conclusion

Let us summarize. At the beginning of this paper, I have argued that we can fruitfully understand the alignment problem as the problem of designing AI systems such that their goals correspond to what their designers want them to be. To motivate the importance of this problem, I have sketched the basic case for why misaligned AI might be an existential risk. Then, I have analyzed instances of alignment failures in current large-language models and game-playing agents. Those cases suggest that alignment problems tend to have particular sorts of features. Given the features of alignment we have encountered in current systems and the peculiarities of envisioned more advanced AI systems, we can expect that aligning advanced AI systems is a hard technical challenge. In light of challenges which are exacerbated or wholly new when considering more advanced AI systems, there is a real risk of alignment failure.

I have not sufficiently supported every step of the argument that misaligned AI is an existential risk. For instance, an exhaustive discussion should consider the supposition that sufficiently intelligent AIs would automatically tend to adopt reasonable values (Petersen, 2017), that AGI is impossible or that humanity would decide to just not build an AGI. Moreover, I have mostly set aside misalignment risks arising from goal misgeneralization where the system fails to generalize a correctly specified goal outside of its training distribution (Langosco et al., 2023). Nevertheless, this paper provides substantive reasons for accepting the conditional claim that, assuming we develop AGI in the next decades, aligning it might be very difficult. Since the potential damage from misaligned AGI is very large, we should not take remaining uncertainty as an excuse to ignore this risk. After all, uncertain risks can be very important, worth thinking about and worth preparing for.

Acknowledgements I thank three anonymous reviewers for helpful comments.

Funding This research was funded by the German ministry for education and research (BMBF) in the context of the K3I-Cycling project. Project number: 033KI216
Open Access funding enabled and organized by Projekt DEAL.

Availability of data and material Not applicable.

[20] The overall strength of the incentives in favor of deploying AGI and the willingness to take risks plausibly depend on many factors, including geopolitical ones (Armstrong et al., 2016; Cave & ÓhÉigeartaigh, 2018). It is important to note that risks from AGI derive also from training the system, not just from deployment.

Code availability Not applicable.

Declarations

Competing interests No potential conflict of interest.

References

Armstrong, S., Bostrom, N., & Shulman, C. (2016). Racing to the precipice: A model of artificial intelligence development. *AI & SOCIETY*, *31*(2), 201–206. https://doi.org/10.1007/s00146-015-0590-y.

Arrhenius, G., Bykvist, K., Campbell, T., & Finneron-Burns, E. (Eds.). (2022). *The Oxford Handbook of Population Ethics* (1st ed.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190907686.001.0001.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., & Kaplan, J. (2022). *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback* (arXiv:2204.05862). arXiv. https://doi.org/10.48550/arXiv.2204.05862.

Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2020). *Emergent Tool Use From Multi-Agent Autocurricula* (arXiv:1909.07528). arXiv. https://doi.org/10.48550/arXiv.1909.07528.

Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., & Steinhardt, J. (2023). *Eliciting Latent Predictions from Transformers with the Tuned Lens* (arXiv:2303.08112). arXiv. http://arxiv.org/abs/2303.08112.

Bostrom, N. (2013). Existential risk Prevention as Global Priority: Existential risk Prevention as Global Priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002.

Bostrom, N. (2014). *Superintelligence. Paths, Dangers, strategies*. Oxford University Press.

Bostrom, N., Dafoe, A., & Flynn, C. (2020). Public Policy and Superintelligent AI: A Vector Field Approach. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 292–326). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0011.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712.

Buckner, C. J. (2021). Black Boxes, or unflattering mirrors? Comparative Bias in the Science of Machine Behavior. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/714960.

Burns, C., Ye, H., Klein, D., & Steinhardt, J. (2022). *Discovering Latent Knowledge in Language Models Without Supervision* (arXiv:2212.03827). arXiv. https://doi.org/10.48550/arXiv.2212.03827.

Burr, C., Cristianini, N., & Ladyman, J. (2018). An analysis of the Interaction between Intelligent Software Agents and human users. *Minds and Machines*, *28*(4), 735–774. https://doi.org/10.1007/s11023-018-9479-0.

Butlin, P. (2023). Reinforcement learning and artificial agency. *Mind & Language*. mila.12458.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (arXiv:2308.08708). arXiv. https://doi.org/10.48550/arXiv.2308.08708.

Carlsmith, J. (2022). *Is Power-Seeking AI an Existential Risk?* (arXiv:2206.13353). arXiv. https://doi.org/10.48550/arXiv.2206.13353.

Carnap, R. (1950). *Logical foundations of probability*. Chicago University Press.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C. R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., & Hadfield-Menell, D. (2023). *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback* (arXiv:2307.15217). arXiv. https://doi.org/10.48550/arXiv.2307.15217.

Cave, S., & ÓhÉigeartaigh, S. S. (2018). An AI Race for Strategic Advantage: Rhetoric and Risks. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 36–40. https://doi.org/10.1145/3278721.3278780.

Center for AI Safety (2023). *Statement on AI Risk*. https://www.safe.ai/statement-on-ai-risk.

Christian, B. (2020). *The Alignment Problem: Machine learning and human values*. W. W. Norton & Co.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). *Deep reinforcement learning from human preferences* (arXiv:1706.03741). arXiv. https://doi.org/10.48550/arXiv.1706.03741.

Cotra, A. (2022). *Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover*. Lesswrong. https://www.lesswrong.com/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to.

Cotra, A. (2021, September 21). *Why AI alignment could be hard with modern deep learning*. Cold Takes. https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/.

Dennett, D. C. (1987). *The intentional stance*.. MIT Press.

Dennett, D. C. (1991). Real patterns. *Journal of Philosophy*, *88*(1), 27–51.

Dung, L. (2023a). How to deal with risks of AI suffering. *Inquiry*, 1–29.https://doi.org/10.1080/0020174X.2023.2238287

Dung, L. (2023b). *The argument for near-term human disempowerment through AI*. https://philpapers.org/rec/DUNTAF-3.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and Consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1.

Friederich, S. (2023). Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00268-7.

Gabriel, I. (2020). Artificial Intelligence, values, and Alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2.

Glock, H. J. (2019). Agency, intelligence and reasons in animals. *Philosophy*, *94*(4), 645–671. https://doi.org/10.1017/S0031819119000275.

Gozalo-Brizuela, R., & Garrido-Merchan, E. C. (2023). *ChatGPT is not all you need. A state of the art review of large generative AI models* (arXiv:2301.04655). arXiv. https://doi.org/10.48550/arXiv.2301.04655.

Grace, K. (2022, August 4). *What do ML researchers think about AI in 2022?* AI Impacts. https://aiimpacts.org/what-do-ml-researchers-think-about-ai-in-2022/.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). *When Will AI Exceed Human Performance? Evidence from AI Experts* (arXiv:1705.08807). arXiv. https://doi.org/10.48550/arXiv.1705.08807.

Greaves, H. (2017). Population Axiology. *Philosophy Compass*, *12*(11), https://doi.org/10.1111/phc3.12442.

Greaves, H., & MacAskill, W. (2021). *The case for strong longtermism*. https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf.

Häggström, O. (2021). *AI, orthogonality and the Muller-Cannon instrumental vs general intelligence distinction* (arXiv:2109.07911). arXiv. https://doi.org/10.48550/arXiv.2109.07911.

Halawi, D., Denain, J. S., & Steinhardt, J. (2023). *Overthinking the Truth: Understanding how Language Models Process False Demonstrations* (arXiv:2307.09476). arXiv. https://doi.org/10.48550/arXiv.2307.09476.

Hubinger, E. (2020). *An overview of 11 proposals for building safe advanced AI* (arXiv:2012.07532). arXiv. https://doi.org/10.48550/arXiv.2012.07532.

Hubinger, E. (2021). *How do we become confident in the safety of a machine learning system?*https://www.alignmentforum.org/posts/FDJnZt8Ks2djouQTZ/how-do-we-become-confident-in-the-safety-of-a-machine.

Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., & Amodei, D. (2018). *Reward learning from human preferences and demonstrations in Atari* (arXiv:1811.06521). arXiv. https://doi.org/10.48550/arXiv.1811.06521.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), https://doi.org/10.1038/s41586-021-03819-2. Article 7873.

Karnofsky, H. (2021, August 10). *Forecasting Transformative AI, Part 1: What Kind of AI?* Cold Takes. https://www.cold-takes.com/transformative-ai-timelines-part-1-of-4-what-kind-of-ai/.

Langosco, L., Koch, J., Sharkey, L., Pfau, J., Orseau, L., & Krueger, D. (2023). *Goal Misgeneralization in Deep Reinforcement Learning* (arXiv:2105.14111). arXiv. https://doi.org/10.48550/arXiv.2105.14111.

MacAskill, W. (2022). *What we owe the future*. Oneworld Publications.

Mowshowitz, Z. (2022, December 2). Jailbreaking ChatGPT on Release Day [Substack newsletter]. *Don't Worry About the Vase*. https://thezvi.substack.com/p/jailbreaking-the-chatgpt-on-release.

Müller, V. C., & Cannon, M. (2022). Existential risk from AI and orthogonality: Can we have it both ways? *Ratio*, *35*(1), 25–36. https://doi.org/10.1111/rati.12320.

Ngo, R., Chan, L., & Mindermann, S. (2022). *The alignment problem from a deep learning perspective* (arXiv:2209.00626). arXiv. http://arxiv.org/abs/2209.00626.

OpenAI (2016, December 22). *Faulty Reward Functions in the Wild*. https://openai.com/blog/faulty-reward-functions/.

OpenAI (2023). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774.

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hachette Books.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback* (arXiv:2203.02155). arXiv. https://doi.org/10.48550/arXiv.2203.02155.

Pan, A., Bhatia, K., & Steinhardt, J. (2022). *The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models* (arXiv:2201.03544). arXiv. https://doi.org/10.48550/arXiv.2201.03544.

Petersen, S. (2017). Superintelligence as Superethical. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2. 0: New Challenges in Philosophy, Law, and Society* (pp. 322–337). Oxford University Press. https://philarchive.org/rec/PETSAS-12.

Petersen, S. (2020). Machines learning values. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 413–436). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0015.

Railton, P. (2020). Ethical Learning, Natural and Artificial. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 45–78). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0002.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141. https://doi.org/10.1145/3351095.3372879.

Russell, S. (2019). *Human compatible: Artificial Intelligence and the Problem of Control*. Viking.

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Saad, B., & Bradley, A. (2022). Digital suffering: Why it's a problem and how to prevent it. *Inquiry : A Journal of Medical Care Organization, Provision and Financing*, *0*(0), 1–36. https://doi.org/10.1080/0020174X.2022.2144442.

Shanahan, M. (2023). *Talking About Large Language Models* (arXiv:2212.03551). arXiv. https://doi.org/10.48550/arXiv.2212.03551.

Skalse, J., Howe, N. H. R., Krasheninnikov, D., & Krueger, D. (2022). *Defining and Characterizing Reward Hacking* (arXiv:2209.13085). arXiv. https://doi.org/10.48550/arXiv.2209.13085.

Sparrow, R. (2023). Friendly AI will still be our master. Or, why we should not want to be the pets of super-intelligent computers. *AI & SOCIETY*. https://doi.org/10.1007/s00146-023-01698-x.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., & Wei, J. (2022). *Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them* (arXiv:2210.09261). arXiv. https://doi.org/10.48550/arXiv.2210.09261.

Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2020). Alignment for Advanced Machine Learning Systems. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 342–382). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0013.

Toromanoff, M., Wirbel, E., & Moutarde, F. (2019). *Is Deep Reinforcement Learning Really Superhuman on Atari? Leveling the playing field* (arXiv:1908.04683). arXiv. https://doi.org/10.48550/arXiv.1908.04683.

Torres, P. (2019). Existential risks: A philosophical analysis. *Inquiry : A Journal of Medical Care Organization, Provision and Financing*, *0*(0), 1–26. https://doi.org/10.1080/0020174X.2019.1658626.

Vold, K., & Harris, D. R. (2021). How Does Artificial Intelligence Pose an Existential Risk? In C. Véliz (Ed.), *The Oxford Handbook of Digital Ethics* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198857815.013.36.

Wallach, W., & Vallor, S. (2020). Moral Machines: From Value Alignment to Embodied Virtue. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 383–412). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0014.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models* (arXiv:2206.07682). arXiv. https://doi.org/10.48550/arXiv.2206.07682.