A Review of the Evidence for Existential Risk from AI via Misaligned Power-Seeking

Rose Hadshar
AI Impacts
rosehadshar@gmail.com

Abstract

Rapid advancements in artificial intelligence (AI) have sparked growing concerns among experts, policymakers, and world leaders regarding the potential for increasingly advanced AI systems to pose existential risks. This paper reviews the evidence for existential risks from AI via misalignment, where AI systems develop goals misaligned with human values, and power-seeking, where misaligned AIs actively seek power. The review examines empirical findings, conceptual arguments and expert opinion relating to specification gaming, goal misgeneralization, and power-seeking. The current state of the evidence is found to be concerning but inconclusive regarding the existence of extreme forms of misaligned power-seeking. Strong empirical evidence of specification gaming combined with strong conceptual evidence for power-seeking make it difficult to dismiss the possibility of existential risk from misaligned power-seeking. On the other hand, to date there are no public empirical examples of misaligned power-seeking in AI systems, and so arguments that future systems will pose an existential risk remain somewhat speculative. Given the current state of the evidence, it is hard to be extremely confident either that misaligned power-seeking poses a large existential risk, or that it poses no existential risk. The fact that we cannot confidently rule out existential risk from AI via misaligned power-seeking is cause for serious concern.

1 Executive summary

Concerns that artificial intelligence could pose an existential risk are growing.

This report reviews the evidence for existential risk from AI, focusing on arguments that future AI systems will pose an existential risk through misalignment and power-seeking:

- *Misalignment*: Some capable AI systems will develop goals which are misaligned with human goals.
  - *Specification gaming*: Some capable AI systems will learn designer-specified goals which diverge from intended goals in unforeseen ways.
  - *Goal misgeneralization*: Some capable AI systems will develop goals which are perfectly correlated with intended goals in training, but diverge once the systems are deployed.
- *Power-seeking*: Some capable, misaligned AI systems will seek power in order to achieve their goals.

Our findings are based on a review of relevant literature, a series of interviews with AI researchers working on existential risk from AI (AI Impacts, 2023d), and a new database of empirical evidence for some claims about existential risk from AI (Hadshar, 2023).

We find that the current state of the evidence for existential risk from misaligned power-seeking is concerning but inconclusive.

- There is strong empirical evidence of specification gaming and related phenomena, both in AI systems and other contexts, but it remains unclear whether specification gaming will be sufficiently extreme to pose an existential risk.
- For goal misgeneralization, the evidence is more speculative. Examples of goal misgeneralization to date are sparse, open to interpretation, and not in themselves harmful. It's unclear whether the evidence for goal misgeneralization is weak because it is not in fact a phenomenon which will affect AI systems, or because it will only affect AI systems once they are more goal-directed than at present.
- There is also limited empirical evidence of power-seeking, but there are strong conceptual arguments and formal proofs which justify a stronger expectation that power-seeking will arise in some AI systems.

Given the current state of the evidence, it is hard to be very confident either that misaligned power-seeking poses a large existential risk, or that it poses no existential risk.

That we cannot confidently rule out existential risk from AI via misaligned power-seeking is cause for serious concern.

Contents

2 Introduction

Many claim that artificial intelligence could pose an existential risk - that **AI could lead to human extinction, or to a catastrophe which destroys humanity's potential.**[1]

Individual researchers have been making this claim for the last decade (Bostrom, 2014; Christian, 2020; Ord, 2020; Russell, 2019). More recently, the number of voices raising concerns about existential risk from AI has grown. In May 2023, hundreds of experts signed an open letter stating that "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war" (Centre for AI Safety, 2023). Politicians have also begun to speak about the need to manage existential risk. For example, the UK's Science, Innovation and Technology Committee has identified "the existential challenge" of AI as "a major threat to human life" as one of twelve areas for policymakers to address (UK Parliament's Science and Committee, 2023).

The argument that AI could pose an existential risk has been well made elsewhere (Bostrom, 2014; Carlsmith, 2022; Hendrycks et al., 2023; Ord, 2020). The increasing prominence of the argument that AI could pose an existential risk, combined with the growing evidence base for some aspects of this argument, make now a good time to review the strength of the evidence for existential risk from AI.

2.1 Scope

There are several different pathways to existential risk from AI.

The 2023 UK AI Safety Summit focuses on two of these pathways:[2]

- "**Misuse risks**,[3] for example where a bad actor is aided by new AI capabilities in biological or cyber-attacks, development of dangerous technologies, or critical system interference"
- "**Loss of control risks** that could emerge from advanced systems that we would seek to be aligned with our values and intentions" (UK Parliament's Science and Committee, 2023)

A particular class of loss of control risks is **risks from misaligned power-seeking** (Carlsmith, 2022). The basic argument for existential risk from misaligned power-seeking is that:[4]

- (*Preconditions*) In the not-too-distant future, some AI systems will be sufficiently capable to pose an existential risk.
- (*Misalignment*) Some capable AI systems will develop goals which are misaligned with human goals.
- (*Power-seeking*) Some capable, misaligned AI systems will seek power in order to achieve their goals.
- (*Existential consequences*) This misaligned power-seeking will lead to human disempowerment, which will constitute an existential catastrophe.

This report reviews the evidence for existential risk from future AI systems via misalignment and power-seeking.

The following table breaks down the argument for existential risk from misaligned power-seeking further, and highlights the areas which are in the scope of this report.

Appendix B gives a shallow review of the evidence for some further claims about existential risk from AI which are outside of the scope of this report.

[1] Ord defines an existential catastrophe as "the destruction of humanity's long-term potential" (Ord, 2020).

[2] Some scholars have also pointed out a third pathway to existential risk from AI, via multi-agent interactions. See Critch and Krueger (2020); Drexler (2019); Manheim (2019), and the Alignment of Complex Systems Research Group.

[3] See Hendrycks et al. (2023) for an introduction to misuse risks, which they term 'Malicious use".

[4] See Appendix A for a discussion of the more detailed argument given in Carlsmith (2022).

Table 1: The argument for existential risk from misaligned power-seeking

Preconditions: In the not-too-distant future, some AI systems will be sufficiently capable to pose an existential risk.	• *Timelines:* The relevant AI systems will be developed in the not-too-distant future. • *Capabilities:* Some AI systems will be highly capable, in the sense that they are able to perform many important tasks at or above human level. • *Goal-directedness:* Some AI systems will be goal-directed, in that they pursue goals consistently over long time periods. • *Situational awareness:*[5] Some AI systems will be aware that they are AI systems, and whether they are in training or deployment.
Misalignment:[6] Some capable AI systems will develop goals which are misaligned with human goals.	• **Specification gaming:**[7] Some capable AI systems will learn designer-specified goals which diverge from intended goals in unforeseen ways. • **Goal misgeneralization:**[8] Some capable AI systems will develop goals which are perfectly correlated with intended goals in training, but diverge once the systems are deployed.
Power-seeking:[9] Some capable, misaligned AI systems will seek power in order to achieve their goals.	
Existential consequences: This misaligned power-seeking will lead to human disempowerment, which will constitute an existential catastrophe.	• *Disempowerment:* This misaligned power-seeking will lead to permanent human disempowerment. • *Existential catastrophe:* Permanent human disempowerment will constitute an existential catastrophe.

2.2 Methodology

This report is based on:

1. **A review of the relevant literature on misaligned power-seeking**

2. **A series of interviews with AI researchers working on existential risk from AI**

   We interviewed six AI researchers about the strength of the evidence for existential risk from AI. Summaries and recordings of some of the interviews can be found here.

Note that the sample size is small, and we did not interview AI researchers who are skeptical of existential risk from AI.[10]

3. **A new database of empirical evidence for some claims about existential risk from AI**

   The full database can be accessed here. It covers empirical evidence only, and includes evidence relating to specification gaming, goal misgeneralization and power-seeking (as well as deceptive alignment, self-improvement, and other claims relating to existential risk from AI).

   The database draws significantly from existing databases on specification gaming (Krakovna, 2020) and goal misgeneralization (Shah et al., 2022a).

[10]We didn't have the resources to interview a representative sample, and decided that we would get the most relevant information from speaking with researchers who work on AI existential risk and so are familiar with the evidence.

3 A review of the evidence for existential risk from misaligned power-seeking

Most of the AI existential risk researchers we interviewed regarded the evidence for misaligned power-seeking as at least somewhat speculative or uncertain. [11] Below, we review the evidence for misaligned power-seeking, including both conceptual and empirical evidence.

3.1 The strength of the empirical evidence

In general, the empirical evidence is weaker than the conceptual arguments for these claims about existential risk from AI. This is discussed in the relevant sections, but there are also some general points to make about the relative weakness of empirical evidence for misaligned power-seeking.

Firstly, there are other properties of AI systems which might prove to be preconditions of misaligned power-seeking, but which current systems have not yet attained. It is plausible that systems will only display misaligned power-seeking at higher levels of general capabilities for example,[12] or that misaligned power-seeking requires a higher level of goal-directedness than current systems have.[13]

Secondly, several of the AI researchers we interviewed clarified that the empirical evidence so far forms only a small or very small part of their reasons for concern about misaligned power-seeking, with more weight placed on conceptual arguments.[14]

3.2 The evidence for misalignment

In this report, we consider two routes to capable AI systems developing goals which are misaligned with human goals:

- **Specification gaming**,[15] where some capable AI systems learn designer-specified goals which diverge from intended goals in unforeseen ways.

[11] "The main best objection I get from really smart people on this is that most of the evidence is of a weaker or more speculative form than what we are used to using to evaluate policies, at least really expensive policies like the ones AI doomers are advocating. They basically say, if I believed you based on these sorts of arguments, I would also have to believe lots of other people saying crazy sounding things. And I think they're right that this is actually a weaker form of evidence that's easier to spoof." [36:07] (AI Impacts, 2023a)

"I think that evidence for goal-directedness and correspondingly power-seeking is weaker. There's kind of a cluster of arguments that are based on systems being goal-directed, both real goal misgeneralization and intentional power-seeking, and so on. And that's something that we're more uncertain about... deceptive alignment is also part of that cluster because that also relies on the system developing more goal-directedness." [56:25] (AI Impacts, 2023c)

"The arguments about misalignment risk are definitely more uncertain in that they are doing more extrapolation. Both arguments are doing extrapolation. I think the misalignment stuff is sometimes doing a bit more of a difficult extrapolation, because it's extrapolating these generalization properties which is just notoriously hard to do. I think that means that the case is just much more uncertain, but the case that the stakes are big is very good." [47:16] (AI Impacts, 2023b)

[12] "The story of you train an AI to fetch a coffee and then it realizes that the only way it can do that is to take over the world is a story about misgeneralization. And it's happening at a very high level of abstraction. You're using this incredibly intelligent system which is reasoning at a very high level about things and it's making the error at that high level... And I think the state of the evidence is... we've never observed a misgeneralization failure at such a high level of abstraction, but that's what we would expect because we don't have AIs that can even reason at that kind of level of abstraction." [28:36] AI Impacts (2023b)

[13] "What I'm expecting is happening here is that current systems are not goal-directed enough to show real power-seeking. And so the power-seeking threat model becomes more reliant on these kind of extrapolations of when there are systems which are more capable, they'll probably be at least somewhat more goal-directed and then once we have goal-directedness, we can more convincingly argue that power-seeking is going to be a thing because we have theory and so on, but there's a lot of uncertainty about it because we don't know how much systems will become more goal-directed." [54:35] (AI Impacts, 2023c)

[14] "[Hadshar] Empirical details about capabilities that AI systems have now don't sound very important to your world view. [Researcher] Exactly." [30:08] (AI Impacts, 2023a)

"I think that theoretical or conceptual arguments do have a lot of weight. Maybe I would put that at 60% and empirical examples at 40%, but I'm pulling this out of the air a little bit." [24:00] (AI Impacts, 2023c)

[15] "Specification gaming is a behavior that satisfies the literal specification of an objective without achieving the intended outcome." (Krakovna et al., 2020). Specification gaming is related to proxy gaming (Hendrycks et al., 2023), side effects (Amodei et al., 2016; Leike et al., 2017), reward gaming (Leike et al., 2017), reward

- **Goal misgeneralization**,[16] where some capable AI systems develop goals which are perfectly correlated with intended goals in training, but diverge once the systems are deployed.

3.2.1 The evidence for specification gaming

One route to AI systems developing misaligned goals is specification gaming, where AI systems learn the goals which they are given, but these goals are misspecified and come apart from intended goals.

"Specification gaming is a behavior that satisfies the literal specification of an objective without achieving the intended outcome." (Krakovna et al., 2020) If sufficiently powerful AI systems were to be deployed in high-stakes settings, then the difference between the literal specification and the intended outcome could become extreme, leading to catastrophic outcomes (Bostrom, 2014; Pueyo, 2018).

Specification gaming is a well-established phenomenon, both in general and in the context of AI systems.

In non-AI contexts, there are numerous examples of variants of specification gaming,[17] in economics (Braganza, 2022; Chrystal and Mizen, 2003; Goodhart, 1984; Kelly and Snower, 2021; Lucas, 1976), education (Berliner and Nichols, 2005; Campbell, 1979; Elton, 2004; Fire and Guestrin, 2019; Koretz, 2008; Strathern, 1997; Stroebe, 2016), healthcare (O'Mahony, 2017; Poku, 2016) and other areas.[18] It is clear that at least in human and social systems, such dynamics are widespread.

In the context of AI systems, there are both theoretical demonstrations of specification gaming given certain model assumptions (Beale et al., 2020; Hennessy and Goodhart, 2023; Manheim and Garrabrant, 2019; Zhuang and Hadfield-Menell, 2021), and many empirical examples of specification gaming in AI systems, both in toy environments and in deployment (Krakovna et al., 2020).[19]

For example, OpenAI trained an agent to play the game CoastRunners. The agent was rewarded for hitting targets along the course of a boat race. But instead of racing to the finish line, the agent discovered a loophole where it could race in a circle, repeatedly crashing and setting itself on fire, to earn maximum points (Jack Clark, 2016).

While a majority of clear examples of specification gaming in AI systems arise in toy environments like CoastRunners (Krakovna, 2020), there are already some examples of deployed AI systems engaging in specification gaming, and of this behavior leading to harm, particularly in the areas of bias and misinformation.

For example, a healthcare screening system deployed in 2019 was trained to predict health care costs. As less is spent on Black patients' care because of unequal access to healthcare, the algorithm rated Black patients as less sick than White patients even where Black patients had more underlying chronic illnesses (Obermeyer et al., 2019).

Falsehoods generated by large language models can also be viewed as the result of specification gaming, though here the case is less clear. Language models trained to accurately predict the next token frequently generate false content (Collective, 2023a,b; Heaven, 2022), but as one of our

hacking (Amodei et al., 2016; Skalse et al., 2022), reward misspecification (Ngo et al., 2023), and Goodhart's law (Hennessy and Goodhart, 2023; Manheim and Garrabrant, 2019; Thomas and Uminsky, 2022).

[16] "Goal misgeneralization is a specific form of robustness failure for learning algorithms in which the learned program competently pursues an undesired goal that leads to good performance in training situations but bad performance in novel test situations." (Shah et al., 2022b). Goal misgeneralization is related to goal drift (Hendrycks et al., 2023) and distributional shift (Amodei et al., 2016; Leike et al., 2017).

[17] For discussions about a cluster of related concepts including Goodhart's Law and proxy failure, see Amodei et al. (2016); John et al. (2023); Manheim and Garrabrant (2019); Thomas and Uminsky (2022).

[18] See Table 1 in John et al. (2023) for a collection of examples.

[19] The database linked to from this post contains over 70 examples of specification gaming. See also Hadshar (2023).

"One form of the problem has also been studied in the context of feedback loops in machine learning systems (particularly ad placement), based on counterfactual learning and contextual bandits. The proliferation of reward hacking instances across so many different domains suggests that reward hacking may be a deep and general problem, and one that we believe is likely to become more common as agents and environments increase in complexity." (Amodei et al., 2016). "Reward hacking—where RL agents exploit gaps in misspecified reward functions—has been widely observed" (Pan et al., 2022).

interviewees pointed out, it is a matter of judgment whether this is best interpreted as specification gaming or as a simple capability failure.[20]

The evidence is strong that AI systems will be subject to specification gaming to some degree. It remains unclear whether specification gaming will be sufficiently serious to pose an existential risk. In order to cause large-scale harms, misspecified goals would need to be subtle enough that systems were still deployed in high-stakes settings, but diverge extremely from intended goals in deployment. To date, no examples of specification gaming in AI systems have been catastrophic, so there is no direct evidence of this degree of harm from specification gaming.

There are some tentative signs that specification might become a more serious problem as models become more capable. In initial experiments, larger language models and language models with more RLHF are more prone to sycophantic answers, and to expressing a desire to seek power and avoid shutdown (Perez et al., 2022). Insofar as these behaviors are indeed caused by specification gaming,[21] this is cause for concern. Another study has found that when goals are misspecified, more capable RL agents will diverge more from intended goals than less capable agents, suggesting that specification gaming may worsen as capabilities improve. The same study also found that the divergence between intended and misspecified goals was sometimes very sudden, which might make it hard to anticipate and prevent such problems arising in deployment (Pan et al., 2022).

Overall, the evidence for specification gaming is strong, though it remains unclear whether the scale of the problem will be sufficient to pose an existential risk.

3.2.2 The evidence for goal misgeneralization

Another route to AI systems developing misaligned goals is goal misgeneralization, where systems develop goals which are perfectly correlated with intended goals in training, but diverge once the systems are deployed.

"Goal misgeneralization is a specific form of robustness failure for learning algorithms in which the learned program competently pursues an undesired goal that leads to good performance in training situations but bad performance in novel test situations." (Shah et al., 2022b)

The underlying mechanism behind goal misgeneralization is distributional shift, where there are systematic differences between the training distribution and the test distribution. Distributional shift is a very widely documented phenomenon in AI systems (Leike et al., 2017; Quinonero-Candela et al., 2022), and out-of-distribution robustness remains unsolved (Hendrycks et al., 2021; Liu et al., 2023). This provides a reason to expect goal misgeneralization to arise.

However, the empirical evidence for goal misgeneralization is currently weak, in spite of the prevalence of distributional shift.

There are examples of goal misgeneralization in AI systems (DeGrave et al., 2021; Langosco et al., 2023; Shah et al., 2022b). However, these examples do not conclusively show that goal misgeneralization will arise in a harmful way.

Firstly, all of the examples of goal misgeneralization we have found take place in demonstration, rather than in deployed systems. Sometimes these demonstrations involve very obvious and crude differences between the training data and the test data. For instance, Langosco et al. (2023) train a CoinRun agent exclusively on mazes where the cheese is always in the upper right hand corner, and show in testing that the agent learns to navigate to the upper right rather than to the cheese. This shows that goal misgeneralization can occur when the training data is very different to the test data - but doesn't provide evidence for goal misgeneralization in more realistic settings. We have not found any evidence of real-world harm from goal misgeneralization so far.

Secondly, it is currently not possible to demonstrate conclusively that examples of goal misgeneralization actually involve systems learning a goal which is correlated in training but not deployment. It is

[20] "With some of the language model examples, I think you can ask the question, is this really specification gaming, or is it capability failure, or something like that? I think sometimes there's a bit of a judgment call there." [29:45] (AI Impacts, 2023c)

[21] That is, the systems are following the specified goal of generating text which receives high positive feedback from humans, but this comes apart from the goal of generating helpful, honest and harmless text. See also Krakovna (2020).

only possible to observe the behavior of the system in question, not its inner workings, so we cannot know what goal (if any) a system has learned. Examples to date only conclusively show behavioral or functional goal misgeneralization.[22]

Furthermore, it's often hard to distinguish goal misgeneralization from capability misgeneralization, where the system's capabilities also fail to generalize.[23] In the abstract, goal misgeneralization is distinct from capability misgeneralization: "a system's capabilities generalize but its goal does not generalize as desired. When this happens, the system competently pursues the wrong goal." (Shah et al., 2023) But in real-world settings, the wrong goal may often lead to capability failure. A system which learns to competently predict that tumors with rulers are malignant based on its training data will fail to competently predict actual malignancy when tested on more diverse data (Narla et al., 2018). Insofar as goal misgeneralization comes with capability misgeneralization, AI systems which learn very misgeneralized goals are unlikely to be deployed.

There are several possible explanations of the weakness of evidence on goal misgeneralization so far.

Goal misgeneralization might require a level of goal-directedness which current systems don't yet have,[24] or an ability to reason at higher levels of abstraction.[25] Reliably identifying goal misgeneralization might also require more advanced interpretability techniques.[26] Alternatively, the distinction between behavioral and 'actual' goal misgeneralization may be misplaced: if sufficiently capable systems engage in behaviors which look like goal misgeneralization, then functionally they are misaligned whether or not their internal representations match our description of goal misgeneralization.

So there are some reasons to expect the current evidence of goal misgeneralization to be weak, even if the phenomenon eventually arises strongly. Nevertheless, so far the evidence for goal misgeneralization remains reasonably speculative.[27]

3.3 The evidence for power-seeking

The presence of misaligned goals in and of itself need not pose an existential risk. But if AI systems with misaligned goals successfully and systematically seek power, the result could be existential.

In Carlsmith (2022), power-seeking is defined as "active efforts by an AI system to gain and maintain power in ways that designers didn't intend, arising from problems with that system's objectives."

[22] "I think right now the examples we have are more like behavioral goal misgeneralization where you just have different behaviors that are all the same in training but then they become decoupled in the new setting but we don't know how the behavior is going to generalize. We call it goal misgeneralization maybe more as a shorthand. The behavior has different ways of generalizing that are kind of coherent. We can present it as the system learned the wrong goal, but we can't actually say that it has learned a goal. Maybe it's just following the wrong heuristic or something. I think the current examples are a demonstration of the more obvious kind of effect where the training data doesn't distinguish between all the ways that the behavior could generalize." [37:11] (AI Impacts, 2023c)

[23] "I think it's a less well understood phenomenon... it can be hard to distinguish capability misgeneralization from goal misgeneralization." [33:16] (AI Impacts, 2023c)

[24] "Specifying something as goal misgeneralization also requires some assumption that the system is goal-directed to some degree and that can also be debatable." [33:16] (AI Impacts, 2023c)

[25] "The story of you train an AI to fetch a coffee and then it realizes that the only way it can do that is to take over the world is a story about misgeneralization. And it's happening at a very high level of abstraction. You're using this incredibly intelligent system which is reasoning at a very high level about things and it's making the error at that high level... And I think the state of the evidence is... we've never observed a misgeneralization failure at such a high level of abstraction, but that's what we would expect because we don't have AIs that can even reason at that kind of level of abstraction." [28:36] (AI Impacts, 2023b)

[26] "The mechanism is a lot less well understood. I think to really properly diagnose goal misgeneralization we would need better interpretability tools." [36:30] (AI Impacts, 2023c)

[27] "I think [the evidence for goal misgeneralization] is not as strong [as for specification gaming]." [33:16] (AI Impacts, 2023c) "These generalization failures at new levels of abstraction are notoriously hard to predict. You have to try and intuit what an extremely large scale neural net will learn from the training data and in which ways it will generalize... I'm relatively persuaded that misgeneralization will continue to happen at higher levels of abstraction, but whether that actually is well described by some of the typical power-seeking stories I'm much less confident and it's definitely going to be a judgment call." [28:36] (AI Impacts, 2023b)

Carlsmith loosely defines power as "the type of thing that helps a wide variety of agents pursue a wide variety of objectives in a given environment." (Carlsmith, 2022) We can take Bostrom's categories of instrumental goals as illustrative of this "type of thing":

- Self-preservation
- Goal-content integrity[28]
- Cognitive enhancement
- Technological perfection[29]
- Resource acquisition (Bostrom, 2012)

The conceptual argument that some AI systems will seek power seems strong.[30] Bostrom's instrumental convergence thesis is simple and intuitively plausible: "as long as they possess a sufficient level of intelligence, agents having any of a wide range of final goals will pursue similar intermediary goals because they have instrumental reasons to do so." (Bostrom, 2012)

There are formal proofs that the instrumental convergence thesis holds for various kinds of AI systems. Turner et al. (2023) prove that "most reward functions make it optimal to seek power by keeping a range of options available" in the context of Markov decision processes. Turner and Tadepalli (2022) extend this result to a class of sub-optimal policies, showing that "many decision-making functions are retargetable, and that retargetability is sufficient to cause power-seeking tendencies". Krakovna and Kramar (2023) further show that agents which learn a goal are likely to engage in power-seeking.

The formal and theoretical case for power-seeking in sufficiently capable and goal-directed AI systems is therefore relatively strong.

However, the empirical evidence of power-seeking in AI systems is currently weak. There are some demonstrations of RL agents engaging in power-seeking behaviors in toy environments (for example, Hadfield-Menell et al. (2017)), but no convincing examples of AI systems in the real world seeking power in this way to date.[31]

Perez et al. (2022) show language models giving "answers that indicate a willingness to pursue potentially dangerous subgoals: resource acquisition, optionality preservation, goal preservation, powerseeking, and more." But indicating willingness is not the same as actually engaging in power-seeking behaviors. Language models might express power-seeking desires merely because their training data contains similar text, and not because they will ever directly seek power.

Sycophancy, where language models agree with their users regardless of the accuracy of the statements, could be taken as an example of power-seeking behavior. But as with the results of Perez et al. (2022), sycophancy is likely to be simply an imitation of the training data, rather than an intentional behavior.[32]

If the theoretical arguments for power-seeking are strong, why is the empirical evidence to date weak?

As with goal misgeneralization, one plausible explanation is that power-seeking behavior depends on a level of goal-directedness or capability in general which current models don't yet have.[33]

[28] "An agent is more likely to act in the future to maximize the realization of its present final goals if it still has those goals in the future. This gives the agent a present instrumental reason to prevent alterations of its final goals." (Bostrom, 2012)

[29] "An agent may often have instrumental reasons to seek better technology, which at its simplest means seeking more efficient ways of transforming some given set of inputs into valued outputs." (Bostrom, 2012)

[30] "I think some of the other theoretical arguments like instrumental convergence also generally seems like a very clear argument, and we can observe some of these effects in human systems and corporations and so on." [25:23] (AI Impacts, 2023c)

[31] "I don't think there's really empirical evidence [for power-seeking]... To me it's very uncertain." [28:36] (AI Impacts, 2023b)

[32] "Looking at current systems, sycophancy can be considered as a form of power-seeking. Although I think that's also maybe debatable. It's building more influence with the user by agreeing with their views, but it's probably more of a heuristic that is just somehow reinforced than intentional power-seeking." [49:35] (AI Impacts, 2023c)

[33] "What I'm expecting is happening here is that current systems are not goal-directed enough to show real power-seeking. And so the power-seeking threat model becomes more reliant on these kind of extrapolations of

Overall, with strong conceptual arguments but no public empirical evidence, it seems plausible but unproven that some AI systems will seek power.

when there are systems which are more capable, they'll probably be at least somewhat more goal-directed and then once we have goal-directedness, we can more convincingly argue that power-seeking is going to be a thing because we have theory and so on, but there's a lot of uncertainty about it because we don't know how much systems will become more goal-directed." [54:35] (AI Impacts, 2023c)

4 Conclusion: The current strength of the evidence for existential risk from misaligned power-seeking

The current state of the evidence for existential risk from misaligned power-seeking is concerning but inconclusive.

There is strong empirical evidence of specification gaming and related phenomena, both in AI systems and other contexts. We can be reasonably confident therefore that specification gaming will arise to some extent in future AI systems, but it remains unclear whether specification gaming will be sufficiently extreme to pose an existential risk.

For goal misgeneralization, the evidence is more speculative. Distributional shift, which is a pre-requisite of goal misgeneralization, is a well-documented phenomenon, but the examples of goal misgeneralization to date are sparse, open to interpretation, and not in themselves harmful. It's unclear whether there is weak evidence for goal misgeneralization because it is not in fact a phenomenon which will affect AI systems to a harmful degree, or because it will only affect AI systems once they are more goal-directed than at present.

There is also limited empirical evidence of power-seeking, but there are strong conceptual arguments and formal proofs which justify a stronger expectation that power-seeking will arise in some AI systems.

Strong empirical evidence of specification gaming combined with strong conceptual arguments for power-seeking make it difficult to dismiss the possibility of existential risk from misaligned power-seeking. On the other hand, we are not aware of any empirical examples of misaligned power-seeking in AI systems, and so arguments that future systems will pose an existential risk must remain somewhat speculative.

Given the current state of the evidence, it is hard to be extremely confident either that misaligned power-seeking poses a large existential risk, or that it poses no existential risk.

That we cannot confidently rule out existential risk from AI via misaligned power-seeking is cause for serious concern.

5 Acknowledgements

Thanks to Katja Grace and Harlan Stewart in particular; to Michael Aird, Adam Bales, Rick Korzekwa, Fazl Barez, Sam Clark, Max Dalton, and many others for various levels of feedback and support; and to all the researchers we interviewed.

6 References

AI Impacts. 2021. What do coherence arguments imply about the behavior of advanced AI? https://wiki.aiimpacts.org/doku.php?id=agency:what_do_coherence_arguments_imply_about_the_behavior_of_advanced_ai&s[]=goal&s[]=directed.

AI Impacts. 2023a. Interview on the strength of the evidence for AI risk claims with an anonymous AI alignment researcher. https://wiki.aiimpacts.org/arguments_for_ai_risk/is_ai_an_existential_threat_to_humanity/interviews_on_the_strength_of_the_evidence_for_ai_risk_claims/summary_of_an_interview_on_the_strength_of_the_evidence_for_ai_risk_claims_with_anonymous_ai_alignment_researcher.

AI Impacts. 2023b. Interview with Jacob Hilton on the strength of the evidence for AI risk claims. https://wiki.aiimpacts.org/arguments_for_ai_risk/is_ai_an_existential_threat_to_humanity/interviews_on_the_strength_of_the_evidence_for_ai_risk_claims/summary_of_an_interview_on_the_strength_of_the_evidence_for_ai_risk_claims_with_jacob_hilton. Online.

AI Impacts. 2023c. Interview with Victoria Krakovna on the strength of the evidence for AI risk claims. https://wiki.aiimpacts.org/arguments_for_ai_risk/is_ai_an_existential_threat_to_humanity/interviews_on_the_strength_of_the_evidence_for_ai_risk_claims/summary_of_an_interview_on_the_strength_of_the_evidence_for_ai_risk_claims_with_victoria_krakovna.

AI Impacts. 2023d. Interviews on the strength of the evidence for AI risk claims. https://wiki.aiimpacts.org/arguments_for_ai_risk/is_ai_an_existential_threat_to_humanity/interviews_on_the_strength_of_the_evidence_for_ai_risk_claims".

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. http://arxiv.org/abs/1606.06565.

Adam Bales. 2023. Will AI avoid exploitation? Artificial general intelligence and expected utility theory. *Philosophical Studies*.

Nicholas Beale, Heather Battey, Anthony C. Davison, and Robert S. MacKay. 2020. An unethical optimization principle. *Royal Society Open Science*, 7(7).

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. 2023. Taken out of context: On measuring situational awareness in LLMs. http://arxiv.org/abs/2309.00667.

David C Berliner and Sharon L Nichols. 2005. The inevitable corruption of indicators and educators through high-stakes testing. Technical report, Arizona State University.

Nick Bostrom. 2012. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, 22(2):71–85.

Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies*. OUP Oxford, Oxford, United Kingdom.

Oliver Braganza. 2022. Proxyeconomics, a theory and model of proxy-based competition and cultural evolution. *Royal Society Open Science*, 9(2).

Donald T. Campbell. 1979. Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1):67–90.

Joseph Carlsmith. 2022. Is Power-Seeking AI an Existential Risk? http://arxiv.org/abs/2206.13353.

Centre for AI Safety. 2023. Statement on AI Risk. https://www.safe.ai/statement-on-ai-risk#open-letter.

Brian Christian. 2020. *The Alignment Problem – Machine Learning and Human Values*. W. W. Norton & Company, New York, NY.

Alec Chrystal and Paul Mizen. 2003. Goodhart's Law: its origins, meaning and implications for monetary policy. In *Central Banking, Monetary Theory and Practice: Essays in Honour of Charles Goodhart*. Edward Elgar Publishing.

The Responsible AI Collective. 2023a. Incident 503: Bing AI Search Tool Reportedly Declared Threats against Users. https://incidentdatabase.ai/cite/503/.

The Responsible AI Collective. 2023b. Incident 511: Microsoft's Bing Failed to Fetch Movie Showtimes Results Due to Date Confusion. https://incidentdatabase.ai/cite/511/.

Andrew Critch and David Krueger. 2020. AI Research Considerations for Human Existential Safety (ARCHES). http://arxiv.org/abs/2006.04948.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619.

K Eric Drexler. 2019. Reframing Superintelligence. https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf.

EJT. 2023. There are no coherence theorems. https://forum.effectivealtruism.org/posts/FoRyordtA7LDoEhd7/there-are-no-coherence-theorems.

Lewis Elton. 2004. Goodhart's Law and Performance Indicators in Higher Education. *Evaluation & Research in Education*, 18(1-2):120–128.

Michael Fire and Carlos Guestrin. 2019. Over-optimization of academic publishing metrics: observing Goodhart's Law in action. *GigaScience*, 8(6).

C. A. E. Goodhart. 1984. Problems of Monetary Management: The UK Experience. In *Monetary Theory and Practice: The UK Experience*. Macmillan Education UK.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2017. The Off-Switch Game. http://arxiv.org/abs/1611.08219.

Dylan Hadfield-Menell and Gillian K. Hadfield. 2019. Incomplete Contracting and AI Alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422. ACM.

Rose Hadshar. 2023. Empirical evidence for existential AI risk factors. https://airtable.com/embed/appWYvYLkBIDckhAo/shrkuKrEf4zhdVBrD/tbl3KurpJxkFVcNJJ?backgroundColor=red&viewControls=on.

Will Douglas Heaven. 2022. Why Meta's latest large language model survived only three days online. https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, Montreal, QC, Canada. IEEE.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An Overview of Catastrophic AI Risks. http://arxiv.org/abs/2306.12001.

Christopher A. Hennessy and Charles A. E. Goodhart. 2023. Goodhart's Law and Machine Learning: A Structural Perspective. *International Economic Review*, 64(3):1075–1086.

Dario Amodei Jack Clark. 2016. Faulty reward functions in the wild. https://openai.com/research/faulty-reward-functions.

Yohan J. John, Leigh Caldwell, Dakota E. McCoy, and Oliver Braganza. 2023. Dead rats, dopamine, performance metrics, and peacock tails: proxy failure is an inherent risk in goal-oriented systems. *Behavioral and Brain Sciences*, pages 1–68.

Colm Kelly and Dennis J Snower. 2021. Capitalism recoupled. *Oxford Review of Economic Policy*, 37(4):851–863.

Daniel Koretz. 2008. *Measuring Up: What Educational Testing Really Tells Us*. Harvard University Press.

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

Viktoria Krakovna. 2020. Specification gaming examples in AI. https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml.

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

Lauro Langosco, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau, and David Krueger. 2023. Goal Misgeneralization in Deep Reinforcement Learning. http://arxiv.org/abs/2105.14111.

Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. 2017. AI Safety Gridworlds. http://arxiv.org/abs/1711.09883.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2023. Towards Out-Of-Distribution Generalization: A Survey. http://arxiv.org/abs/2108.13624.

Robert E. Lucas. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46.

David Manheim. 2019. Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence. *Big Data and Cognitive Computing*, 3(2).

David Manheim and Scott Garrabrant. 2019. Categorizing Variants of Goodhart's Law. http://arxiv.org/abs/1803.04585.

Akhila Narla, Brett Kuprel, Kavita Sarin, Roberto Novoa, and Justin Ko. 2018. Automated Classification of Skin Lesions: From Pixels to Practice. *Journal of Investigative Dermatology*, 138(10):2108–2110.

Richard Ngo, Lawrence Chan, and Sören Mindermann. 2023. The alignment problem from a deep learning perspective. http://arxiv.org/abs/2209.00626.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

S O'Mahony. 2017. Medicine and the Mcnamara Fallacy. *Journal of the Royal College of Physicians of Edinburgh*, 47(3):281–287.

Toby Ord. 2020. *The Precipice*. Bloomsbury Publishing.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. http://arxiv.org/abs/2201.03544.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. http://arxiv.org/abs/2212.09251.

Michael Poku. 2016. Campbell's Law: implications for health care. *Journal of Health Services Research & Policy*, 21(2):137–139.

Salvador Pueyo. 2018. Growth, degrowth, and the challenge of artificial superintelligence. *Journal of Cleaner Production*, 197:1731–1736.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2022. *Dataset Shift in Machine Learning*. MIT Press.

Stuart Russell. 2019. *Human Compatible: AI and the Problem of Control*. Allen Lane, London.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022a. Goal misgeneralization examples in AI. https://docs.google.com/spreadsheets/d/e/2PACX-1vTo3RkXUAigb25nP7gjpcHriR6XdzA_L5loOcVFj_u7cRAZghWrYKH2L2nU4TA_Vr9KzBX5Bjpz9G_l/pubhtml.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022b. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. http://arxiv.org/abs/2210.01790.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2023. Goal Misgeneralisation: Why Correct Specifications Aren't Enough For Correct Goals. https://deepmindsafetyresearch.medium.com/goal-misgeneralisation-why-correct-specifications-arent-enough-for-correct-goals-cf96ebc60924.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and Characterizing Reward Gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

Marilyn Strathern. 1997. 'Improving ratings': audit in the British University system. *European Review*, 5(3):305–321.

Wolfgang Stroebe. 2016. Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations. *Perspectives on Psychological Science*, 11(6):800–816.

Rachel L. Thomas and David Uminsky. 2022. Reliance on metrics is a fundamental challenge for AI. *Patterns*, 3(5).

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2023. Optimal Policies Tend to Seek Power. http://arxiv.org/abs/1912.01683.

Alexander Matt Turner and Prasad Tadepalli. 2022. Parametrically Retargetable Decision-Makers Tend To Seek Power. http://arxiv.org/abs/2206.13477.

Innovation UK Parliament's Science and Technology Committee. 2023. AI offers significant opportunities but twelve governance challenges must be addressed says science, innovation and technology committee. https://committees.parliament.uk/committee/135/science-innovation-and-technology-committee/news/197236/ai-offers-significant-opportunities-but-twelve-governance-challenges-must-be-addressed-says-science-innovation-and-technology-committee/.

Simon Zhuang and Dylan Hadfield-Menell. 2021. Consequences of Misaligned AI. http://arxiv.org/abs/2102.03896.

7 Appendix A: Carlsmith's argument for existential risk via power-seeking AI

The following table maps between the premises of Carlsmith (2022)'s argument, and the claims used in this report (see Table 1). Claims within the scope of this report are bolded.

Note that the claims used in this report are not identical to Carlsmith's premises, though they are closely related.

Carlsmith		**Claims used in this report**
By 2070:		(*Preconditions: Timelines*) The relevant AI systems will be developed in the not-too-distant future.
1. It will become possible and financially feasible to build AI systems with the following properties:	*Advanced capability*: they outperform the best humans on some set of tasks which when performed at advanced levels grant significant power in today's world (tasks like scientific research, business/military/political strategy, engineering, and persuasion/manipulation).	(*Preconditions: Capabilities*) Some AI systems will be highly capable, in the sense that they are able to perform many important tasks at or above human level
	Agentic planning: they make and execute plans, in pursuit of objectives, on the basis of models of the world.	(*Preconditions: Goal-directedness*) Some AI systems will be goal-directed, in that they pursue goals consistently over long time periods.
	Strategic awareness: the models they use in making plans represent with reasonable accuracy the causal upshot of gaining and maintaining power over humans and the real-world environment.	(*Preconditions: Situational awareness*) Some AI systems will be aware that they are AI systems, and whether they are in training or deployment.
	(Call these "APS"—Advanced, Planning, Strategically aware—systems.)	
2. There will be strong incentives to build and deploy APS systems.		
3. It will be much harder to build APS systems that would not seek to gain and maintain power in unintended ways (because of problems with their objectives) on any of the inputs they'd encounter if deployed, than to build APS systems that would do this, but which are at least superficially attractive to deploy anyway.		(*Misalignment*) Some capable AI systems will develop goals which are misaligned with human goals. (*Misalignment: Specification gaming*) Some capable AI systems will learn designer-specified goals which diverge from intended goals in unforeseen ways. (*Misalignment: Goal misgeneralization*) Some capable AI systems will develop goals which are perfectly correlated with intended goals in training, but diverge once the systems are deployed.
4. Some deployed APS systems will be exposed to inputs where they seek power in unintended and high-impact ways (say, collectively causing >$1 trillion dollars of damage), because of problems with their objectives.		(*Power-seeking*) Some capable, misaligned AI systems will seek power in order to achieve their goals.

5. Some of this power-seeking will scale (in aggregate) to the point of permanently disempowering all of humanity.	(*Existential consequences: Disempowerment*) This misaligned power-seeking will lead to permanent human disempowerment.
6. This disempowerment will constitute an existential catastrophe.	(*Existential consequences: Existential catastrophe*) Permanent human disempowerment will constitute an existential catastrophe.

8 Appendix B: Some evidence for other claims about existential risk from AI

We systematically reviewed the evidence for claims about misalignment and power-seeking. However, in the course of our research and interviews, we came across some evidence for other relevant claims.

This appendix contains some of the evidence for goal-directedness, situational awareness, and deceptive alignment. It should not be treated as a comprehensive review of the state of the evidence on these topics.

8.1 Some evidence for goal-directedness

Roughly, goal-directedness refers to a property of AI systems to persistently pursue a goal. [34] Goal-directedness has not been well-defined so far, and so reviewing the evidence for goal-directedness is hampered by unclarity about the concept. [35]

That said, it seems plausible that goal-directedness is a direct precondition for goal misgeneralization and for power-seeking,[36] so it is an important claim to assess.

Coherence theorems offer one kind of conceptual evidence for goal-directedness, but the extent to which they apply to future AI systems is contested (Bales, 2023; EJT, 2023; AI Impacts, 2021).[37]

There is limited empirical evidence of goal-directedness in systems so far.[38] One of the researchers we interviewed noted that language models may be particularly unsuited to goal-directedness.[39]

However, individual researchers we interviewed believe that:

- To the extent that language models can simulate humans, they will have the ability to simulate goal-directedness.[40]
- There is a clear trend towards systems acting more autonomously.[41]

[34] In Carlsmith (2022), goal-directedness is referred to as "agentic planning", where AI systems "make and execute plans, in pursuit of objectives, on the basis of models of the world."

[35] "Right now it's really hard to distinguish between real goal-directedness and learned heuristics… I think part of the problem with goal-directedness is we don't really understand the phenomenon that well." [44:00] (AI Impacts, 2023c)

[36] "Specifying something as goal misgeneralization also requires some assumption that the system is goal-directed to some degree and that can also be debatable." [33:16] (AI Impacts, 2023c) "What I'm expecting is happening here is that current systems are not goal-directed enough to show real power-seeking. And so the power-seeking threat model becomes more reliant on these kind of extrapolations of when there are systems which are more capable, they'll probably be at least somewhat more goal-directed and then once we have goal-directedness, we can more convincingly argue that power-seeking is going to be a thing because we have theory and so on, but there's a lot of uncertainty about it because we don't know how much systems will become more goal-directed." [54:35] (AI Impacts, 2023c)

[37] "Some of the theoretical arguments make the case that goal-directedness is an attractor. I think that's something that's more debatable, less clear to me. There have been various discussions on LessWrong and elsewhere about to what extent do coherence arguments imply goal-directedness. And I think the jury is still out on that one." [42:36] (AI Impacts, 2023c)

[38] "I think the evidence so far at least for language models, there isn't really convincing evidence of goal-directedness." [44:00] (AI Impacts, 2023c)

[39] "It's also possible goal-directedness is kind of hard. And especially, maybe language models are just a kind of system where goal-directedness comes less naturally than other systems like reinforcement learning systems or even with humans or whatever." [40:26] (AI Impacts, 2023c)

[40] "I think generally the kind of risk scenarios that we are most worried about would involve the system acting intentionally and deliberately towards some objectives but I would expect that intent and goal-directedness comes in degrees and if we see examples of increasing degrees of that then I think that does constitute evidence of that being possible. Although it's not clear whether it will go all the way to really deliberate systems, but I think especially to the extent that these systems can simulate humans… they have the ability to simulate deliberate intentional action and planning because that's something that humans can do." [20:20] (AI Impacts, 2023c)

[41] "We are already capable of getting AI systems to do simple things relatively autonomously. I don't think it's a threshold where now it's autonomous, now it's not… I think it's a spectrum and it's just very clearly ramping up. We already have things that have a little autonomy but not very much. I think it's just a pretty straightforward trend at this point." [24:39] (AI Impacts, 2023b)

One researcher we interviewed highlighted goal-directedness as one of their key uncertainties about existential risk from AI.[42]

8.2 Some evidence for situational awareness

"A model is situationally aware if it's aware that it's a model and can recognize whether it's currently in testing or deployment." (Berglund et al., 2023)

This is important to arguments about existential risk from AI as situational awareness is plausibly a precondition for successful misaligned power-seeking: a model may need to understand its own situation at a sophisticated level in order to make plans which successfully disempower humans. In particular, situational awareness seems like a precondition for deceptive alignment.

There is some empirical work demonstrating situational awareness in large language models, but the results are inconclusive (Berglund et al., 2023; Ngo et al., 2023; Perez et al., 2022). Berglund et al. (2023) find that language models can perform out-of-context reasoning tasks, but only with particular training set ups and data augmentation. Perez et al. (2022) run various experiments to test awareness, and find that "the models we evaluate are not aware of at least some basic details regarding themselves or their training procedures." On the other hand, Langosco et al. (2023) use the same questions as Perez et al. (2022) but find that their model answers 85% accurately.

[42] "I think we might see more goal-directed systems which produce clearer examples of internal goal misgeneralization, but also I wouldn't be that surprised if we don't see that. I think that's one of the big uncertainties I have about level of risk. How much can we expect goal-directedness to emerge?" [40:26] (AI Impacts, 2023c)