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