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.