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).