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.