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.