4 Conclusion: The current strength of the evidence for existential risk from misaligned power-seeking

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. We can be reasonably confident therefore that specification gaming will arise to some extent in future AI systems, but it remains unclear whether specification gaming will be sufficiently extreme to pose an existential risk.

For goal misgeneralization, the evidence is more speculative. Distributional shift, which is a pre-requisite of goal misgeneralization, is a well-documented phenomenon, but the examples of goal misgeneralization to date are sparse, open to interpretation, and not in themselves harmful. It's unclear whether there is weak evidence for goal misgeneralization because it is not in fact a phenomenon which will affect AI systems to a harmful degree, 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.

Strong empirical evidence of specification gaming combined with strong conceptual arguments for power-seeking make it difficult to dismiss the possibility of existential risk from misaligned power-seeking. On the other hand, we are not aware of any empirical examples of misaligned power-seeking in AI systems, and so arguments that future systems will pose an existential risk must 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.

That we cannot confidently rule out existential risk from AI via misaligned power-seeking is cause for serious concern.