Contents

1 Executive summary **2**

2 Introduction **4**

2.1 Scope 4

2.2 Methodology 5

3 A review of the evidence for existential risk from misaligned power-seeking **7**

3.1 The strength of the empirical evidence 7

3.2 The evidence for misalignment 7

3.2.1 The evidence for specification gaming 8

3.2.2 The evidence for goal misgeneralization 9

3.3 The evidence for power-seeking 10

4 Conclusion: The current strength of the evidence for existential risk from misaligned power-seeking **13**

5 Acknowledgements **14**

6 References **14**

7 Appendix A: Carlsmith's argument for existential risk via power-seeking AI **19**

8 Appendix B: Some evidence for other claims about existential risk from AI **21**

8.1 Some evidence for goal-directedness 21

8.2 Some evidence for situational awareness 22