5 Acknowledgements

Thanks to Katja Grace and Harlan Stewart in particular; to Michael Aird, Adam Bales, Rick Korzekwa, Fazl Barez, Sam Clark, Max Dalton, and many others for various levels of feedback and support; and to all the researchers we interviewed.

6 References

AI Impacts. 2021. What do coherence arguments imply about the behavior of advanced AI? https://wiki.aiimpacts.org/doku.php?id=agency:what_do_coherence_arguments_imply_about_the_behavior_of_advanced_ai&s[]=goal&s[]=directed.

AI Impacts. 2023a. Interview on the strength of the evidence for AI risk claims with an anonymous AI alignment researcher. https://wiki.aiimpacts.org/arguments_for_ai_risk/is_ai_an_existential_threat_to_humanity/interviews_on_the_strength_of_the_evidence_for_ai_risk_claims/summary_of_an_interview_on_the_strength_of_the_evidence_for_ai_risk_claims_with_anonymous_ai_alignment_researcher.

AI Impacts. 2023b. Interview with Jacob Hilton on the strength of the evidence for AI risk claims. https://wiki.aiimpacts.org/arguments_for_ai_risk/is_ai_an_existential_threat_to_humanity/interviews_on_the_strength_of_the_evidence_for_ai_risk_claims/summary_of_an_interview_on_the_strength_of_the_evidence_for_ai_risk_claims_with_jacob_hilton. Online.

AI Impacts. 2023c. Interview with Victoria Krakovna on the strength of the evidence for AI risk claims. https://wiki.aiimpacts.org/arguments_for_ai_risk/is_ai_an_existential_threat_to_humanity/interviews_on_the_strength_of_the_evidence_for_ai_risk_claims/summary_of_an_interview_on_the_strength_of_the_evidence_for_ai_risk_claims_with_victoria_krakovna.

AI Impacts. 2023d. Interviews on the strength of the evidence for AI risk claims. https://wiki.aiimpacts.org/arguments_for_ai_risk/is_ai_an_existential_threat_to_humanity/interviews_on_the_strength_of_the_evidence_for_ai_risk_claims".

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. http://arxiv.org/abs/1606.06565.

Adam Bales. 2023. Will AI avoid exploitation? Artificial general intelligence and expected utility theory. *Philosophical Studies*.

Nicholas Beale, Heather Battey, Anthony C. Davison, and Robert S. MacKay. 2020. An unethical optimization principle. *Royal Society Open Science*, 7(7).

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. 2023. Taken out of context: On measuring situational awareness in LLMs. http://arxiv.org/abs/2309.00667.

David C Berliner and Sharon L Nichols. 2005. The inevitable corruption of indicators and educators through high-stakes testing. Technical report, Arizona State University.

Nick Bostrom. 2012. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, 22(2):71–85.

Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies*. OUP Oxford, Oxford, United Kingdom.

Oliver Braganza. 2022. Proxyeconomics, a theory and model of proxy-based competition and cultural evolution. *Royal Society Open Science*, 9(2).

Donald T. Campbell. 1979. Assessing the impact of planned social change. *Evaluation and Program Planning*, 2(1):67–90.