Joseph Carlsmith. 2022. Is Power-Seeking AI an Existential Risk? http://arxiv.org/abs/2206.13353.

Centre for AI Safety. 2023. Statement on AI Risk. https://www.safe.ai/statement-on-ai-risk#open-letter.

Brian Christian. 2020. *The Alignment Problem – Machine Learning and Human Values*. W. W. Norton & Company, New York, NY.

Alec Chrystal and Paul Mizen. 2003. Goodhart's Law: its origins, meaning and implications for monetary policy. In *Central Banking, Monetary Theory and Practice: Essays in Honour of Charles Goodhart*. Edward Elgar Publishing.

The Responsible AI Collective. 2023a. Incident 503: Bing AI Search Tool Reportedly Declared Threats against Users. https://incidentdatabase.ai/cite/503/.

The Responsible AI Collective. 2023b. Incident 511: Microsoft's Bing Failed to Fetch Movie Showtimes Results Due to Date Confusion. https://incidentdatabase.ai/cite/511/.

Andrew Critch and David Krueger. 2020. AI Research Considerations for Human Existential Safety (ARCHES). http://arxiv.org/abs/2006.04948.

Alex J. DeGrave, Joseph D. Janizek, and Su-In Lee. 2021. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619.

K Eric Drexler. 2019. Reframing Superintelligence. https://www.fhi.ox.ac.uk/wp-content/uploads/Reframing_Superintelligence_FHI-TR-2019-1.1-1.pdf.

EJT. 2023. There are no coherence theorems. https://forum.effectivealtruism.org/posts/FoRyordtA7LDoEhd7/there-are-no-coherence-theorems.

Lewis Elton. 2004. Goodhart's Law and Performance Indicators in Higher Education. *Evaluation & Research in Education*, 18(1-2):120–128.

Michael Fire and Carlos Guestrin. 2019. Over-optimization of academic publishing metrics: observing Goodhart's Law in action. *GigaScience*, 8(6).

C. A. E. Goodhart. 1984. Problems of Monetary Management: The UK Experience. In *Monetary Theory and Practice: The UK Experience*. Macmillan Education UK.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2017. The Off-Switch Game. http://arxiv.org/abs/1611.08219.

Dylan Hadfield-Menell and Gillian K. Hadfield. 2019. Incomplete Contracting and AI Alignment. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 417–422. ACM.

Rose Hadshar. 2023. Empirical evidence for existential AI risk factors. https://airtable.com/embed/appWYvYLkBIDckhAo/shrkuKrEf4zhdVBrD/tbl3KurpJxkFVcNJJ?backgroundColor=red&viewControls=on.

Will Douglas Heaven. 2022. Why Meta's latest large language model survived only three days online. https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8320–8329, Montreal, QC, Canada. IEEE.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An Overview of Catastrophic AI Risks. http://arxiv.org/abs/2306.12001.

Christopher A. Hennessy and Charles A. E. Goodhart. 2023. Goodhart's Law and Machine Learning: A Structural Perspective. *International Economic Review*, 64(3):1075–1086.