Dario Amodei Jack Clark. 2016. Faulty reward functions in the wild. https://openai.com/research/faulty-reward-functions.

Yohan J. John, Leigh Caldwell, Dakota E. McCoy, and Oliver Braganza. 2023. Dead rats, dopamine, performance metrics, and peacock tails: proxy failure is an inherent risk in goal-oriented systems. *Behavioral and Brain Sciences*, pages 1–68.

Colm Kelly and Dennis J Snower. 2021. Capitalism recoupled. *Oxford Review of Economic Policy*, 37(4):851–863.

Daniel Koretz. 2008. *Measuring Up: What Educational Testing Really Tells Us*. Harvard University Press.

Victoria Krakovna and Janos Kramar. 2023. Power-seeking can be probable and predictive for trained agents. http://arxiv.org/abs/2304.06528.

Viktoria Krakovna. 2020. Specification gaming examples in AI. https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml.

Viktoria Krakovna, Jonathan Uesato, Vlad Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. 2020. Specification gaming: the flip side of AI ingenuity. https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity.

Lauro Langosco, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau, and David Krueger. 2023. Goal Misgeneralization in Deep Reinforcement Learning. http://arxiv.org/abs/2105.14111.

Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. 2017. AI Safety Gridworlds. http://arxiv.org/abs/1711.09883.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2023. Towards Out-Of-Distribution Generalization: A Survey. http://arxiv.org/abs/2108.13624.

Robert E. Lucas. 1976. Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1:19–46.

David Manheim. 2019. Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence. *Big Data and Cognitive Computing*, 3(2).

David Manheim and Scott Garrabrant. 2019. Categorizing Variants of Goodhart's Law. http://arxiv.org/abs/1803.04585.

Akhila Narla, Brett Kuprel, Kavita Sarin, Roberto Novoa, and Justin Ko. 2018. Automated Classification of Skin Lesions: From Pixels to Practice. *Journal of Investigative Dermatology*, 138(10):2108–2110.

Richard Ngo, Lawrence Chan, and Sören Mindermann. 2023. The alignment problem from a deep learning perspective. http://arxiv.org/abs/2209.00626.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

S O'Mahony. 2017. Medicine and the Mcnamara Fallacy. *Journal of the Royal College of Physicians of Edinburgh*, 47(3):281–287.

Toby Ord. 2020. *The Precipice*. Bloomsbury Publishing.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. http://arxiv.org/abs/2201.03544.