Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering Language Model Behaviors with Model-Written Evaluations. http://arxiv.org/abs/2212.09251.

Michael Poku. 2016. Campbell's Law: implications for health care. *Journal of Health Services Research & Policy*, 21(2):137–139.

Salvador Pueyo. 2018. Growth, degrowth, and the challenge of artificial superintelligence. *Journal of Cleaner Production*, 197:1731–1736.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2022. *Dataset Shift in Machine Learning*. MIT Press.

Stuart Russell. 2019. *Human Compatible: AI and the Problem of Control*. Allen Lane, London.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022a. Goal misgeneralization examples in AI. https://docs.google.com/spreadsheets/d/e/2PACX-1vTo3RkXUAigb25nP7gjpcHriR6XdzA_L5loOcVFj_u7cRAZghWrYKH2L2nU4TA_Vr9KzBX5Bjpz9G_l/pubhtml.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022b. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. http://arxiv.org/abs/2210.01790.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2023. Goal Misgeneralisation: Why Correct Specifications Aren't Enough For Correct Goals. https://deepmindsafetyresearch.medium.com/goal-misgeneralisation-why-correct-specifications-arent-enough-for-correct-goals-cf96ebc60924.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and Characterizing Reward Gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

Marilyn Strathern. 1997. 'Improving ratings': audit in the British University system. *European Review*, 5(3):305–321.

Wolfgang Stroebe. 2016. Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations. *Perspectives on Psychological Science*, 11(6):800–816.

Rachel L. Thomas and David Uminsky. 2022. Reliance on metrics is a fundamental challenge for AI. *Patterns*, 3(5).

Alexander Matt Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2023. Optimal Policies Tend to Seek Power. http://arxiv.org/abs/1912.01683.

Alexander Matt Turner and Prasad Tadepalli. 2022. Parametrically Retargetable Decision-Makers Tend To Seek Power. http://arxiv.org/abs/2206.13477.

Innovation UK Parliament's Science and Technology Committee. 2023. AI offers significant opportunities but twelve governance challenges must be addressed says science, innovation and technology committee. https://committees.parliament.uk/committee/135/science-innovation-and-technology-committee/news/197236/ai-offers-significant-opportunities-but-twelve-governance-challenges-must-be-addressed-says-science-innovation-and-technology-committee/.