Casey O. Barkan
Schmidt Science Fellows Program Application

References for Fellowship Research Proposal

[1] Katja Grace, Harlan Stewart, Julia Fabienne Sandk ühler, Stephen Thomas, *et al.*, "Thousands of AI authors on the future of AI," arXiv preprint arXiv:2401.02843 (2024).

[2] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, *et al.*, "Managing extreme AI risks amid rapid progress," Science 384, 842–845 (2024).

[3] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, *et al.*, "Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet," Anthropic (2024).

[4] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, *et al.*, "Escalation risks from language models in military and diplomatic decision-making," in The 2024 ACM Conference on Fairness, Accountability, and Transparency (2024) pp. 836–898.

[5] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey, "Sparse autoencoders find highly interpretable features in language models," arXiv preprint arXiv:2309.08600 (2023).

[6] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, *et al.*, "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971 (2023).

[7] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, *et al.*, "Mistral 7b," arXiv preprint arXiv:2310.06825 (2023).

[8] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, *et al.*, "Towards monosemanticity: Decomposing language models with dictionary learning," Transformer Circuits Thread 2 (2023).