Piantadosi, S. T. and F. Hill (2022). Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Ramsey, F. P. (1926). Truth and probability. *Histoy of Economic Thought Chapters*, 156–198.

Rogers, A., O. Kovaleva, and A. Rumshisky (2021). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics 8*, 842–866.

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Schervish, M. J., T. Seidenfeld, and J. B. Kadane (2002). Measuring incoherence. *Sankhyā: The Indian Journal of Statistics, Series A*, 561–587.

Shalev-Shwartz, S. and S. Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shanahan, M. (2022). Talking about large language models. *arXiv preprint arXiv:2212.03551*.

Staffel, J. (2020). *Unsettled thoughts: A theory of degrees of rationality*. Oxford University Press, USA.

Stiennon, N., L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems 33*, 3008–3021.

Templeton, A., T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan (2024). Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Thorpe, S. (1989). Local vs. distributed coding. *Intellectica 8*(2), 3–40.

Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases. *Science 185*(4157), 1124–1131.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM 27*(11), 1134–1142.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks 10*(5), 988–999.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Wald, A. (1947). An essentially complete class of admissible decision functions. *The Annals of Mathematical Statistics*, 549–555.

Ward, F., F. Toni, F. Belardinelli, and T. Everitt (2024). Honesty is the best policy: defining and mitigating ai deception. *Advances in Neural Information Processing Systems 36*.

Zhong, Z., Z. Liu, M. Tegmark, and J. Andreas (2024). The clock and the pizza: Two stories in mechanistic explanation of neural networks.

Advances in Neural Information Processing Systems 36.

Zhou, J., F. Chen, and A. Holzinger (2020). Towards explainability for ai fairness. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pp. 375–386. Springer.

Zhou, J. and T. Joachims (2023). How to explain and justify almost any decision: Potential pitfalls for accountability in ai decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 12–21.