Alain, G. and Y. Bengio (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Azaria, A. and T. Mitchell (2023). The internal state of an llm knows when it's lying.

Bai, Y., S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. (2022). Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623.

Bender, E. M. and A. Koller (2020). Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5185–5198.

Bricken, T., A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review 78*(1), 1–3.

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang (2023). Sparks of artificial general intelligence: Early experiments with gpt-4.

Burns, C., H. Ye, D. Klein, and J. Steinhardt (2022). Discovering latent knowledge in language models without supervision.

Campbell, J., R. Ren, and P. Guo (2023). Localizing lying in llama: Understanding instructed dishonesty on true-false questions through prompting, probing, and patching.

Christensen, D. (1991). Clever bookies and coherent beliefs. *The Philosophical Review 100*(2), 229–247.

Christiano, P. F., J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems 30*.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American journal of physics 14*(1), 1–13.

Cunningham, H., A. Ewart, L. Riggs, R. Huben, and L. Sharkey (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Dafoe, A., E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel (2020). Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*.

Davidson, D. (1970). Mental events. In L. Foster and J. W. Swanson (Eds.), *Experience and Theory*, pp. 79–101. Humanities Press.

Davidson, D. (1973). Radical interpretation. *Dialectica 27*(3/4), 313–328.

Davidson, D. (1974). On the very idea of a conceptual scheme. In *Proceedings and addresses of the American Philosophical Association*, Volume 47, pp. 5–20.

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, Volume 7, pp. 1–68.

Dong, Q., L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui (2022). A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory 52*(4), 1289–1306.

Elhage, N., T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, et al. (2022). Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Evans, O., O. Cotton-Barratt, L. Finnveden, A. Bales, A. Balwit, P. Wills, L. Righetti, and W. Saunders (2021). Truthful ai: Developing and governing ai that does not lie. *arXiv preprint arXiv:2110.06674*.

Farquhar, S., V. Varma, Z. Kenton, J. Gasteiger, V. Mikulik, and R. Shah (2023). Challenges with unsupervised llm knowledge discovery. *arXiv preprint arXiv:2312.10029*.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association 102*(477), 359–378.

Grote, T. (2021). Trustworthy medical ai systems need to know when they don't know. *Journal of medical ethics 47*(5), 337–338.

Grote, T. (2023). The allure of simplicity: On interpretable machine learning models in healthcare. *Philosophy of Medicine 4*(1).

Hammond, P. J. (1988). Consequentialist foundations for expected utility. *Theory and decision 25*, 25–78.

Harding, J. (2023). Operationalising representation in natural language processing. *arXiv preprint arXiv:2306.08193*.

Hedden, B. (2015). *Reasons without persons: Rationality, identity, and time*. OUP Oxford.

Ibeling, D., T. Icard, K. Mierzewski, and M. Mossé (2023). Probing the quantitative–qualitative divide in probabilistic reasoning. *Annals of Pure and Applied Logic*, 103339.

Jeffrey, R. C. (1990). *The logic of decision*. University of Chicago press.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science 65*(4), 575–603.

Kundu, S., Y. Bai, S. Kadavath, A. Askell, A. Callahan, A. Chen, A. Goldie, A. Balwit, A. Mirhoseini, B. McLean, et al. (2023). Specific versus general principles for constitutional ai. *arXiv preprint arXiv:2310.13798*.

Leitgeb, H. (2017). *The stability of belief: How rational belief coheres with probability*. Oxford University Press.

Levinstein, B. A. and D. A. Herrmann (2024). Still no lie detector for language models: Probing empirical and conceptual roadblocks. *Philosophical Studies*, 1–27.

Lewis, D. (1974). Radical interpretation. *Synthese*, 331–344.

Li, K., A. K. Hopkins, D. Bau, F. Viégas, H. Pfister, and M. Wattenberg (2023). Emergent world representations: Exploring a sequence model trained on a synthetic task.

Li, K., O. Patel, F. Viégas, H. Pfister, and M. Wattenberg (2023). Inference-time intervention: Eliciting truthful answers from a language model.

Mandelkern, M. and T. Linzen (2023). Do language models refer? *arXiv preprint arXiv:2308.05576*.

Marks, S. and M. Tegmark (2023). The geometry of truth: Emergent linear structure in large language model representations of true/false datasets.

Nanda, N., L. Chan, T. Lieberum, J. Smith, and J. Steinhardt (2023). Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Nanda, N., A. Lee, and M. Wattenberg (2023). Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*.

Olsson, C., N. Elhage, N. Nanda, N. Joseph, N. DasSarma, T. Henighan, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, S. Johnston, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah (2022). In-context learning and induction heads. *Transformer Circuits Thread*. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems 35*, 27730–27744.

Park, P. S., S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks (2024). Ai deception: A survey of examples, risks, and potential solutions. *Patterns 5*(5).

Patel, R. and E. Pavlick (2021). Mapping language models to grounded conceptual spaces. In *International conference on learning representations*.

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A 381*(2251), 20220041.

Pettigrew, R. (2016). *Accuracy and the Laws of Credence*. Oxford University Press.