REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.

Ashby, W. R. (1956). An Introduction to Cybernetics. Chapman & Hall.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus–norepinephrine function: Adaptive gain and optimal performance. Annual Review of Neuroscience, 28, 403–450.

Berger, J. O. (1985). Statistical Decision Theory and Bayesian Analysis. Springer.

Bewley, T. F. (2002). Knightian decision theory. Part I. Decisions in Economics and Finance, 25, 79–110.

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

Brachman, R. J., & Levesque, H. J. (2004). Knowledge Representation and Reasoning. Morgan Kaufmann.

Carver, C. S., & Scheier, M. F. (1998). On the Self-Regulation of Behavior. Cambridge University Press.

Chang, K.-W., et al. (2023). HELM: Holistic Evaluation of Language Models. Stanford Center for Research on Foundation Models.

Christiano, P., et al. (2017). Deep Reinforcement Learning from Human Preferences. Advances in Neural Information Processing Systems (NeurIPS).

Cover, T. M., & Thomas, J. A. (2006). Elements of Information Theory (2nd ed.). Wiley-Interscience.

Csiszár, I., & Körner, J. (2011). Information Theory: Coding Theorems for Discrete Memoryless Systems (2nd ed.). Cambridge University Press.

Duhem, P. (1954). The Aim and Structure of Physical Theory. Princeton University Press. (Original work published 1906).

Dunbar, R. (1998). The Social Brain Hypothesis. Evolutionary Anthropology, Wiley.

El-Yaniv, R. (2010). On the Foundations of Noise-Free Selective Classification. Journal of Machine Learning Research.

Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. Quarterly Journal of Economics, 75, 643–669.

Flavell, J. H. (1979). Metacognition and Cognitive Monitoring: A New Area of Cognitive–Developmental Inquiry. American Psychologist.

Floridi, L. (2011). The Philosophy of Information. Oxford University Press.

Fox, N. A., Henderson, H. A., Rubin, K. H., Calkins, S. D., & Schmidt, L. A. (2005). Continuity and discontinuity of behavioral inhibition and exuberance: Psychophysiological and behavioral influences across the first four years of life. Child Development, 76(1), 1–19.

Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? Nature Reviews Neuroscience, 11(2), 127–138.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. International Conference on Machine Learning (ICML).

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational Rationality: A Converging Paradigm for Intelligence in Brains, Minds, and Machines. Science, 349(6245), 273–278.

Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. Journal of Mathematical Economics, 18, 141–153.

Godfrey-Smith, P. (2016). Other Minds: The Octopus and the Evolution of Intelligent Life. Farrar, Straus and Giroux.

Grünwald, P. D. (2007). The Minimum Description Length Principle. MIT Press.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. International Conference on Machine Learning (ICML), 1321–1330.

Hansen, L. P., & Sargent, T. J. (2008). Robustness. Princeton University Press.

Harnad, S. (1990). The Symbol Grounding Problem. Physica D: Nonlinear Phenomena, 42(1–3), 335–346.

Hendrycks, D., & Gimpel, K. (2017). A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. International Conference on Learning Representations (ICLR).

Hendrycks, D., et al. (2019). Deep anomaly detection with outlier exposure. International Conference on Learning Representations (ICLR).

Hendrycks, D., et al. (2021). Unsolved Problems in ML Safety. arXiv preprint.

Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press.

Kagan, J. (1972). Motives and Development. Yale University Press.

Kagan, J. (1994). Galen's Prophecy: Temperament in Human Nature. Basic Books.

Kagan, J. (2007). Temperament as a Biological Phenomenon. Guilford Press.

Kagan, J., & Snidman, N. (1997). Early childhood predictors of adult anxiety disorders. Biological Psychiatry, 41(3), 183–189.

Kant, I. (1781). Critique of Pure Reason. Hartknoch.

Kaplan, J., et al. (2020). Scaling laws for neural language models. arXiv:2001.08361.

Kendall, A., & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? Advances in Neural Information Processing Systems (NeurIPS).

Knight, F. H. (1921). Risk, Uncertainty, and Profit. Houghton Mifflin.

Knill, D. C., & Richards, W. (1996). Perception as Bayesian Inference. Cambridge University Press.

Kolmogorov, A. N. (1965). Three Approaches to the Quantitative Definition of Information. Problems of Information Transmission, 1(1), 1–7.

Kruska, D. (1988). Mammalian Domestication and Its Effect on Brain Structure and Behavior. Brain, Behavior and Evolution. Karger.

Lakatos, I. (1976). Falsification and the Methodology of Scientific Research Programmes. Cambridge University Press.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. Advances in Neural Information Processing Systems (NeurIPS).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521, 436–444.

Liang, P., et al. (2022). Holistic Evaluation of Language Models. Transactions on Machine Learning Research.

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. Association for Computational Linguistics (ACL).

Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. MIT Press.

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On Faithfulness and Factuality in Abstractive Summarization. Association for Computational Linguistics (ACL).

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. International Conference on Machine Learning (ICML), 625–632.

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. Advances in Neural Information Processing Systems (NeurIPS).

Ovadia, Y., Fertig, E., Ren, J., et al. (2019). Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. Advances in Neural Information Processing Systems (NeurIPS).

Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press.

Popper, K. (1959). The Logic of Scientific Discovery. Routledge.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2009). Dataset Shift in Machine Learning. MIT Press.

Quine, W. V. O. (1951). Two dogmas of empiricism. Philosophical Review, 60, 20–43.

Ranzato, M., Chopra, S., Auli, M., & Zaremba, W. (2016). Sequence Level Training with Recurrent Neural Networks. International Conference on Learning Representations (ICLR).

Rissanen, J. (1989). Stochastic Complexity in Statistical Inquiry. World Scientific.

Russell, S., & Norvig, P. (2021). Artificial Intelligence: A Modern Approach (4th ed.). Pearson.

Savage, L. J. (1954). The Foundations of Statistics. Wiley.

Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379–423; 623–656.

Simon, H. A. (1957). Models of Man: Social and Rational. Wiley.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems (NeurIPS).

Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.

Tishby, N., Pereira, F. C., & Bialek, W. (1999). The Information Bottleneck Method. Proceedings of the Allerton Conference on Communication, Control, and Computing.

Tishby, N., & Zaslavsky, N. (2015). Deep Learning and the Information Bottleneck Principle. IEEE Information Theory Workshop.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605.

Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS).

Wenger, K., Hossein Abadi, K., Fozard, D., Tirdad, K., Dela Cruz, A., & Sadeghian, A. (2023). A Novel Application of XAI in Squinting Models: A Position Paper. SSRN Electronic Journal (preprint).

Yao, S., Zhao, J., Yu, D., et al. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. International Conference on Learning Representations (ICLR).

Zhang, C., et al. (2017). Understanding deep learning requires rethinking generalization. International Conference on Learning Representations (ICLR).