

# Explainability

Readings on what is meant by “explainability” (and related terms like “transparency” and “interpretability”) in data science and to what extent achieving explainability (or transparency or interpretability) in algorithms is morally important.

<b>Title</b>	<b>Citation</b>
Transparency in Complex Computational Systems	Creel ( <a href="#">2020</a> )
How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms	Burrell ( <a href="#">2015</a> )
Explainable AI: A Review of Machine Learning Interpretability Methods	Linardatos, Papastefanopoulos, & Kotsiantis ( <a href="#">2020</a> )
Transparency’s Ideological Drift	Pozen ( <a href="#">2018</a> )
Philosophy of Science at Sea: Clarifying the Interpretability of Machine Learning	Beisbart & R��z ( <a href="#">2022</a> )
The Right to an Explanation	Vredenburg ( <a href="#">2021</a> )
Algorithmic and Human Decision Making: For a Double Standard of Transparency	G��nther & Kasirzadeh ( <a href="#">2022</a> )
The Mythos of Model Interpretability	Lipton ( <a href="#">2016</a> )
Epistemic Values in Feature Importance Methods: Lessons from Feminist Epistemology	Hancox-Li & Kumar ( <a href="#">2021</a> )
The Fate of Explanatory Reasoning in the Age of Big Data	Cabrera ( <a href="#">2020</a> )
Interpreting Interpretability: Understanding Data Scientists’ Use of Interpretability Tools for Machine Learning	Kaur et al. ( <a href="#">2020</a> )
“Explaining” Machine Learning Reveals Policy Challenges	Coyle & Weller ( <a href="#">2020</a> )
Why Should I Trust You? Explaining the Predictions of Any Classifier	Ribeiro, Singh, & Guestrin ( <a href="#">2016</a> )

## References

- Beisbart, C., & Rätz, T. (2022). Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass*, 17(6), e12830. <https://doi.org/10.1111/phc3.12830>
- Burrell, J. (2015). How the machine 'thinks:' understanding opacity in machine learning algorithms. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2660674>
- Cabrera, F. (2020). The fate of explanatory reasoning in the age of big data. *Philosophy & Technology*, 34(4), 645–665. <https://doi.org/10.1007/s13347-020-00420-9>
- Coyle, D., & Weller, A. (2020). "Explaining" machine learning reveals policy challenges. *Science*, 368(6498), 1433–1434. <https://doi.org/10.1126/science.aba9647>
- Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, 87(4), 568–589. <https://doi.org/10.1086/709729>
- Günther, M., & Kasirzadeh, A. (2022). Algorithmic and human decision making: For a double standard of transparency. *AI and Society*, 37(1), 375–381. <https://doi.org/10.1007/s00146-021-01200-5>
- Hancox-Li, L., & Kumar, I. E. (2021). Epistemic values in feature importance methods: Lessons from feminist epistemology. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 817–826. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445943>
- Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3313831.3376219>
- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23, 18. <https://doi.org/10.3390/e23010018>
- Lipton, Z. (2016). The mythos of model interpretability. *Communications of the ACM*, 61. <https://doi.org/10.1145/3233231>
- Pozen, D. E. (2018). Transparency's ideological drift. *Yale Law Journal*, 128, 100–165.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938. Retrieved from <http://arxiv.org/abs/1602.04938>
- Vredenburg, K. (2021). The right to explanation. *Journal of Political Philosophy*, 30(2), 209–229. <https://doi.org/10.1111/jopp.12262>