

Pragmatic Explainable Machine Learning

Sakina Hansen

London School of Economics and Political Science
s.a.hansen1@lse.ac.uk

Research Summary

This extended abstract describes my PhD research, which addresses the research problem of how to evaluate and design explainable artificial intelligence (XAI) methods for use in high-stakes settings. My research is motivated by the observation that while a wide range of XAI methods are now deployed across disciplines, there is no clear way to determine which method is appropriate for a given context, where the wrong choice can produce misleading interpretations with serious consequences.

My PhD comprises several interdisciplinary and multi-method projects united by a pragmatic, context-aware approach to XAI and AI ethics. The philosophy strand of this work develops a pragmatic-rhetorical account of explainability, in which explanation methods are assessed not by fidelity alone but relative to specific goals and rhetorical situations. A statistics strand applies this pragmatic orientation to algorithmic fairness, where I demonstrate that there are causal requirements that non-causal XAI techniques cannot generally satisfy. To address this, my co-authors and I develop causal dependence plots (CDPs), establish theoretical results for non-causal explanation methods, and introduce fairness detectability as a quantitative diagnostic for evaluating explanations against causal baselines. These contributions are validated through experiments. Planned future work will apply these methods to analyse sentencing data from Crown and Magistrates Courts in England and Wales, illustrating how causal and goal-aligned explainability can support statistically robust and policy-relevant reasoning about ethnic, gender, and class disparities. A third strand extends the pragmatic perspective to AI ethics more broadly through qualitative research on how AI risks are represented in media discourse and perceived by practitioners in organizational contexts, viewing AI ethics as a situated and context-dependent practice.

This work connects directly to several AIES themes. It contributes to ethical theories, models and frameworks for AI and data through the pragmatic-rhetorical account, which offers a normative basis for evaluating XAI methods relative to goals and contexts. It advances work on interpretability, explainability, and transparency through causal dependence

plots and fairness detectability as methodological tools. It further engages with the themes of AI governance and power through qualitative research on AI discourse and AI in organizational contexts.

Research Questions & Contributions

My PhD is structured around three research questions: (RQ1) How should we evaluate and choose which XAI method to use for a given goal or task?

(RQ2) What are the limitations of non-causal XAI, and in what circumstances should we prefer causal over non-causal methods?

(RQ3) How is XAI, and AI ethics more generally, shaped and perceived across different areas of society, and how does this influence AI governance?

Methodologically, my work contributes:

- *Causal dependence plots (CDPs)*: an XAI method that directly incorporates causal knowledge (Loftus, Bynum, and Hansen 2024)
- *Fairness detectability*: a metric and visualisation tool for evaluating whether an XAI method is appropriate

Empirically, my work contributes

- Simulation experiments and theoretical results that show that non-causal XAI methods are often (1) blind to indirect unfairness when the predictor is invariant and (2) incapable of identifying a specific causal effect
- Qualitative findings about a) the differing ways charity workers understand AI ethics and b) how AI risks are represented in right-wing news media

Conceptually and normatively my work contributes:

- A pragmatic-rhetorical account of XAI that evaluates XAI methods against how well they enable context-specific goals
- The concept of a red-techno scare — a contemporary form of technologically mediated anti-communism — and an alternative discourse centred on AI for human flourishing

Progress to Date

Philosophical Research

Explaining the outputs of machine learning (ML) systems has become a central concern as these technologies increas-

ingly inform scientific inquiry and social decision-making. While growing critiques highlight the limitations of existing XAI methods, including risks of misleading interpretation and misuse, how XAI should be deployed to support responsible reasoning remains an open question. Recent work has proposed that the value of XAI methods arises from their capacity to provide understanding through idealized models (Fleisher 2022; Sullivan 2024). Building on this insight, I combine an idealization account with Jan Faye’s pragmatic theory of explanation (Faye 2007) to develop a pragmatic-rhetorical theory of XAI. This account proposes that explanations should be evaluated by how well they achieve context-specific goals within particular rhetorical situations, and that XAI should therefore be assessed by the effects of using it in line with pragmatic goals and criteria, not solely by its epistemic properties. On this basis, I introduce a conceptual framework for goal-oriented XAI. The framework distinguishes three central explanatory goals (hypothesis generation, evidence corroboration, and intervention) and provides structured questions for evaluating XAI methods relative to these goals. Together, these contributions offer a normative foundation for assessing when and how XAI can meaningfully support scientific and social reasoning.

Data Science Research

With my co-authors, we developed causal dependence plots (CDPs), a feature importance method which instead of assuming independence between features, uses a causal model to measure the impact of a particular variable on the prediction (Loftus, Bynum, and Hansen 2024). This differs from standard methods based on independence or holding other predictors constant, such as partial dependence plots. CDPs include a variety of interpretive plots showing total, direct, and indirect causal effects. We demonstrate with simulations and real data experiments how CDPs can be combined modularly with methods for causal learning or sensitivity analysis. To develop the relationship between XAI and algorithmic fairness, we categorize three goals of XAI for fairness: establishing correct epistemic footing, enabling algorithm critique, and enabling change through intervention or policy. Framed by these goals, we use data simulations with popular causal and non-causal XAI methods to highlight limitations through knowledge of the underlying data generating process, develop impossibility propositions for non-causal XAI based on these results, and consequently propose fairness detectability to measure the performance of a particular XAI method with respect to a baseline causal effect.

Qualitative Research

I examine Fox News’ discourse on AI risks since ChatGPT’s release using a bespoke discourse analysis methodology. I present empirical evidence that Fox News represents AI risk primarily as a Chinese threat to the nation, as ideological overreach by “the Left,” and as an individual’s responsibility. I argue that these discursive patterns constitute what I term a red-techno scare, a contemporary form of anti-communism. This discourse shapes the scope of what are considered legitimate governance responses. Approaching the analysis from

a normative and reflexive lens, I propose an alternative discourse for AI risks focused on AI for human flourishing (Hansen 2026). I also conducted a pilot study consisting of semi-structured interviews with employees who work with data or technology at three UK-based charities. I found that they view AI’s ethical risks primarily through its potential to be an exclusionary tool, and its benefits primarily through its potential as an efficiency tool (Hansen 2025). Lastly, in a co-design project for a non-profit that supports survivors of gender-based violence, colleagues and I developed responsible AI guidance as they sought to translate its value-driven approach into LLM-enabled products (Uleah et al. 2025).

Future Work & Impact

The final chapter will apply the methods developed in this research to data from the Crown and Magistrates Courts in England and Wales. The goal is to measure ethnic, gender, and class disparities in sentencing between 2011 and 2023. This work serves both as an applied stress test for the methods developed in the thesis and as a contribution to data-driven criminology, addressing calls for ML methods suited to complex criminological data (Brennan and Oliver 2013).

Together, these strands of research demonstrate the value of interdisciplinary approaches to AI ethics. The combination of philosophical grounding, technical rigour, and empirical engagement with how AI is understood outside of academia informs novel responsible AI practices.

References

- Brennan, T.; and Oliver, W. L. 2013. The Emergence of Machine Learning Techniques in Criminology: Implications of Complexity in our Data and in Research Questions. *Criminology & Public Policy*, 12(3): 551–562.
- Faye, J. 2007. The pragmatic-rhetorical theory of explanation. In *Rethinking explanation*, 43–68. Springer.
- Fleisher, W. 2022. Understanding, Idealization, and Explainable AI. *Episteme*, 19(4): 534–560.
- Hansen, S. 2025. Exclusion or efficiency: understanding perspectives about AI ethics among charity workers in the United Kingdom. *Proceedings of Machine Learning Research*, 294.
- Hansen, S. 2026. “Woke AI is a big threat to free speech”: Red-Techno Scares in AI Risk Discourse on Fox News. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, 1607–1629.
- Loftus, J. R.; Bynum, L. E.; and Hansen, S. 2024. Causal dependence plots. *Advances in Neural Information Processing Systems*, 37: 112656–112683.
- Sullivan, E. 2024. SIDEs: Separating Idealization from Deceptive ‘Explanations’ in xAI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1714–1724. Rio de Janeiro Brazil: ACM. ISBN 979-8-4007-0450-5.
- Uleah, K.; Worth, S.; Blum-Dumontet, E.; Hansen, S.; Jorgensen, M.; Khor, A.; Krishnamurthy-Spencer, N.; ; Russo, M.; and Scott, K. 2025. From principles to practice, CHAYN’s guide to building a feminist AI: A framework for building trauma-informed, survivor-centered LLM products.