

# Explainable AI methods and their interplay with privacy protection

Francesca Naretto

University of Pisa, Italy, francesca.naretto@unipi.it

## Extended Abstract

In recent years, Machine Learning (ML) models have achieved remarkable predictive accuracy, leading to their adoption in various domains, such as health care, bank related services, and even autonomous driving cars. However, these models are often referred to as black-boxes since they lack transparency in their decision-making processes. This opacity raises concerns regarding trustworthiness, a fundamental requirement, emphasized in regulations such as the Artificial Intelligence (AI) Act in Europe, as well as policies in China, Japan, and the United States [1, 2, 3]. To develop trustworthy AI systems, it is crucial to interpret and understand how ML models reach their decisions. This challenge is addressed by Explainable Artificial Intelligence (xAI), a field that has gained increasing attention but still faces numerous open challenges, particularly in ensuring stable and reliable explanations.

At the same time, there is another ethical concern: privacy. Since the introduction of the GDPR [4], data privacy has been approached from multiple perspectives. In particular, researchers have been investigating how to provide access to data, allowing for pattern extraction and knowledge discovery tasks, while safeguarding individuals' privacy. Technically, various privacy attacks have been proposed, leading to the development of privacy protection methodologies designed to defend against these threats. Nowadays, there are also privacy attacks targeting ML models. Even when training data remains private, adversaries can exploit model queries to infer sensitive information, such as membership inference or attribute disclosure, posing significant risks [5].

Given the importance of both xAI and Privacy in ML, my Ph.D. thesis explores the intersection of explainability and privacy, providing the following two key contributions:

- A novel variant of a local rule-based explanation method that enhances the stability and actionability of explanations.
- A study on the synergies and tensions between data privacy and explainability, analyzing how explanations can both enhance privacy awareness and, conversely, expose sensitive information.

## Explainability: Stability and Actionability in xAI

The first part of my thesis focuses on improving xAI methods. A comprehensive survey of post-hoc explanation techniques for tabular data, images and text was conducted, analyzing state-of-the-art approaches from both theoretical and practical perspectives [6]. This study tested various metrics on publicly available methods to assess the quality of explanations, highlighting key limitations such as instability and lack of robustness. Many explainability techniques introduce randomness, causing explanations for the same record to vary across runs, hence reducing user trust in black-box models.

To address this issue, my thesis introduces an enhanced local rule-based explanation method that generates stable and actionable explanations for tabular data [7]. This approach uses factual logic rules to explain model decisions and counterfactual logic rules to suggest minimal modifications needed to change a ML model's output. The methodology relies on decision trees trained on synthetically generated neighbors, exploiting a genetic algorithm, ensuring local interpretability while maintaining



high fidelity to the original model. Experimental results confirm that this approach provides more stable and reliable explanations compared to existing methods.

## Privacy and Explainability: A Dual Perspective

The second part of the thesis examines the relationship between privacy and explainability, considering both privacy-aware explanations and privacy risks introduced by explanations.

**Enhancing Privacy Awareness Through Explainability.** This thesis presents EXPERT[8, 9, 10], a framework that predicts privacy risks and provides explanations for individuals' data exposure. Existing privacy risk evaluation methods are computationally expensive, requiring frequent re-evaluation when new data arrives. Instead, EXPERT uses ML models to classify individuals as high-risk or low-risk in terms of privacy, enabling real-time assessments. Although complex black-box models are used for classification, explainability techniques are applied post-hoc to interpret privacy risk factors. The framework is validated on human mobility data, providing visual explanations that highlight risk areas on a map, assisting both end-users in understanding their exposure and data providers in evaluating privacy risks at a population level.

**Privacy Risks of Explainability Methods.** Explainability methods can also compromise privacy by revealing patterns learned by ML models. A key example is the Membership Inference Attack (MIA) [5], which determines whether a given record was part of the model's training set. While existing MIA techniques require access to probability vectors or dataset statistics, this thesis introduces ALOA [11], a membership attack that is agnostic to both, making it more practical and effective in real-world settings. Experimental results demonstrate that ALOA achieves privacy exposure levels comparable to or higher than state-of-the-art attacks, even with fewer assumptions.

This thesis examines also how explainability itself can be exploited in privacy attacks. Many explanations rely on surrogate models, which may inadvertently expose sensitive information. To assess this risk, this work introduces REVEAL[12], a framework for evaluating privacy exposure in global and local explainers. Results show that global explainers pose a significantly higher privacy risk than local ones, emphasizing the need for privacy-aware XAI approaches.

## Conclusions and Future Directions

This work highlights the complex interplay between explainability and privacy, demonstrating that explanations can be both a tool for privacy awareness and a potential privacy threat. The findings contribute to the ongoing development of trustworthy AI by proposing solutions that improve explanation stability while also addressing privacy risks.

Future work will explore privacy-preserving explainability methods to ensure that AI systems remain both interpretable and secure. This includes investigating techniques such as differentially private explanations and adversarial robustness against attacks targeting explainability models. Expanding the evaluation to additional domains, beyond mobility and tabular data, would further validate the generalizability of the proposed solutions.

## Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

## References

- [1] E. Union, The artificial intelligence act, <https://artificialintelligenceact.eu/the-act/>, ????
- [2] S. C. on Artificial Intelligence, The national artificial intelligence research and development strategic plan: 2019 update, in: Executive Office of the President of the United States, Curran Associates, Inc., 2019.

- [3] China, New generation artificial intelligence development plan, <https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/>, ????
- [4] E. Union, The general data protection regulation, <https://www.garanteprivacy.it/>, ????
- [5] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE Symposium on Security and Privacy (SP), 2017.
- [6] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, *Data Mining and Knowledge Discovery Journal* (2023). doi:<https://doi.org/10.1007/s10618-023-00933-9>.
- [7] R. Guidotti, A. Monreale, S. Ruggieri, F. Naretto, F. Turini, D. Pedreschi, F. Giannotti, Stable and actionable explanations of black-box models through factual and counterfactual rules, *Data Mining and Knowledge Discovery* (2024). doi:<https://doi.org/10.1007/s10618-022-00878-5>.
- [8] F. Naretto, R. Pellungrini, D. Fadda, S. Rinzivillo, Exphlot: Explainable privacy assessment for human location trajectories, in: *Discovery Science*, 2023. doi:[https://doi.org/10.1007/978-3-031-45275-8\\_22](https://doi.org/10.1007/978-3-031-45275-8_22).
- [9] F. Naretto, R. Pellungrini, F. M. Nardini, F. Giannotti, Prediction and explanation of privacy risk on mobility data with neural networks, in: *ECML PKDD 2020 Workshops*, Springer International Publishing, 2020. doi:<https://doi.org/10.1007/978-3-030-65965-3\34>.
- [10] F. Naretto, R. Pellungrini, A. Monreale, F. M. Nardini, M. Musolesi, Predicting and explaining privacy risk exposure in mobility data, in: *Discovery Science - 23rd International Conference*, DS 2020, Thessaloniki, Greece, October 19-21, 2020, Proceedings, *Lecture Notes in Computer Science*, Springer, 2020.
- [11] A. Monreale, F. Naretto, S. Rizzo, Agnostic label-only membership inference attack, in: *17th International Conference on Network and System Security*, Springer, 2023.
- [12] F. Naretto, A. Monreale, F. Giannotti, Evaluating the privacy exposure of interpretable global and local explainers, in: *Transactions on Data Privacy*, volume 18, 2025.