Conceptualisation and Implementation of Human-centric Privacy Preserving Framework for Explainable AI

Sonal Allana

School of Computer Science, University of Guelph
sallana@uoguelph.ca

Abstract

Explainability has emerged as a pillar of Trustworthy AI for ensuring safety in high-risk application domains. However, the incorporation of explainability to boost the transparency of black-box AI systems can inadvertently introduce unforeseen vulnerabilities. Previous research has drawn attention to privacy leakage, malicious or otherwise, from explainable interfaces leading to identification of individuals and exposure of sensitive personal information. Privacy preservation methods used in response to this leakage are found to adversely affect the utility of the system, including the degradation of model accuracy and explanation quality. The proposed thesis will examine the advancement of Privacy Enhancing Technologies (PETs) in Explainable AI (XAI) while ensuring that users remain at the core of the design process. The main objectives of this research are: (1) determining defenses for privacy attacks in XAI (2) building interpretable algorithms for private models and (3) examining user requirements for privacy preserving XAI. This research is expected to yield characteristics of privacy preserving XAI, guidelines and recommendations for effectively building privacy compliant XAI while considering the diverse needs of end users. The research outcomes will enable developers and researchers in designing XAI that is safe for deployment and considers the balance between privacy, explainability and utility.

Introduction

AI systems are increasingly ubiquitous, amid widespread adoption in a wide range of industries. In contrast to traditional software, where the logic underlying the outputs are traceable, the outcomes of black-box AI systems are difficult to interpret, even for their own developers. Consequently, the field of Explainable AI (XAI) has seen momentum as a means of bringing transparency to black-boxes. The current literature on Trustworthy AI (Tabassi 2023) identifies several key desiderata for the development of safe AI systems, including validity, robustness, explainability, fairness, and privacy, among others. Although these principles are intended to collectively enhance the overall safety of these systems, inherent tensions may arise between certain desired properties. One such conflict is seen between the properties of explainability and privacy (Shokri, Strobel, and Zick 2021). Explainability can support privacy in several ways, such as in determining if the privacy requirements of a system are met (Doshi-Velez and Kim 2017). Conversely, it may also expose systems to additional privacy risks.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Research Motivation

The additional information provided through XAI interfaces can introduce intentional privacy leaks through adversaries or unintentional leaks through legitimate users. Researchers have demonstrated various privacy attacks exploiting explanations that target training data, sensitive information contained in query inputs or the intellectual property of model owners. Despite this vulnerability, research on fortifying XAI against such attacks remains limited. A comprehensive understanding of privacy preservation methods applicable to different categories of explanations, as well as standardized baselines for privacy assessment in XAI, is currently lacking. Current XAI methods are predominantly model-centric, often overlooking the objectives and goals of end users (Kaplan, Uusitalo, and Lensu 2024). Certain XAI methods are technically complex for layman users and primarily designed to satisfy the needs of researchers and developers rather than those of end users. Research is also lacking in balancing privacy and explainability with other system properties such as utility (Harder, Bauer, and Park 2020). Thus, the conflicting outcomes of including explainability as a non-functional requirement in AI systems, necessitate strategies that balance it with other desiderata in the overall AI system.

Thesis Statement

Since explainability is found to have an adverse impact on the privacy of an AI system, the proposed thesis aims at advancing privacy preservation in explainable methods and gaining an in-depth understanding of privacy threats in the landscape of XAI methods. This goal should be pursued while maintaining a user-centred design approach and ensuring acceptable levels of system utility. Thus the thesis statement for the proposed work is as follows:

Privacy, explainability and utility can co-exist in a human-centric XAI system despite their conflicting relationships.

This thesis statement indicates that although transparency measures can introduce privacy risks and privacy preservation methods may impact system utility, it is possible to achieve a balance between the triad of privacy, explainability and utility. This process should be guided by the active involvement of users, considering their needs, requirements and feedback throughout the design and implementation cycle, thereby realizing the objective of human-centricity.

Research Objectives

The long-term objective of this thesis is the advancement of privacy preservation in XAI systems and the establishment of a human-centric privacy preserving XAI framework. To achieve this, we have identified the following 3 short-term objectives:

- Advancing privacy preservation in post-hoc XAI to mitigate exposure of sensitive information through explanations.
- Providing interpretability to private black-box ensemble models and providing tuning parameters for balancing different properties.
- Determining requirements of explanations that are private, accurate and usable by end users.

The first objective considers an existing XAI-aware privacy attack that exposes sensitive information of users through explanations. We identify suitable countermeasures for the selected attack through empirical evaluation. The second objective considers an existing private ensemble algorithm and introduces an interpretability layer to the black-box. Tuning parameters are integrated to adjust the levels of privacy, explainability and utility. The third objective will gather user perspectives on private explanations designed through the first two objectives and its impact on users' ability to complete their tasks. This will help identify the needs of users and elicit feedback on the proposed privatisation methods of post-hoc XAI and the interpretability of private ensemble models.

Research Progress and Timeline

For the proposed thesis, I have completed a scoping review of current literature (Allana, Kankanhalli, and Dara 2025) to determine the privacy risks in XAI and the proposed defenses. The scoping review, based on PRISMA-ScR, is executed using the following two research questions:

- What are the privacy risks of releasing explanations in AI systems?
- What current methods have researchers employed to achieve privacy preservation in XAI systems?

Based on the knowledge synthesised from the review, I have identified and proposed with the help of my supervisor, the characteristics of privacy preserving XAI to lay the foundation of the proposed framework. Thereafter, I have designed and implemented the experiments corresponding to the first objective (Allana et al. 2025). Currently, I am working towards completing the experiments corresponding to the second objective in collaboration with an industry partner. The role of the industry partner is to provide advice on the integration of their proposed private algorithm in the system. After completing the experiments and analysing the results, I will commence work on the planning of the third objective to produce a rough draft of the aims and objectives of the user study. The discussion with mentors and peers at the doctoral consortium on the outcomes of the completed experiments, goals of the user study and the integration of the three objectives under the umbrella of the thesis statement, will improve the quality of the research and lay the direction for future research in this area. I expect to complete work on the thesis and prepare to defend towards the end of 2026.

Expected Contributions

Towards achieving the long-term objective of the thesis, my proposed work will lay the foundation of the human-centric privacy preserving XAI framework comprising of baselines for privacy measurement of different XAI categories, characteristics of privacy preserved explanations, recommendations for integrating appropriate PETs with specific methods and guidelines for balancing privacy, utility and explanation quality. It will build a knowledge-base of PETs suitable for specific types of XAI methods and strategies for integration, resulting in XAI that is safe for deployment in critical domains.

Acknowledgements

This research is funded by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant awarded to my advisor, Dr. Rozita Dara.

References

Allana, S.; Dara, R.; Lin, X.; and Xiong, P. 2025. Towards integration of Privacy Enhancing Technologies in Explainable Artificial Intelligence. *Submitted to Elsevier Knowledge Based Systems*. Under review.

Allana, S.; Kankanhalli, M.; and Dara, R. 2025. Privacy Risks and Preservation Methods in Explainable Artificial Intelligence: A Scoping Review. *Transactions on Machine Learning Research*.

Doshi-Velez, F.; and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning. ArXiv:1702.08608 [cs, stat].

Harder, F.; Bauer, M.; and Park, M. 2020. Interpretable and Differentially Private Predictions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 4083–4090.

Kaplan, S.; Uusitalo, H.; and Lensu, L. 2024. A unified and practical user-centric framework for explainable artificial intelligence. *Knowledge-Based Systems*, 283: 111107.

Shokri, R.; Strobel, M.; and Zick, Y. 2021. On the Privacy Risks of Model Explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 231–241. Virtual Event USA: ACM. ISBN 978-1-4503-8473-5.

Tabassi, E. 2023. AI Risk Management Framework: AI RMF (1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD.