Privacy Risks and Preservation Methods in Explainable Artificial Intelligence: A Scoping Review

Anonymous authors
Paper under double-blind review

Abstract

Explainable Artificial Intelligence (XAI) has emerged as a pillar of Trustworthy AI and aims to bring transparency in complex models that are opaque by nature. Despite the benefits of incorporating explanations in models, an urgent need is found in addressing the privacy concerns of providing this additional information to end users. In this article, we conduct a scoping review of existing literature to elicit details on the conflict between privacy and explainability. Using the standard methodology for scoping review, we extracted 57 articles from 1,943 studies published from January 2019 to December 2024. The review addresses 3 research questions to present readers with more understanding of the topic: (1) what are the privacy risks of releasing explanations in AI systems? (2) what current methods have researchers employed to achieve privacy preservation in XAI systems? (3) what constitutes a privacy preserving explanation? Based on the knowledge synthesized from the selected studies, we categorize the privacy risks and preservation methods in XAI and propose the characteristics of privacy preserving explanations to aid researchers and practitioners in understanding the requirements of XAI that is privacy compliant. Lastly, we identify the challenges in balancing privacy with other system desiderata and provide recommendations for achieving privacy preserving XAI. We expect that this review will shed light on the complex relationship of privacy and explainability, both being the fundamental principles of Trustworthy AI.

1 Introduction

1.1 Paradigm shift in technology and the need for explanations

Traditional software development processes have metamorphosed into stable and reliable frameworks through decades of fine tuning by software experts. These software systems are built on human designed algorithms and produce a trace of the logic used to generate an output. Even in complex systems, it is possible for software experts to analyze the logic and generate an explanation for a specific result. During the software development lifecycle, engineers focus on creating the algorithm and validating using well designed test cases that closely replicate real world scenarios. In contrast, modern AI systems do not have an underlying human-written algorithm and learn from data fed to them. This data-driven nature creates dependence of the system on data quality (Merhi, 2022) and introduces problems such as lack of fairness when data is biased, or irrelevant results when data is incomplete or outdated (Trocin et al., 2021). During the AI development phase, engineers access training datasets which may contain personally identifiable or sensitive information about individuals. For neural network systems, the development process often involves a trial-and-error approach, where high accuracy is targeted by tweaking the hyperparameters such as the learning rate, epochs, number of layers or activation functions. The lack of an algorithm prevents engineers from tracing through the AI system and interpreting the results. Thus, the basic ability to be explainable and understand input-output behaviors, which is critical to all computer systems (Sundararajan et al., 2017), is often out of reach of AI systems. Explanations for outcomes of AI are crucial in high-risk applications (Mochaourab et al., 2023) in domains such as healthcare (Dhar et al., 2023; Dwivedi et al., 2023; Yang et al., 2022), finance (Dwivedi et al., 2023; Zhang et al., 2022), defense (Dwivedi et al., 2023), justice (Deeks, 2019), energy and power (Machlev et al., 2022) where the impact on human life and well-being is significant (Karimi et al., 2023; McDermid et al., 2021; Nassar et al., 2020) and the inability to do so deters their successful implementation (Nassar et al., 2020; Shrikumar et al., 2017a; Yang et al., 2022).

Trustworthy AI strives to mitigate risks due to possible harms from the data-driven nature of AI systems. Trustworthiness is based on foundation principles of reliability, validity, robustness, privacy, explainability and fairness (Alzubaidi et al., 2023; Tabassi, 2023) to boost user confidence in the system outputs. Among these principles, explainability aims to bring the much-needed transparency in opaque models and can be considered as a non-functional requirement of a software system to mitigate opacity (Chazette et al., 2021). There are numerous benefits of including explanations in AI models. Besides aiding data scientists in getting a better understanding of the data (Hohman et al., 2019) and performing required data cleansing (Chen et al., 2022b), explanations can help developers in detecting errors in input and determining features that can be modified to change the outcome (Datta et al., 2016). When multiple models are available with similar accuracy, an explanation method can help to choose between models (Dhurandhar et al., 2018). Interpretable models can enable knowledge discovery by detecting knowledge or patterns that were missed by uninterpretable ones (Kim et al., 2016). Since humans remain an important component in the decision-making process as end-users and consumers of automated decisions (Terziyan & Vitko, 2022), explanations can give them an understanding of the model outcome, especially when they are adversely affected by the decisions (Ali et al., 2023). Explainability can also facilitate privacy awareness in end-users (Brunotte et al., 2021), enabling them to make right choices for their personal data and aid regulators and compliance officers to understand the compliance of models (McDermid et al., 2021) with applicable regulations. With generative AI (Gen-AI) and large language models (LLMs) entering mainstream, explanations constitute an important design principle (Weisz et al., 2023) in enabling a better mental model for users (Sun et al., 2022) and in communicating its capabilities and limitations to them (Weisz et al., 2023). It can also support users in effective prompt engineering to determine the words that impact the output of a model (Mishra et al., 2023) and in verification of generated content to mitigate the problem of hallucinations (Schneider, 2024).

1.2 Challenges for privacy in explainability

In many high risk application domains of AI, training models on sensitive personal information is inevitable for usefulness of these systems (Veugen et al., 2022). For instance, a lung cancer detection model necessitates training on chest X-ray images, which constitutes personal information of patients. Similarly, a loan evaluation model of a bank, requires access to the financial profiles of customers, which is also personal information of individuals. Usage of personal data impacts the privacy of individuals when they are subject to intentional or unintentional identification and exposure through these systems. Some models are found to memorize data contained in the input (Song et al., 2017) which can be exploited by adversaries for extraction of personal information. Gen-AI models create new content from large multi-modal datasets (Sun et al., 2022) which could potentially contain sensitive personal information (Meskó & Topol, 2023). Due to such privacy risks involved, when personal data is used in training, testing, or inferencing of AI models, they become subject to data regulation and privacy acts (ICO, 2020).

Explainability is a foundational principle of Trustworthy AI, however, recent research has determined that introducing explanations in AI systems is found to conflict with the privacy requirements of the system. Explanation interfaces are found to give adversaries an additional attack surface (Duddu & Boutet, 2022; Liu et al., 2024) to mine the information contained in the model. Privacy attacks can target explanations to retrieve information about membership in the training set (Liu et al., 2024; Naretto et al., 2022; Shokri et al., 2021), build surrogates of the underlying model (Aïvodji et al., 2020; Wang et al., 2022; Yan et al., 2023b), infer sensitive attributes of individuals (Duddu & Boutet, 2022; Luo et al., 2022) and reconstruct the training set (Shokri et al., 2021). This leakage is demonstrated across different types of XAI methods including those that are currently used in commercial production systems. In addition to privacy attacks, the content of explanations may also inadvertently expose information that is proprietary (Milli et al., 2019) and hence valuable and confidential to organizations (Winikoff & Sardelic, 2021) or sensitive to individuals, thus causing breach of data and privacy regulations. Hence researchers have highlighted the urgent need of mitigating privacy leakage through explanation interfaces (Luo et al., 2022; Patel et al., 2022; Yan et al., 2023b). Due to these concerns of the privacy vulnerabilities of explanations, necessary privacy preservation measures are required in XAI systems (Aïvodji et al., 2020; Shokri et al., 2021; Zhao et al., 2021).

1.3 Main contributions

Previous research has identified that the privacy issues in explainability are insufficiently studied (Liu et al., 2024; Luo et al., 2022; Naretto et al., 2022) despite its criticality in achieving safety in AI transparency. To the best of our knowledge, there is currently no work that provides an in-depth understanding of the conflict between privacy and explainability in AI. Hence, we focus this article on these two fundamental desiderata of Trustworthy AI and explore the landscape of privacy risks and preservation methods proposed in literature in the context of XAI. The key questions that we have designed to define the scope of this article are:

RQ1: What are the privacy risks of releasing explanations in AI systems?

RQ2: What current methods have researchers employed to achieve privacy preservation in XAI systems?

RQ3: What constitutes a privacy preserving explanation?

We conducted a scoping review guided by RQ1 and RQ2. Based on the knowledge gathered from the extracted studies, we propose characteristics of privacy preserving XAI and outline them with the help of practical use cases to answer RQ3. Our main contributions in this article are as follows:

- *Categorization of reported privacy risks in XAI:* We review the conflict between privacy and explainability in current literature and categorize the risks.

- *Identification of applicable privacy preservation methods in XAI:* We determine the privacy preservation methods that are applicable to XAI and report the progress achieved by researchers in integrating them in XAI systems.

- *Privacy preserving XAI characteristics:* We propose the desirable characteristics of privacy preserving XAI to provide researchers and practitioners the guidelines for achieving the trade-off between privacy, utility and explainability.

The rest of this article is organized as follows. Section 2 presents a brief background on XAI including its definition, evolution, categorization of explanation approaches and related reviews. In Section 3, we present the details of the scoping review methodology for extracting studies relevant to our research questions. Sections 4 and 5 synthesize the results from the scoping review. In Section 4, we consolidate both intentional and unintentional privacy risks of explanations to answer RQ1. In Section 5, we elaborate the use of privacy preserving methods on explanations and the existing works that utilize them in response to RQ2. Section 6 proposes the characteristics of privacy preserving XAI and answers RQ3. We conclude the article by discussing the results, and highlight the open issues, challenges, and recommendations for future work in Section 7 and conclusions in Section 8.

2 Background

2.1 Definition of XAI

In 2017, DARPA kickstarted its 4-year XAI program to accelerate research in the development of explanation methods and interfaces to enhance understanding and trust of end-users (Gunning & Aha, 2019). The program defined XAI as "AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future" (Gunning & Aha, 2019). The study established users' preference for systems with explanations over systems that provided only decisions. Ribeiro et al. (2016) refer to explanations of predictions as qualitative artifacts that provide the relationship between an input instance and the output prediction.

2.2 Evolution of XAI and emergence of privacy concerns

The field of explainability can be traced to the early 1990s, driven by the lack of transparency in black-box models. Early contributions (Benitez et al., 1997; Craven & Shavlik, 1995; LiMin Fu, 1994; Milaré et al., 2002; Torres & Rocco, 2005) proposed different techniques for extracting interpretable representations from these systems. The rise of deep learning and the improvement in the predictive performance of black-box systems, propelled complex uninterpretable systems into mainstream usage. However, their use in critical domains remains problematic due to their lack of transparency. Regulatory frameworks such as the General Data Protection Regulation (GDPR, 2016), specifically the provisions on individuals' rights related to automated decision-making including profiling, intensified the demand for transparent, explainable models thus resulting in a rapid growth in the field of explainability.
However, the introduction of transparency through XAI methods has also exposed new vectors for privacy leakage through explanation interfaces. Early studies (Milli et al., 2019; Shokri et al., 2020; 2021) described privacy attacks on the training data and the underlying model. In response, researchers have begun to explore defense mechanisms and pioneering works in this field (Harder et al., 2020; Patel et al., 2022) have proposed various strategies for generating privacy preserved explanations. Despite these efforts, privacy risks in XAI remain an open research problem, with novel attacks being identified and defense strategies being actively investigated. Figure 1 outlines the key milestones in the evolution of XAI and highlights the emergence of privacy issues and proposed defenses.

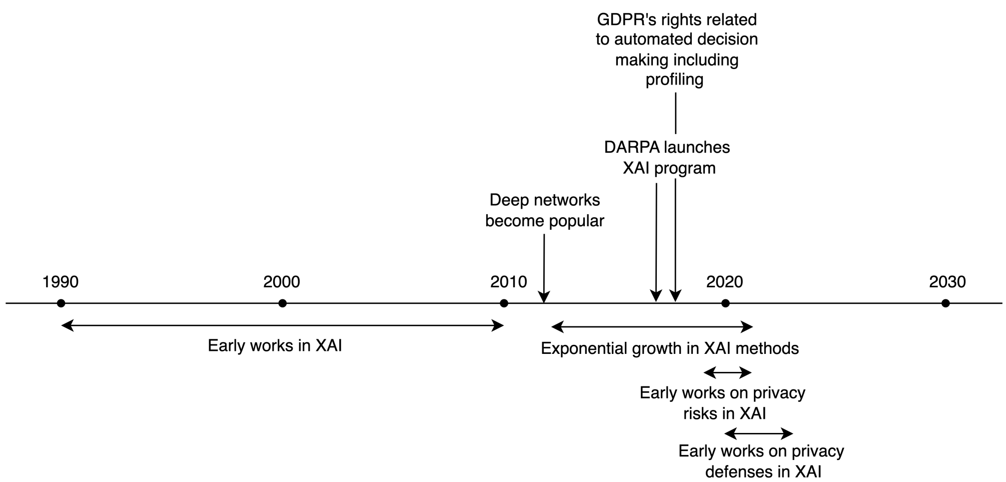

Figure 1: Key milestones and emergence of privacy attacks/defenses in XAI.

2.3 Categorization of XAI

In recent years, several different XAI methods have been proposed. Broadly, explainability can be achieved using inherently interpretable models or applying post-hoc methods on trained models (Harder et al., 2020). Methods specific to certain model types and capabilities, are referred to as model-specific while those independent of the model are referred to as model-agnostic (Dwivedi et al., 2023). In this subsection, we discuss the main categories into which XAI methods are grouped in existing literature (Table 1), based on the underlying mechanism used to derive explanations. Since there is a broad spectrum of available explainability methods, we limit ourselves to a selection of methods to give readers sufficient understanding of the terminologies used in subsequent sections. For a comprehensive review of XAI categories and methods, we refer the reader to other related reviews listed in Section 2.4.

2.3.1 Interpretable methods

These AI models are understandable by design (Arrieta et al., 2020). They have embedded rules or transparent architecture that facilitates the understanding of the input-output logic of the system. They are also referred to as white-box or transparent models. Decision trees, Bayesian networks, linear/logistic regression, k-nearest neighbours, rule based systems and general additive models (Arrieta et al., 2020; Molnar, 2023; Rawal et al., 2022) are some examples of interpretable models. According to Arrieta et al. (2020), different transparent models possess different degrees of transparency given by the properties of simulatability, decomposability and algorithmic transparency.

Though interpretable models are promising in aiding the understandability of a system, they have limitations. The primary deterrent to their successful adoption as explainable-by-design methods, is their lower accuracy (Blanco-Justicia et al., 2020; El Zein et al., 2024; Gunning & Aha, 2019) compared to better performing black-box models such as deep learning systems. When the accuracy gains between these model types is substantial, there is an unwillingness to trade performance with interpretability. Interpretable models also lack natural language explanations, making them unsuitable for use by non-technical users (Biran & McKeown, 2017). For tree-based models, the understandability deteriorates as the complexity, i.e., the depth of the tree increases (Šarčević et al., 2022). Nonetheless, due to their intrinsically transparent architecture, interpretable models are often used as surrogates for black-box models (McDermid et al., 2021). The use of multiple surrogate models is found to facilitate the availability of different types of explanations (Dwivedi et al., 2023) improving the overall interpretability of the system. For trustworthy explanations, surrogates are expected to generate accurate approximations of black-box models, failing which the usefulness of the explanations can deteriorate (Yang et al., 2022).

2.3.2 Example-based methods

These methods use examples, i.e., data instances as samples to explain the model (McDermid et al., 2021). The instances may be from the training set or generated by the method (Jiménez-Luna et al., 2020; Li et al., 2020b). These methods are also referred to as record-based (Shokri et al., 2020), instance-based (Jiménez-Luna et al., 2020) or case-based (Montenegro et al., 2021) methods in literature. They can complement feature-based methods to aid understandability of the end users (Jia et al., 2021) and also improve the interpretability of complex distributions (Kim et al., 2016). They are intuitive and natural in their ability to provide explanations to humans (Jiménez-Luna et al., 2020). Some methods in this category are anchors (Ribeiro et al., 2018), contrastive explanations (Dhurandhar et al., 2018), counterfactuals (Wachter et al., 2017), influence functions (Koh & Liang, 2017) and prototypes and criticisms (Kim et al., 2016).

Example-based explanations, though easily interpretable by end-users, can cause breach of privacy when datapoints are revealed as explanations (Shokri et al., 2021; Veugen et al., 2022). Among the different example-based methods, counterfactuals are effective for understandability, however, they can aid adversaries in determining the change in input required to alter an output to a different classification. Such manipulation of output can have undesirable consequences in critical domains (Machlev et al., 2022).

2.3.3 Knowledge-based methods

These methods utilize knowledge representation techniques in machine learning (ML) models to enhance interpretability (Tiddi & Schlobach, 2022). The integration of background knowledge (Hitzler et al., 2022) facilitates the incorporation of contextual information (Lecue, 2020; Páez, 2019), thus increasing the trustworthiness of explanations. The emerging field of neuro-symbolic (Hitzler et al., 2022) or in-between methods (Ilkou & Koutraki, 2020) explores the integration of symbolic AI approaches rooted in knowledge representation and reasoning with subsymbolic or connectionist based approaches (Hitzler et al., 2020). Neuro-symbolic hybrids aim to combine the resilience of neural approaches with the interpretability of symbolic approaches. Another knowledge-based approach is the use of semantic web technologies for semantic interpretation and automated reasoning from structured knowledge bases (Seeliger et al., 2019). Knowledge graphs and ontologies are the common tools that can be deployed to support explainability. Knowledge graphs have applicability in pre-model and post-model explainabilty contexts for feature extraction, relation identification, inferencing and reasoning (Rajabi & Etminani, 2022). The field of semantic web technologies in explainability is attractive because of its potential in creating knowledge-rich explanations without compromising the model performance (Seeliger et al., 2019).

2.3.4 Feature-based methods

These explanation methods score or measure the effect of individual input features on the output of the model (Arrieta et al., 2020; Bhatt et al., 2020; Dwivedi et al., 2023; Strobel & Shokri, 2022). They are also referred to as feature importance (McDermid et al., 2021), feature relevance (Arrieta et al., 2020) or attribution-based (Liu et al., 2024) methods. They are based on the attribution problem which is the distribution of the output of a model for a specific input to its base features (Sundararajan & Najmi, 2020). Two important categories of feature-based methods identified in literature are perturbation and backpropagation-based methods (Ancona et al., 2018; McDermid et al., 2021).

- *Perturbation-based methods* remove, alter, or mask an input feature or set of features and observe the difference with the original output (McDermid et al., 2021). Some perturbation-based methods are LIME (Ribeiro et al., 2016), permutation feature importance (Breiman, 2001), SHAP (Lundberg & Lee, 2017) and MASK (Fong & Vedaldi, 2017).
- *Backpropagation-based methods* compute input attributions in forward and backward passes of the network (Ancona et al., 2018). The use of the gradient of the output with the respective input features (McDermid et al., 2021; Strobel & Shokri, 2022) is a common approach in these methods and is referred to as gradient-based approach. Methods used on images that determine the global importance of pixels, generate saliency maps, and are referred to as pixel-level attribution methods (Kapishnikov et al., 2019; Molnar, 2023). Some examples of backpropagation-based methods are gradient (Simonyan et al., 2014), gradient x input (Shrikumar et al., 2017b), guided backpropagation (Springenberg et al., 2015) and integrated gradients (Sundararajan et al., 2017).

Compared to backpropagation, perturbation-based methods require running the model with different sets of input, hence they are slower (Ancona et al., 2018) and increasing the number of features increases the performance time (Kapishnikov et al., 2019). Moreover, when perturbation-based methods are used in neural networks, obtaining reliable results for all permutations is challenging due to non-linearity and dependence of the outcome on the exact set of features (Kapishnikov et al., 2019). Though feature-based explanations are widely used by many Machine Learning as a Service (MLaaS) platforms (Luo et al., 2022), the explanations though useful to researchers, may be difficult to understand by end-users (Veugen et al., 2022).

Table 1 Broad XAI categories and a selection of early works.

XAI Category	XAI Method		Model-specific/agnostic	Study
Interpretable	Decision trees, Bayesian networks, linear/logistic regression, k-nearest neighbours, rule-based systems, general additive models		Model-specific	-
Example-based	Anchors		Model-agnostic	Ribeiro et al. (2018)
	Contrastive explanations		Model-agnostic	Dhurandhar et al. (2018)
	Counterfactuals		Model-agnostic	Wachter et al. (2017)
	Influence functions		Model-agnostic	Koh & Liang (2017)
	Prototypes and criticisms		Model-agnostic	Kim et al. (2016)
Knowledge-based	Semantic web technologies		Model-agnostic	Seeliger et al. (2019)
	Neuro-symbolic approaches		Model-specific	Hitzler et al. (2022)
Feature-based	Perturbation-based	LIME	Model-agnostic	Ribeiro et al. (2016)

XAI Category		XAI Method	Model-specific/agnostic	Study
		Permutation Feature Importance	Model-agnostic	Breiman (2001)
		SHAP	Model-agnostic	Lundberg & Lee (2017)
		MASK	Model-agnostic	Fong & Vedaldi (2017)
	Backpropagation-based	Gradient	Model-specific	Simonyan et al. (2014)
		Gradient x Input	Model-specific	Shrikumar et al. (2017b)
		Guided Backpropagation	Model-specific	Springenberg et al. (2015)
		Integrated Gradients	Model-specific	Sundararajan et al. (2017)

2.4 Related reviews

XAI is currently an active research area and detailed reviews have captured the state of the art in the field. Though current literature has reviews covering different aspects of XAI, to the best of our knowledge there is a lack of comprehensive review that considers the tension of privacy with explainability. Our work addresses this gap and offers a unique contribution compared to other existing reviews. In this subsection, we identify related reviews on XAI and summarize their focus areas.

An in-depth overview of the core concepts and taxonomies in XAI was provided by Arrieta et al. (2020). Mohseni et al. (2021) conducted an interdisciplinary survey and proposed a comprehensive framework for design and evaluation of XAI methods. Dwivedi et al. (2023) covered a wide breadth of explanation algorithms, programming frameworks and software toolkits for XAI development. Ali et al. (2023) examined explainability through the lens of trustworthiness detailing evaluation metrics, available software packages and XAI datasets. Bodria et al. (2023) systematically categorized explanation methods and benchmarked prominent methods using quantitative metrics. Muralidhar et al. (2023) reviewed transparency elements from human computer interaction (HCI) in the context of explanations while Cambria et al. (2023) investigated presentation methods and usage of natural language with XAI.
Beyond these broad surveys, domain specific reviews have also emerged. For example, XAI in healthcare has been surveyed by Payrovnaziri et al. (2020) and Yang et al. (2022); in cybersecurity by Capuano et al. (2022) and in energy and power systems by Machlev et al. (2022). Methodology focussed reviews also exist, covering counterfactuals (Guidotti, 2022), data-driven knowledge-aware XAI systems (Li et al., 2020b), knowledge-graph based XAI (Rajabi & Etminani, 2022; Tiddi & Schlobach, 2022) and semantic web technologies for explanations (Seeliger et al., 2019). Recent advances include the intersection of explainability and federated learning (FL), termed as Federated XAI (Fed-XAI), reviewed by López-Blanco et al. (2023) and categorisation of explanation techniques for transformer-based language models based on training paradigms as surveyed by Zhao et al. (2024).

The focus of this review diverges from prior reviews by specifically examining the privacy risks arising from including explainability in AI systems. Further, we review strategies used by researchers in mitigating the privacy leakage in XAI. Our review employs an established scoping review methodology guided by clearly defined research questions. The resulting taxonomy of XAI privacy risks and corresponding mitigation methods are distilled from the understanding of existing literature across the privacy and XAI communities. This methodology enables a structured and rigorous approach to addressing the research questions through the analysis of the selected studies.

3 Method

We conducted a scoping review based on the Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) (Tricco et al., 2018). This section elaborates the process followed and the identified research trends.

3.1 Literature Selection and Extraction

A 4-step process was employed comprising of identification, screening, eligibility, and extraction, as illustrated in Figure 2. In the initial step of identification, Elsevier Engineering Village (Engineering Village) search platform was used and the search was conducted on Compendex and Inspec databases. These databases index publications from leading computer science publishers, including IEEE, ACM, Springer and Elsevier. A search string was formulated using the two main concepts of privacy and explainability and applied on the title, subject, and abstract fields. For reproducibility, the search and inclusion criteria used to retrieve the relevant studies is as follows:

- *Search string:* (privacy OR confidential* OR "membership inference" OR "model inversion" OR "model extract*" OR "model reconstruct*" OR "property inference") AND (explainab* OR explanat* OR interpretab* OR XAI OR recourse OR "transparency report").
- *Period of publication:* January 1, 2019, to December 31, 2024. The start year was chosen based on the seminal works (Milli et al., 2019; Shokri et al., 2020; 2021) published on this topic.
- *Date of most recent search:* Jan 6, 2025
- *Type of publications included:* journal articles, conference articles, book chapters, articles in press.
- *Type of publications excluded:* preprints, unpublished papers, dissertations, books, standards, report chapters, notes, report reviews, editorials, erratum, retracted documents.
- *Language:* English
- *Inclusion criteria:* Study should describe at least one privacy risk or privacy preservation method in XAI.

The search results comprising of 3,766 studies were exported from Engineering Village and imported into Covidence (Covidence) review management software. During the import process, the software merged duplicate studies from the databases, retaining only unique records. After deduplication, 1,943 studies were forwarded for screening wherein the title and abstract were examined to determine relevance to RQ1 or RQ2 while considering the inclusion criteria. Out of 1,943 studies, 69 studies were moved to the next step to determine eligibility wherein the full text of the identified articles were examined with respect to RQ1 and RQ2. During this stage, studies that addressed only security issues in explainability, privacy issues in ML, or survey papers were eliminated. This resulted in removal of 17 studies. At this stage 5 relevant studies, absent in the original search results, were identified through forward and backward searches. These were added to the pool resulting in extraction of 57 studies.

3.2 Research Trends

Each extracted study was categorized under the appropriate research question. The distribution of these studies for RQ1 (i.e., XAI privacy risks) and RQ2 (i.e., XAI privacy preservation) by year, can be seen in Figure 3(a). An upward trend in the reported privacy risks associated with XAI methods is evident over the period under review. Correspondingly, there has been a noticeable increase in the number of studies exploring the use of various privacy preservation methods in XAI as observed from Figure 3(b). Among these techniques, differential privacy and anonymization emerge as the most commonly employed approaches. With respect to the identified privacy risks, three types of attacks, namely, membership inference, model inversion

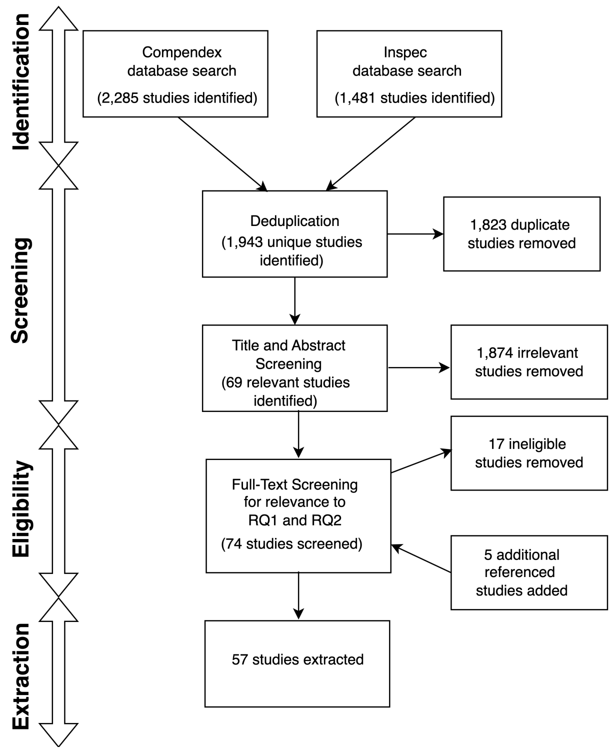

Figure 2: Scoping review process as per PRISMA-ScR.

and model extraction, appear with comparable frequency across literature (Figure 3(c)). Notably, property inference attacks have not been examined in the context of XAI systems. Figure 3(d) presents the categories of XAI targeted by different privacy attacks. Feature-based and example-based XAI are more frequently targeted to such attacks in comparison to interpretable methods. To the best of our knowledge, no privacy attacks have been reported on knowledge-based approaches.

4 Privacy Risks in XAI

Traditionally privacy is referred to as the “right to be left alone” (Warren & Brandeis, 1890) and the “claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others” (Westin, 1967). In the modern context, with availability, collection, and collation of copious information about individuals through online and offline sources, the concept of information privacy is more applicable and refers to the ability of individuals to exert control on their own data (Curzon et al., 2021). Clarke (1999) has defined information privacy as the “claims of individuals that data about themselves should generally not be available to other individuals and organizations, and that, where data is possessed by another party, the individual must be able to exercise a substantial degree of control over that data and its use”. In this article, we refer to this latter definition of privacy.

Trustworthy AI is built on the foundational principle of explainability, which supports the gaining of insights into the decision making processes of black-box AI systems (Tabassi, 2023). However, the relationship between privacy and explainability has contrasting aspects. On the one hand, explainability aids privacy in several ways such as in creating privacy awareness in users (Brunotte et al., 2021), in ascertaining that privacy of a system is achieved (Doshi-Velez & Kim, 2017; Müftüoğlu et al., 2022), and in determining correlations with identifiable data for removal (Hohman et al., 2019). On the other hand, explanations can reveal sensitive information contained in models and training data (Harder et al., 2020; Rawal et al., 2022; Zhao et al., 2021) thus leading to privacy risks (Kuppa & Le-Khac, 2021). Thus, there are conflicting outcomes (Guerra-Manzanares et al., 2023; Nguyen et al., 2023; Sanderson et al., 2023; Spartalis et al., 2024) of in-

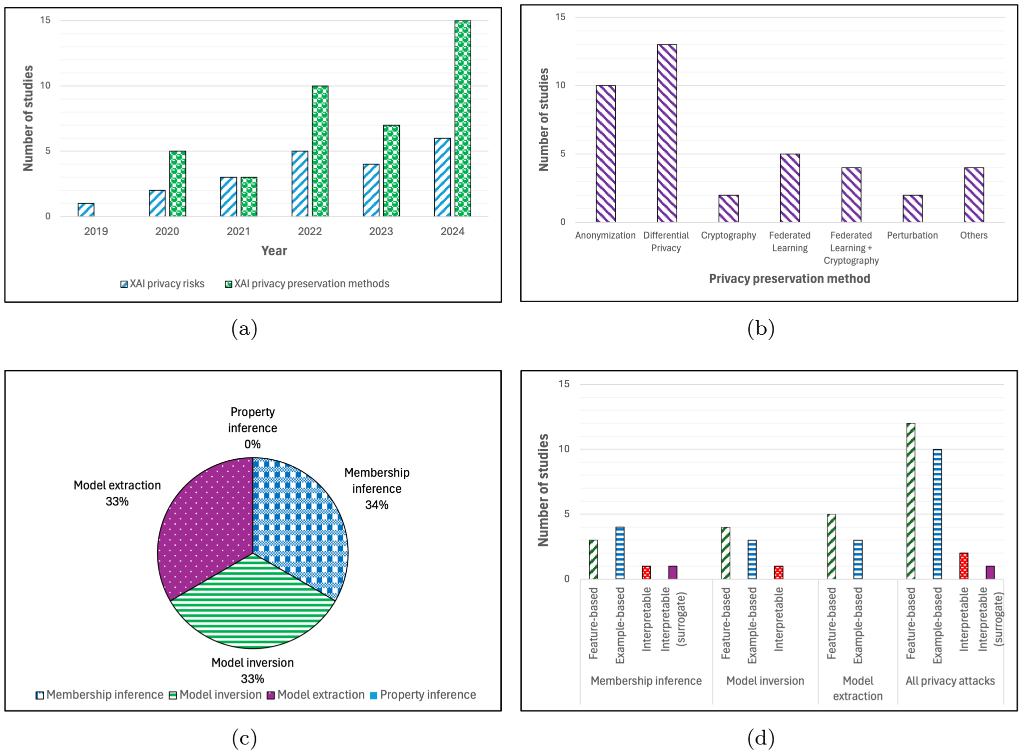

Figure 3: Research trends identified from extracted studies (a) Studies on XAI privacy risks and preservation methods (b) Privacy preservation methods in XAI (c) Privacy attacks in XAI (d) Privacy attacks by XAI categories.

cluding explainability as a non-functional requirement in AI systems. Figure 4 summarises the privacy risks that can manifest in AI systems due to the inclusion of explanation techniques.

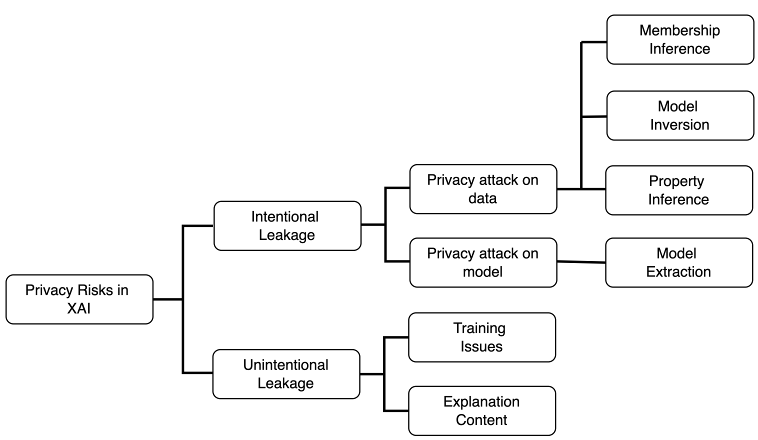

Figure 4: Proposed taxonomy of privacy risks in XAI.

4.1 Intentional Privacy Leakage

This subsection reviews intentional privacy risks in the form of privacy attacks. As XAI systems are fundamentally AI models augmented with explainability features (Figure 5), they remain susceptible to malicious threats that affect conventional AI models. Prior research has identified security attacks, such as evasion and poisoning (Pitropakis et al., 2019), that compromise the integrity of AI. However, the present review focusses on privacy attacks that aim to compromise the personal data of individuals or the confidentiality of the underlying model. In the XAI context, model explanations further aid (Zhao et al., 2021) the identification or exposure of personal information of individuals or the intellectual property of the model owner. Table 2 provides an overview of these privacy risks and the studies addressing them.

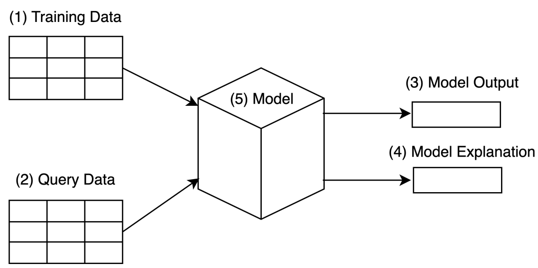

Figure 5: Assets that pose privacy risks in XAI.

4.1.1 Membership Inference

This is a privacy risk of identification of an individual in the training set of a model (Shokri et al., 2017; Zarifzadeh et al., 2024) (Figure 6). An adversary can execute this attack with black-box or white-box access to the model (Veale et al., 2018) after it has been deployed. This risk is particularly relevant in sensitive application domains leading to exposure of individuals' information used in training the model (Shokri et al., 2017). For example, with prior knowledge of an individual's personal details, such as age, gender and medical history, an adversary may determine if the individual was part of the training data of a disease detection model, suggesting a high likelihood that the individual has the disease. (Hu et al., 2022). Overfitting of the target model is identified as a main cause of membership inference (Jia et al., 2019; Yeom et al., 2018).

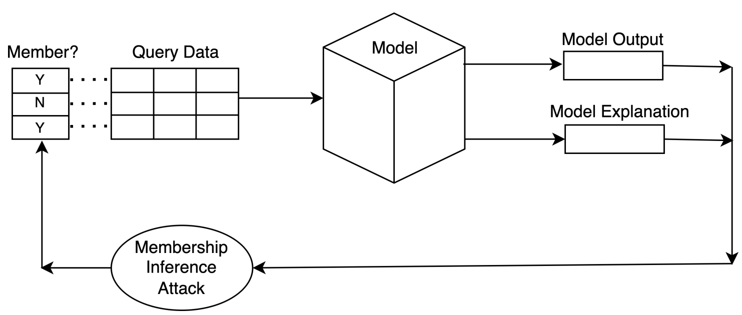

Figure 6: Membership inference exploiting explanations.

The seminal work on membership inference on feature-based and example-based XAI systems was presented by Shokri et al. (2020; 2021). The study used various backpropagation and perturbation methods to show the vulnerability of feature-based systems. The proposed attack used variances in the prediction and explanation vectors to differentiate between members and non-members. Liu et al. (2024) introduced a membership inference on feature-based XAI using model performance and robustness metrics. The study observed higher loss in confidence on perturbation of important features for members and utilized this observation in training an attack model, in addition to using the performance loss from the model. Ma et al. (2024) extended membership inference to label-only settings on Shapley value explanations. This method, which builds on earlier work on label-only attacks using hard prediction labels (Choquette-Choo et al., 2021), improved neighbourhood sampling using explanations thus reducing the number of queries.

In the example-based category, Shokri et al. (2020; 2021) investigated influence functions on logistic regression models. Since influence functions generate explanations in the form of actual datapoints, the study observed that attackers could obtain certainty about membership, thus leading to stronger attacks. More recently, Cohen & Giryes (2024) considered self-influence functions instead, that show the influence of a datapoint on its own prediction. The proposed attack required white-box access to the target model parameters, activations, and gradients. The selection of an appropriate threshold range for self-influence scores associated with members was critical for this attack and was achieved by maximizing the balanced accuracy on the training set.

Kuppa & Le-Khac (2021) used a different type of example-based explanation, namely, counterfactuals, for membership inference. The authors trained shadow models using counterfactual samples and auxiliary datasets. A threshold on the difference in predictions of the attack and target models was used to determine membership. Pawelczyk et al. (2023) also targeted counterfactuals and proposed two types of attacks. The first relied on the distances between data points and their counterfactuals to differentiate between members and non-members. The second used a loss-based approach using a likelihood-ratio test (Carlini et al., 2022) that improved the attack.

Interpretable models using decision trees, and surrogate models created using Trepan algorithm (Craven & Shavlik, 1995), were evaluated for membership inference by Naretto et al. (2022). The study also examined the effect of overfitting on the attack. The success of membership inference was determined to be higher on both interpretable and surrogate models compared to black-box models. Further, surrogates of overfitted models exhibited higher susceptibiliy to the attack than those derived from well-regularized models.

Membership inference attacks in machine learning models have been explored extensively in existing literature (Hu et al., 2022) and attack strategies have exploited confidence scores and predictions (Shokri et al., 2017). However, the recent attacks that exploit explanations suggest that XAI interfaces provide a new avenue for adversaries to launch this attack. Such attacks have targeted feature-based, example-based, and interpretable (including surrogates) XAI methods. The effectiveness of the attack is influenced by factors such as dataset type (Shokri et al., 2021), dimension (Pawelczyk et al., 2023; Shokri et al., 2021), model architecture (Shokri et al., 2021) and overfitting (Pawelczyk et al., 2023). Some attacks have proven effective in the absence of knowledge of the training dataset or target architectures (Liu et al., 2024), underscoring their practical threat potential.

While interpretable models are often recommended as surrogates for explaining black-box models (McDermid et al., 2021), as demonstrated by these attacks, the layer of interpretability can introduce a backdoor to the target system and lead to privacy leaks (Naretto et al., 2022). In the example-based category, influence functions expose data instances, particularly outliers, due to their distinct characteristics and higher influence on the training process (Shokri et al., 2021). Among feature-based methods, those using perturbations exhibit higher resilience to membership inference due to use of out-of-distribution points, however, this can also result in reduced explanation fidelity ((Shokri et al., 2021). Conversely, feature-based methods with better explanation quality are also found to be susceptible to higher leakage (Liu et al., 2024) suggesting a conflict between privacy and utility.

4.1.2 Model Inversion

This category of privacy risk is found to be of two types, namely, data reconstruction and training class inference (Zhang et al., 2023). These attacks can be conducted with black-box or white-box access to the model (Fredrikson et al., 2015; Veale et al., 2018) after it has been deployed. In data reconstruction (Figure 7), individuals' data used in training or querying (Zhao et al., 2021) the model is recovered partly or completely and constitutes a risk of exposure (Dionysiou et al., 2023). Attribute inference is a type of data reconstruction that can determine the values of certain attributes, generally those sensitive to individuals (Yeom et al., 2018) such as gender, age, race, and others. In the second type of model inversion, i.e., training class inference, it is possible to recover a representative record for a required target class (Dionysiou et al., 2023; Yang et al., 2019; Zhang et al., 2023).

Model inversion attacks have been documented in XAI on example-based, feature-based, and interpretable systems. Shokri et al. (2020; 2021) demonstrated a data reconstruction attack on influence functions in logistic regression models and found the attack dependent on data dimensionality. The authors designed different heuristics for low and high dimension data to improve coverage and efficiently recover more training

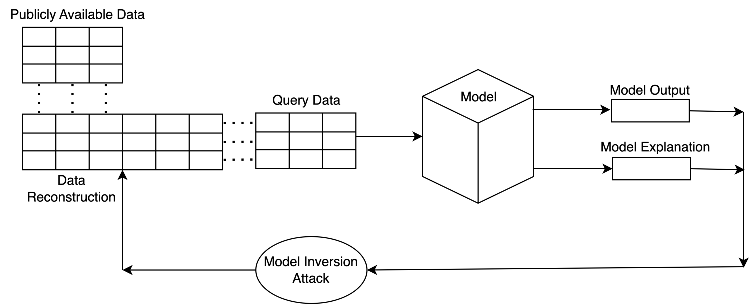

Figure 7: Model inversion exploiting explanations.

points. Goethals et al. (2023) showed an explanation linkage attack using native counterfactuals generated from actual instances of the training data. The attack demonstrated the vulnerability of counterfactuals in leaking private attributes.

Private images were found to be susceptible to recovery through saliency maps by Zhao et al. (2021) leading to inadvertent exposure. The study found XAI systems that provided class-specific multiple explanations particularly prone to leakage. The authors also used attention transfer to highlight similar risks for non-explanation models. Other studies (Duddu & Boutet, 2022; Luo et al., 2022) have focused on attribute inference of tabular data using feature-based XAI. The former trained attack models using predictions and explanations to infer sensitive features. The latter used Shapley values and effectively executed the attack with limited number of queries on cloud ML services. Toma & Kikuchi (2024) further showed that the efficacy of the proposed attack was dependent on the combination of black-box architecture and XAI method. Their findings indicate that linear models using Shapley values were particularly vulnerable to attribute inference.

Ferry et al. (2024) designed a probabilistic white-box attack applicable to transparent models, such as decision trees and rule lists, and quantified the information about the training data contained in the model. The work found that models built using greedy algorithms leak more information compared to those built using optimal strategies. The authors also observed the attack's capacity for misuse in launching other inference attacks such as membership and property inference.

The above attacks from current literature, demonstrate the leakage of privacy through XAI methods leading to exposure of personal data or sensitive attributes of individuals through explanations. Model explanations provide an effective attack surface compared to predictions (Duddu & Boutet, 2022; Zhao et al., 2021) and constitute a privacy risk indicating the contradiction between the need for explanations in Trustworthy AI and protecting privacy (Zhao et al., 2021). Data reconstruction attacks impact active users of AI systems rather than training data as in membership inference, putting end-users at risk (Zhao et al., 2021) and thus having a higher impact. In certain proposed model inversion attacks, sensitive attributes can be retrieved from models trained on non-sensitive attributes (Duddu & Boutet, 2022) while other proposed attacks demonstrate higher leakage from more important attributes (Luo et al., 2022) and recovery of entire training datasets (Shokri et al., 2021). In addition, the above works highlight the misuse of XAI techniques even for models that do not provide explanations (Zhao et al., 2021).

A tension exists between preserving privacy and maintaining utility of the XAI system. For instance, synthetic counterfactuals created by perturbing actual samples are shown to provide resilience to inversion in comparison to using native counterfactuals. However, their usage is found to affect the plausibility and runtime of explanations (Goethals et al., 2023), suggesting degrading utility. The use of multiple diverse explanations are usually recommended (Vo et al., 2023) for improving understandibility of explanations, however they are also found to contribute to leakage of privacy. Consequently restricting the access to explanation APIs has been suggested as a countermeasure (Zhao et al., 2021), however such restrictions may reduce the utility to end-users.

4.1.3 Model Extraction

This type of risk breaches the confidentiality of the target model and is a threat to the intellectual property of the model owner (Figure 8). It is also referred to as model stealing since the functionality of the model can be replicated to a significant degree of accuracy and fidelity (Jegorova et al., 2022). Since the extracted models can further leak personal data through membership inference and model inversion (Song et al., 2017), model extraction can indirectly lead to identification and exposure of individuals. This attack is usually used as a starting point for initiating other types of attacks (Aïvodji et al., 2020; Miura et al., 2024). In a typical model extraction attack, the adversary has black-box access to a deployed victim model and uses an unlabeled dataset to query it, thus generating labels to build an attack dataset (Yan et al., 2023b). This dataset is then used for training the cloned model. In contrast, data-free model extraction eliminates the need for an attack dataset. Instead, adversaries leverage generative models to synthesize the datasets, which is advantageous when the input data is difficult to obtain (Miura et al., 2024).

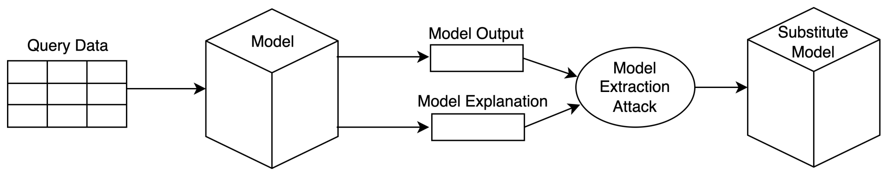

Figure 8: Model extraction exploiting explanations.

Model extraction attacks using explanations have been proposed across feature-based and example-based XAI methods. In the seminal work on the topic, gradient explanations, in the form of saliency maps, were found to be vulnerable by Milli et al. (2019). The use of explanations improved the attack by reducing the number of queries compared to attacks relying solely on model predictions. Miura et al. (2024) also leveraged gradient-based explanations but used data-free approach to train generative models for creating the attack dataset. The inclusion of explanations improved the quality of the generative model, and the accuracy of the cloned model improved with the diversity of the generated samples. Similarly, Yan et al. (2023a) employed data-free extraction wherein explanation loss was used to guide the generative model. Accuracy of the cloned model was improved by matching the victim model's predictions and explanations.

A different approach of extraction on feature-based XAI, used multitask learning to learn both classification and explanation tasks of the victim model (Yan et al., 2022). Further, a model agnostic technique on gradient and pertubation-based XAI (Yan et al., 2023b), showed that explanations provided auxiliary information that enabled more efficient attacks, reducing the query budget compared to prediction-only strategies. The attack could also be applied to non-explanation models and achieved accuracy equivalent to those of explanation models.

Besides the above extraction attacks targeting various feature-based XAI, from the example-based category, counterfactuals have been mainly targeted for this attack. In an extraction attack proposed by Aïvodji et al. (2020), they were used to approximate the decision boundary of the victim model with high accuracy and fidelity under low query budgets. Multiple and diverse counterfactuals were found to aid the extraction process by divulging additional information to adversaries. An improvisation of the attack, to reduce the number of queries further, mitigate decision boundary shift and achieve higher agreement with the victim model, was proposed by Wang et al. (2022). The method used the original counterfactual explanation with its own counterfactual as training pairs, to extract additional datapoints to train the cloned model. In another approach, Kuppa & Le-Khac (2021) proposed iterative querying of the victim model to capture the training data distribution. The method utilized distillation loss to transfer knowledge from the victim to the cloned model and was found to be successful due to the optimization of various properties such as diversity, proximity, feasibility, and sparsity.

The misuse of XAI techniques for high fidelity model cloning poses a threat to the confidentiality of model owners. As demonstrated by the aforementioned attacks, explanation-based extraction attacks offer substantial advantages over traditional prediction-only approaches by facilitating model replication with reduced number of queries (Milli et al., 2019; Miura et al., 2024). The reduction in the number of queries benefits the adversary, especially in pay-by-query models. Certain attacks are also possible with partial knowledge of the data distribution (Aïvodji et al., 2020) or in absence of overlap between the attack and training datasets (Yan et al., 2022). In addition, in scenarios where attackers do not possess the input datasets, data-free extraction attacks are possible and the use of explanations is shown to improve the attack accuracy (Miura et al., 2024). Moreover, the diversity of the generated input datasets in such attacks is found to improve the accuracy of the cloned models (Miura et al., 2024).

In addition to the direct threat to explanation models, XAI techniques can be misused for extraction of non-explanation models (Yan et al., 2023b). The use of diverse explanations, intended to build user trust in explanations, can lead to further leakage of privacy (Aïvodji et al., 2020). Similarly, the optimization of counterfactuals to satisfy various properties to improve explanation quality, can reveal information to adversaries about class-specific decision boundaries thus aiding the attack (Kuppa & Le-Khac, 2021) and leading to the conflict of explainability and privacy with utility.

4.1.4 Property Inference

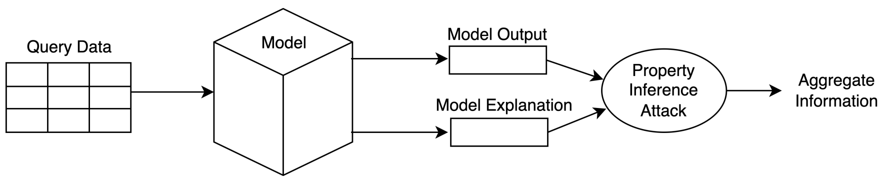

Figure 9: Property inference exploiting explanations.

This type of risk pertains to inference of properties from the training data such as global statistics or aggregates (Mahloujifar et al., 2022), which model owners did not intend on sharing (Ganju et al., 2018) (Figure 9). The inferred property does not need to correspond to features in the training data or be correlated to any feature. For instance, the adversary may deduce the gender distribution in the training set (Naretto et al., 2022) or infer the collective employee sentiments through internal company emails used to train a spam classifier (Mahloujifar et al., 2022). Such inferences can lead to exposure of sensitive information and constitutes a privacy risk. Although property inference is a known issue in AI models, to the best of our knowledge, no attacks have yet been documented that exploit explanations for this purpose.

Table 2 Studies on intentional privacy leakage in XAI systems.

Privacy Risk	**XAI Category**	**XAI Method**	**Study**
Membership inference	Interpretable	Decision tree	Naretto et al. (2022)
	Interpretable (surrogate)	Trepan	Naretto et al. (2022)
	Example-based	Influence functions	Shokri et al. (2020; 2021)
		Counterfactuals	Kuppa & Le-Khac (2021); Pawelczyk et al. (2023)
		Self-influence functions	Cohen & Giryes (2024)
	Feature-based	Gradient, integrated gradients, guided backpropagation, LRP, LIME, SmoothGrad	Shokri et al. (2020; 2021)
		Integrated gradients, SmoothGrad, VarGrad, Grad-CAM++, SHAP, LIME	Liu et al. (2024)
		Shapley values	Ma et al. (2024)

Privacy Risk	XAI Category	XAI Method	Study
Model inversion	Interpretable	Decision tree, rule list	Ferry et al. (2024)
	Example-based	Influence functions	Shokri et al. (2020; 2021)
		Counterfactuals (native)	(Goethals et al., 2023)
	Feature-based	Gradient, gradient x input, class activation maps (CAM), Grad-CAM, LRP	(Zhao et al., 2021)
		Integrated gradients, DeepLIFT, GradientSHAP, SmoothGrad	Duddu & Boutet (2022)
		Shapley values	Luo et al. (2022); Toma & Kikuchi (2024)
Model extraction	Example-based	Counterfactuals	Aïvodji et al. (2020); Kuppa & Le-Khac (2021); Wang et al. (2022)
	Feature-based	Gradient	Milli et al. (2019)
		Gradient, Grad-CAM, MASK	Yan et al. (2022)
		Gradient, Grad-CAM, MASK, LIME	Yan et al. (2023b)
		Grad-CAM, LIME	Yan et al. (2023a)
		Gradient, integrated gradients, SmoothGrad	(Miura et al., 2024)
Property inference	Not reported	Not reported	-

4.2 Unintentional Privacy Leakage

This subsection discusses unintentional privacy leakage in XAI that occur without malicious intent (Jegorova et al., 2022). Certain leakages can occur due to the natural mechanisms of the training process or through the content of explanations.

4.2.1 Training issues

Training issues such as, overfitting and memorization, identified in AI models can lead to privacy leakage. Overfitting is found to aid membership and attribute inference attacks (Yeom et al., 2018). Memorization leads to the model remembering subsets of training data (Song et al., 2017) and occurs during training before overfitting begins (Jegorova et al., 2022). It can cause leakage when data owners deploy models with codebases and training pipelines developed by third parties, such as in MLaaS, allowing sensitive information to be leaked from training data (Song et al., 2017).

4.2.2 Explanation content

The content of explanations may contain values of sensitive fields. For instance, in example-based explanations such as influence functions, training datapoints potentially containing sensitive fields, are directly revealed to end-users. Karimi et al. (2023) provide another example of unintentional leakage through example-based explanations, i.e., contrastive explanations, which can lead to inference of sensitive details of individuals whose partial attributes are known. Additionally, interpretable models used as surrogates, can reveal properties of the training data or additional information about the black-box beyond what the model owner intended to share (Blanco-Justicia et al., 2020). In addition to direct content-based leakage, risks may also arise from the inadvertent exposure of explanations to unauthorised users (Kuppa & Le-Khac, 2021). For example, during troubleshooting of error cases, developers or quality engineers may inadvertently access sensitive personal information in the explanation. Moreover, even when direct identifiers are absent, explanations that contain proxy or correlated features can still enable indirect inference.

5 Privacy Preservation Methods in XAI

To address the privacy risks outlined in Section 4, a growing body of research has emphasized the need for privacy preserving XAI techniques (Aïvodji et al., 2020; Shokri et al., 2021; Zhao et al., 2021). In response to these concerns, several studies have proposed methods to generate explanations while mitigating privacy concerns. Many of these approaches draw upon established principles and methods from the broader domain of privacy preserving ML (PPML), adapting them to specific challenges posed by explanation techniques. This section synthesizes the key contributions in literature to enhance the privacy guarantees of XAI systems in alignment with the objectives of RQ2. We categorize the proposed solutions under the main approaches in PPML. Table 3 summarizes these approaches and methods, and Figure 10 presents a consolidated mapping of privacy preserving strategies to specific types of privacy attacks discussed earlier.

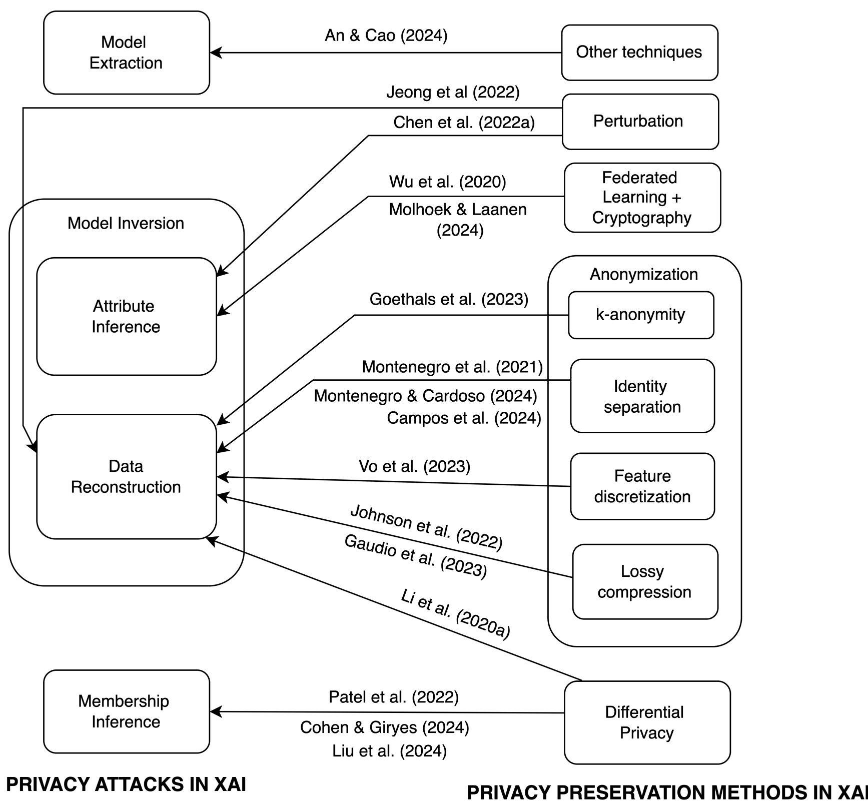

Figure 10: Proposed privacy preservation methods for specific privacy attacks in XAI.

5.1 Differential privacy

Differential privacy (DP) (Dwork et al., 2006) is a widely recognised method that provides a quantifiable definition of privacy and the incremental privacy loss from publishing confidential data (McKay Bowen & Garfinkel, 2021). A mechanism is differentially private if it can hide the participation of any single individual in a dataset (Harder et al., 2020). This can be achieved by using noise and is typically associated with an adverse effect on the accuracy of the system (Harder et al., 2020). By adjusting the privacy budget, ϵ, from 0 to ∞, practitioners can manage this trade-off between privacy and accuracy (McKay Bowen & Garfinkel, 2021). Given its robust privacy guarantees, early research in privacy preserving explanations has adopted DP using various strategies to safeguard the training data in interpretable, feature-based, and example-based XAI. In the context of XAI, an explanation is differentially private if it can obscure any single individual in the training dataset (Patel et al., 2022). The technique can be applied at various stages, including the explanation generation algorithm (Patel et al., 2022), the training process of the target model (Cohen & Giryes, 2024; Liu et al., 2024) or directly on the training data (Bozorgpanah & Torra, 2024; Ezzeddine et al., 2024).

Decision trees are popular due to their simplicity and inherent interpretable qualities, however they are prone to privacy leakage (Ferry et al., 2024; Naretto et al., 2022). A number of algorithms have been developed for building decision trees based on DP guarantees (Fletcher & Islam, 2020) offering different trade-offs in privacy and utility. The interpretability of private trees and their subsequent usefulness to XAI users, depend on factors such as the privacy budget per query, tree depth, pruning strategies and termination criteria (Fletcher & Islam, 2020). Nori et al. (2021) applied DP to another type of interpretable model, namely, Explainable Boosting Machines (EBM), to prevent privacy leakage of training data. The resulting privatized system demonstrated good accuracy at low privacy budgets while facilitating correction of errors introduced by noise, the removal of bias and the enforcement of constraints such as monotonicity. Building on this, Baek & Chung (2024) further optimized the utilization of privacy budget in these models to improve accuracy, incorporating techniques such as gradient error optimization and pruning of non-essential features.

Harder et al. (2020) proposed an interpretable model using differentially private locally linear maps with Gaussian mechanism per output class. The filters learned by the model from input images were observed to have higher interpretability compared to feature-based methods, such as integrated gradients and Smooth-Grad. However, increasing the number of such maps per class dropped accuracy due to the distribution of privacy budget over additional parameters. In a different approach, Li et al. (2020a) proposed an interpretable model in the form of feedforward-designed convolutional neural network (FF-CNN) made privacy preserving by using DP guarantees on subspace approximation with adjusted bias (Saab). Their findings indicated the the integration of DP was effective in mitigating the risk of reconstruction of input images while maintaining classification accuracy.

For feature-based explanations generated by local linear approximations around the point of interest, Patel et al. (2022) introduced a differentially private approach for loss calculation in the explanation algorithm. The study also proposed an adaptive method of reusing previous explanations for prudent usage of the privacy budget. Nguyen et al. (2023) employed local DP to restrict adversaries from learning the top influential features through aggregated scores in feature-based XAI. While originally proposed to defend against a backdoor security attack exploiting explanations (Severi et al., 2021), the random perturbation of influential features under local DP guarantees was observed to preserve the privacy of those features while maintaining explanation fidelity. Bozorgpanah & Torra (2024) also applied local DP to mask the training dataset and investigated its impact on privacy and utility of feature-based explanations. They introduced an irregularity metric to measure the feature distortion due to privatization of the original dataset and the change in explanation values. Their findings indicate that the use of additive noise on the training dataset caused irregularities, thereby reducing the utility of the explanations. In a related study, Ezzeddine et al. (2024) added calibrated noise to training datasets and evaluated the impact on SHAP explanations using various distance metrics. They observed the change in SHAP values in the privatized systems correlated with the privacy budget and data dependent. Abbasi et al. (2024) used a different approach on the data and employed DP for synthetic data generation for training of different model architectures. They used similarity scores to track the change in explanations while utility loss evaluated the drop in accuracy, thus quantifying the triad of privacy, utility and explainability.

In addition to the aforementioned studies, researchers have also examined the use of DP in the training process of the target model in feature-based (Liu et al., 2024) and example-based (Cohen & Giryes, 2024; Mochaourab et al., 2023) XAI. These investigations have determined mitigation against membership inference attacks for high privacy budgets (Cohen & Giryes, 2024; Liu et al., 2024). The introduction of DP noise serves as a regularization mechanism for target models (Nori et al., 2021) and its mathematical guarantee enables quantification of privacy, making it a compelling choice as a privacy enhancing technology. In the context of XAI, in addition to its application in mitigating membership inference from explanations (Cohen & Giryes, 2024; Liu et al., 2024; Patel et al., 2022), it is also found to mitigate reconstruction of sensitive inputs (Li et al., 2020a). However, the improvement in mitigation of attacks at high privacy budgets and hence degrading accuracy (Cohen & Giryes, 2024; Liu et al., 2024) can be a setback to the use of this technique. In addition to adversely affecting accuracy, its introduction can also deteriorate explanation quality in terms of fidelity (Liu et al., 2024; Patel et al., 2022) with pronounced effects on minority groups (Patel et al., 2022). The technique is also ineffective against attribute inference attacks when there are existing strong correlations between different attributes (Chen et al., 2022a). In such cases, the privacy enhancing benefits of DP may be insufficient to prevent adversaries from inferring sensitive attributes.

Algorithms such as DPSaab (Li et al., 2020a), have been observed to offer a more favourable trade-off between accuracy and privacy. Practitioners can employ strategies such as reusing previously generated differentially private explanations to utilize the privacy budget effectively (Patel et al., 2022). Methods that use local DP, are observed to achieve high faithfulness of explanations with privacy (Nguyen et al., 2023), thus demonstrating that it is possible to balance multiple desirable properties. Hence judicious use of this technique can ensure that privacy is achieved while maintaining reasonable utility of the model and explanations.

5.2 Cryptography

Cryptographic protocols for privacy preservation in ML use secure algorithms to protect the target model and data. Prominent methods in this category include homomorphic encryption, secret sharing, and secure multi-party computation (Yin et al., 2022). In XAI, these techniques have seen limited application in interpretable and example-based systems. They have also been used in conjunction with other privacy preserving techniques such as federated learning (El Zein et al., 2024; Molhoek & Laanen, 2024; Wu et al., 2020; 2023).

For interpretable tree-based models, Zhao et al. (2023) proposed an additive homomorphic scheme for model owners and query users, for pushing the encrypted model and query data to cloud service providers for inferencing. Adding perturbations to the inference results and query data ensured privacy protection of these assets while maintaining accuracy comparable to non-private inference. In the example-based category, Veugen et al. (2022) proposed a cryptographic method with secure multi-party computation to generate contrastive explanations, while protecting private training data and confidentiality of the model. The algorithm securely trained a binary decision tree to generate fact and foil leaves, which were used as explanations for a query data point. Additionally, a synthetic data point from the foil leaf was provided to the end-user to enhance explainability.

Cryptographic methods, such as homomorphic encryption, introduce significant computational complexity in the system (Liu et al., 2022a). The use of encryption can deteriorate model transparency, limiting the ability of data scientists to correct errors, inspect data, add features or fine tune the model (Dowlin et al., 2016). Therefore it is essential to implement cryptographic protocols in XAI system components in the right use cases to complement other privacy preservation techniques or when other techniques are infeasible or costly.

5.3 Anonymization

Anonymization refers to the process of transformation of data (Majeed & Lee, 2021) to obscure the distinctive features of individuals, thus safeguarding their privacy. The process is associated with the removal or modification of direct and quasi-identifiers (Majeed & Lee, 2021), that can uniquely identify individuals. Various methods of anonymization are used in practice, such as k-anonymity, l-diversity, and t-closeness (Yin et al., 2022). In XAI, different anonymization techniques have been demonstrated in example-based, feature-based, and interpretable methods. Techniques such as disentangled representation learning and lossy compression have been applied on sensitive visual data, such as medical images, to generate privatized explainable-by-design representations.

A dataset is considered k-anonymous if every record is indistinguishable from k-1 other records (Sweeney, 2002b), thus providing a measure of the risk of re-identification of records. K-anonymity can be achieved using methods such as suppression and generalization (Sweeney, 2002a), which obscure the data to remove identifiable features. Though traditionally this technique is applied to target datasets for protection, Goethals et al.

(2023) proposed its usage on native counterfactuals, that are actual datapoints from the training dataset, for protection against model inversion attack through explanations. This strategy of generating k-anonymous counterfactuals was shown to result in lower information loss and higher validity, outperforming counterfactuals derived from k-anonymized datasets. Further analysis by Berning et al. (2024) determined that the effectiveness of k-anonymous counterfactuals is confined to dense areas of the dataset. Its offered privacy protection was also found disproportionate to the value of k. Vo et al. (2023) highlighted another limitation, namely, the computational overhead of generating these counterfactuals requiring querying the explainer for a large number of counterfactuals. The authors proposed an alternative strategy of privatizing diverse counterfactuals by discretization of continuous features. This technique is closely related to generalization in privacy preserving data mining and is particularly effective against linkage attacks.

K-anonymity using microaggregation has been applied on the training dataset in feature-based XAI by Bozorgpanah & Torra (2024). The study found explanations from non-private and private datasets largely aligned, with minor irregularities observed. The alignment indicated that utility was effectively preserved after privatization. Similarly, Blanco-Justicia et al. (2020) applied microaggregation to build local tree-based surrogate explanations from clusters around an instance to be explained. The method enforced k-anonymity by restricting the cluster size and incorporated shallow trees to enable comprehensibility.

An emerging area of study focusses on providing explanations while protecting privacy in the medical domain, where privacy of patients' visual data is crucial. The primary aim of such techniques is transformation of private data through removal of identifying features while retaining explanatory evidence. Strategies such as the use of autoencoders for disentanglement of identifiable characteristics (Montenegro & Cardoso, 2024), Siamese network for increasing identity distance between original and privatized images (Montenegro et al., 2021) and latent diffusion models for generating synthetic images Campos et al. (2024) are proposed. The use of lossy compression by pixel sampling is also observed to remove identification information while being post-hoc explainable (Gaudio et al., 2023; Johnson et al., 2022). This approach also has an added advantage of reducing the image size significantly, thus making medical training datasets smaller (Gaudio et al., 2023).

In critical domains, such as healthcare, anonymizing training and query data can assist in protecting identifiable information. However, the applied techniques should preserve the output quality for utility to diverse end-users (Campos et al., 2024). Unlike DP, anonymization techniques lack proven guarantees, however despite DP's theoretical guarantees, it is unable to scale beyond low resolution image data (Campos et al., 2024). Lossy compression alternatively provides an effective way of privatizing image data, with the benefits of achieving both privacy and explainability while reducing training dataset sizes. It thus enables data sharing with multiple parties in non-private settings (Gaudio et al., 2023) and can serve as an explanation generation method for sensitive image data.

Anonymization techniques, such as k-anonymity, protect privacy of individuals by mitigating re-identification and linkage attacks (Vo et al., 2023). When applying k-anonymity, selecting an appropriate value of k is critical in striking the right balance between accuracy and privacy risk level (Bozorgpanah & Torra, 2024). While higher values of k enhance privacy, explainability may be adversely affected (Berning et al., 2024; Blanco-Justicia et al., 2020). Moreover, the actual level of privacy may also not scale with increasing values of k (Berning et al., 2024). Hence the selection of k that achieves the right trade-off in privacy, explainability and utility is important. Additionally, k-anonymity has limitations such as its dependence on data characteristics, susceptibility to homogeneity attack (Berning et al., 2024), and its vulnerability to privacy leakage when background knowledge is available or diversity is lacking in the private attributes (Goethals et al., 2023). Other techniques such as l-diversity and t-closeness may address some of these challenges, though their applicability to explanations remain unexplored.

The generation of synthetic data for privacy preserving data analysis is explored in previous non-XAI works (Boedihardjo et al., 2023; Liu et al., 2022b). Generating synthetic data that is private, accurate and preserves properties of the true data is a known challenge and NP-hard in the worst case (Ullman & Vadhan, 2011). When models are trained on such data, the explanations through XAI tools are expected to be inherently privacy preserving, hence this approach can be a promising direction in preserving privacy in explainable systems.

5.4 Perturbation

Perturbation of sensitive data is a widely recognized technique in the field of privacy preserving data publishing (Tran et al., 2024; Yin et al., 2022). When explanations contain sensitive information, obfuscating the contents through perturbations can prevent direct exposure. This technique can also be applied to stem indirect leakage of inferencing sensitive attributes through explanations.

Chen et al. (2022a) proposed a generic privacy preserving mechanism applicable to different XAI types such as feature-based and interpretable surrogates. The proposed method perturbed the decision mapping of an algorithm prior to public release of transparency reports. To mitigate privacy leakage while upholding utility, the authors defined a maximum confidence measure in the inference of sensitive attributes of data subjects and a utility measure in terms of faithfulness. Jeong et al. (2022) applied perturbations on saliency map explanations as a defense mechanism for model inversion in image models. The proposed framework comprised of a two-player minimax game between inversion and noise injector networks, in which the inversion network attempted to reconstruct images from saliency maps and the noise injector perturbed explanations to counter the inversion. The use of multiple evaluation metrics to differentiate between original and reconstructed images facilitated the quantification of the privacy of the defense mechanism.

For the prevention of privacy leakage in XAI, researchers have attempted perturbation of two types of model outputs, namely, predictions and explanations. Adding perturbations to model predictions, such as the strategy of adding noise to output confidence scores used by MemGuard (Jia et al., 2019), is found ineffective in mitigating membership inference through explanations (Liu et al., 2024). Perturbation of explanations is also insufficient in defending against data-free model extraction based on explanations (Yan et al., 2023a). However, the strategy has shown promising results in countering model inversion. The use of perturbation techniques at the explanation interface is also attractive due to its ease of implementation that requires no retraining of the model (Jeong et al., 2022). Nevertheless, large magnitude noise can degrade the usefulness of explanations (Jeong et al., 2022), hence perturbations should be carefully calibrated to minimize any adverse impact on explanation quality.

5.5 Federated Learning

Among the distributed privacy enhancing techniques available in PPML, Federated learning (FL) is an architectural solution (El Mestari et al., 2024) that enables training of local models on user devices and exchange of model parameters with a centralized server that co-ordinates the training of a shared global model (Konečný et al., 2016). It thus enables collaborative learning while keeping users' private data at the source (Guerra-Manzanares et al., 2023) and mitigates the privacy risk of multiple parties sharing their sensitive data with other parties (El Zein et al., 2024) or a centralized server (Zhu et al., 2022). In horizontal federated learning (HFL), local datasets have the same feature space but contain different samples while vertical federated learning (VFL), involves datasets with different feature spaces but overlaps in samples (Fiosina, 2022).

To address both privacy and explainability in Trustworthy AI, the combination of FL and XAI, i.e., Fed-XAI is suggested (Bárcena et al., 2022; Corcuera Bárcena et al., 2023; López-Blanco et al., 2023) and refers to the federated learning of XAI models. Many approaches of Fed-XAI using HFL and VFL are proposed in literature. Fiosina (2022) used a HFL approach for forecasting taxi trip duration and applied feature-based explainability methods post-hoc. Chen et al. (2022b) used an explainable VFL framework to optimize counterfactual explanations using a representative query distributed on multiple parties. Both setups demonstrate the use of post-hoc explainability tools in a distributed environment, with FL serving as a privacy preserving setup for collaborative learning of sensitive data owned by multiple parties. Fed-XAI architectures have also leveraged interpretable models locally, such as fuzzy rule-based classifiers (Daole et al., 2024), Takagi-Sugeno (Zhu et al., 2022) and Takagi-Sugeno-Kang (Corcuera Bárcena et al., 2023) fuzzy rule-based models. In these setups, interpretability is achieved using underlying explainable-by-design (Corcuera Bárcena et al., 2023) models.

Though FL aids privacy by default, it is prone to reconstruction and inferencing attacks (Mothukuri et al., 2021; Zhang et al., 2024). The sharing of gradients and model parameters, communication mechanism and aggregation process can lead to leakage of privacy of the participating clients (Zhang et al., 2024). Hence researchers have proposed integration of other privacy preserving techniques, such as cryptography, with FL methods. In one such work, Molhoek & Laanen (2024) generated synthetic data on vertically partitioned data in a FL two-party setup. Counterfactuals built from this synthetic data using secure multi-party computation, were ranked and shared with both parties and were found to be resilient to attribute inference. El Zein et al. (2024) proposed a HFL structure using decision tree models, wherein a global decision tree was collaboratively trained by participants and additive secret-sharing was used in aggregation of intermediate results. A VFL technique, Falcon (Wu et al., 2023), utilized a hybrid approach consisting of partially homomorphic encryption (PHE) and additive secret sharing for exchange of intermediate computations. Another setup, Pivot (Wu et al., 2020), proposed as part of Falcon, used threshold partially homomorphic encryption (TPHE) and additive secret sharing to protect privacy of intermediate exchanges. Though these works successfully integrate cryptographic techniques with FL, research has also determined that the use of cryptographic methods in FL reduces the centralized server's ability to differentiate true model parameters leading to backdoor attacks (Guo et al., 2022). Hence appropriate defense frameworks, such as trust evaluation schemes (Guo et al., 2023), should be incorporated for protection of the FL system from malicious users.

FL enables the training of AI models from diverse, private, and high-quality data (Zhu et al., 2022) located at client systems. It reduces the footprint of user data in the network (Mothukuri et al., 2021) by keeping data at the source and avoids transmission and storage of sensitive information in a centralized location when multiple parties are involved (Wang et al., 2019). Despite its benefits, in its current form FL faces challenges for its risk-free adoption (Mothukuri et al., 2021) including ensuring privacy constraints, merging of local XAI models and dealing with large data streaming that can lead to concept drifts (Bárcena et al., 2022). The introduction of XAI methods in the FL architecture, can also further increase the vulnerability of the system to privacy attacks through explanations. Thus Fed-XAI presently cannot guarantee privacy preservation through XAI components and further research to develop strategies to stem inadvertent privacy leakage through explanations is essential.

5.6 Other techniques

In addition to the main privacy preservation methods commonly employed in PPML, certain non-standard techniques have also been explored to mitigate privacy leakage in certain types of XAI. These approaches aim to enhance privacy preservation by adopting alternative strategies including limiting access to training data, obscuring features, or providing an abstraction of the target models. While not traditionally classified under formal privacy methods, these approaches contribute to reducing privacy leakage and complement other methods.

A client-centric, data-driven approach of generating counterfactuals was proposed by An & Cao (2024) by leveraging previous inferences retrieved by the model user. Due to the generation of counterfactuals locally at the client, the method was shown to be resilient to model extraction while achieving desirable properties such as diversity and succinctness. In another approach to create an interpretable model from a neural network, Marton et al. (2024) described a data-free strategy of distilling the function represented by the model. The method used synthetic data to train a set of neural networks and extracted the parameters to train an Interpretation-Net with an output representation in the form of surrogate decision trees.

Using a knowledge-based approach, Rožanec et al. (2022) applied semantic technologies in the form of domain specific ontology and knowledge graphs, to enhance explanations and describe features on a higher conceptual level. This enabled delinking explanations from features, thus preserving the confidentiality of the underlying model. Further, the integration of feature-based XAI such as LIME, enabled the system to determine features important for predictions. Terziyan & Vitko (2022) also applied semantic techniques to build XAI consisting of decision trees and rules generated from targeted points around the decision boundary of black-box models, without accessing the original training data. Due to the interoperability of semantic rules, the method enabled usage in decentralized setup for collaborative decision making, without individual parties sharing private local data.

These works demonstrate the application of data-free and knowledge-driven techniques in XAI to build privacy-by-design systems for protection of training data and the confidentiality of the model. By disconnecting features from the model and creating abstraction layers for generation of explanations (Rožanec et al., 2022), these strategies are helpful in protecting the underlying assets.

Table 3 Privacy preserving methods applied to XAI systems.

Privacy Preservation Category	Privacy Preserving Algorithm	Protected Asset	XAI Category (Method)	Study
Differential privacy	Various DP training algorithms	Training data	Interpretable (decision trees)	Fletcher & Islam (2020)
	DP locally linear maps	Training data	Interpretable (locally linear maps)	Harder et al. (2020)
	DPSaab	Training data	Interpretable (FF-CNN)	Li et al. (2020a)
	DP-EBM	Training data	Interpretable (EBM)	Baek & Chung (2024); Nori et al. (2021)
	DP explanation generation	Training and query data	Feature-based methods using local linear approximations (LIME, etc.)	Patel et al. (2022)
	Local DP	Training data	Feature-based methods that aggregate scores (SHAP, etc.)	Nguyen et al. (2023)
	DP trained SVM	Training data	Example-based (counterfactuals)	Mochaourab et al. (2023)
	DP-SGD	Training data	Feature-based (Grad-CAM)	Liu et al. (2024)
	DP-RMSProp	Training data	Example-based (self-influence functions)	Cohen & Giryes (2024)
	Local DP	Training data	Feature-based (TreeSHAP)	Bozorgpanah & Torra (2024)
	Local DP	Training data	Feature-based (SHAP)	Ezzeddine et al. (2024)
	DP-WGAN (Wasserstein GAN)	Training data	Various XAI methods from DALEX framework[1]	Abbasi et al. (2024)
Cryptography	Privacy preserving foil trees	Training data, model	Example-based (contrastive explanations)	Veugen et al. (2022)
	Additive homomorphic encryption	Query data, inference results, model	Interpretable (tree-based models)	Zhao et al. (2023)
Anonymization	Microaggregation (MDAV)	Training data, model	Interpretable (decision trees)	Blanco-Justicia et al. (2020)
	Privacy preserving generative model	Training data	Example-based (case-based)	Montenegro et al. (2021)

[1]DALEX framework is available on `https://github.com/modeloriented/dalex`

Privacy Preservation Category	Privacy Preserving Algorithm	Protected Asset	XAI Category (Method)	Study
	HeartSpot (lossy compression)	Training data	Feature-based (saliency maps)	Johnson et al. (2022)
	Discretization of features (generalization)	Training data	Example-based (counterfactuals)	Vo et al. (2023)
	CF-K (k-anonymity of counterfactuals)	Training data	Example-based (native counterfactuals)	Berning et al. (2024); Goethals et al. (2023)
	DeepFixCX (lossy compression)	Training data	Feature-based (saliency maps)	Gaudio et al. (2023)
	Microaggregation (MDAV)	Training data	Feature-based (TreeSHAP)	Bozorgpanah & Torra (2024)
	Disentangled representation learning	Training data	Example-based (case-based)	Montenegro & Cardoso (2024)
	Latent diffusion models	Training data	Example-based (case-based)	Campos et al. (2024)
Perturbation	GNIME	Training and query data	Feature-based (saliency maps)	Jeong et al. (2022)
	Linear-Time Optimal Privacy Scheme	Training and query data	Various XAI methods (interpretable surrogates, feature-based, etc.)	Chen et al. (2022a)
Federated Learning	Pivot (VFL, additive secret sharing, TPHE)	Training data	Tree-based models (transparent)	Wu et al. (2020)
	HFL	Training data	Feature-based methods (DeepLIFT, integrated gradients, LIME, etc.)	Fiosina (2022)
	HFL	Training data	Interpretable (Takagi-Sugeno,Takagi–Sugeno–Kang, fuzzy rule-based classifier)	Corcuera Bárcena et al. (2023); Daole et al. (2024); Zhu et al. (2022)
	VFL	Training data	Counterfactuals	Chen et al. (2022b)
	Falcon (VFL, additive secret sharing, PHE)	Explanations and training data	Feature-based (LIME)	Wu et al. (2023)
	PrivaTree (HFL, additive secret sharing)	Training data	Decision trees (transparent)	El Zein et al. (2024)
	VFL, SMC, Synthetic data	Query data	Example-based (counterfactuals)	Molhoek & Laanen (2024)
Other techniques	Semantic XAI	Training data	Interpretable (decision trees, semantic rules)	Terziyan & Vitko (2022)

Privacy Preservation Category	Privacy Preserving Algorithm	Protected Asset	XAI Category (Method)	Study
	Semantic technologies (knowledge graphs, ontologies)	Model	Feature-based (LIME, etc.)	Rožanec et al. (2022)
	Guarded counterfactuals	Training data, model	Example-based (counterfactuals)	An & Cao (2024)
	Interpretation-Nets	Training data	Interpretable (decision trees)	Marton et al. (2024)

6 Privacy Preserving XAI Characteristics

In the preceding sections, we have examined the privacy risks in XAI arising from both intentional and unintentional causes. We have also reviewed applicable privacy preserving methods to safeguard the additional attack surface exposed by explanations. In this section, drawing on the insights from investigation into RQ1 and RQ2, we aim to address RQ3 by identifying key characteristics that XAI should possess to mitigate the identified risks. These characteristics provide a framework for understanding the essential properties of privacy preserving XAI, taking into account the vulnerable assets that require protection and the various stakeholders involved during the AI lifecycle. The proposed characteristics offer guidelines to both researchers and practitioners to assess the effectiveness of existing privacy preserving XAI methods and guide the development of new approaches that prioritize privacy by design. By incorporating these qualities, XAI can strive to achieve the optimal balance between the triad of privacy, explainability and utility.

We present the characteristics (Figure 11) by considering three use cases outlined in Table 4. To facilitate understanding, a simplified example of an online loan application system that leverages an AI model with XAI capabilities is considered. The system uses seven input features where salary, net worth, and age, are protected features that require privacy preservation. The use cases describe the following scenarios:

- Use Case 1 considers intentional privacy leakage through an adversary.
- Use Case 2 involves interaction of a layman end-user, i.e., a bank's customer, with the XAI system. The end-user is provided an explanation of an automated decision directly through the system and indirectly through a human. Let's assume that in this use case, the loan was denied because the applicant salary was $<$ 40K and age was $>$ 50 years.
- Use Case 3 considers the interaction of technical support, i.e., AI developer and quality engineer, with the XAI system.

Table 4 Use cases for privacy preserving XAI in an online loan application system.

Property	Details
System	Online loan application system
Model owner	Bank
Model input features	salary, net worth, age, length of credit history, occupation, working hours per week, education
Sensitive features	salary, net worth, age

Property	Details
Use Case 1	Adversary with black-box access to the system.
Actors	Adversary
Overview	An adversary secures black-box access to the bank's model through the online application system. The adversary attempts different queries and observes the outputs generated by the system.
Query data	(i) randomly generated queries. (ii) targeted queries using prior information.
Use Case 2	Customer accessing explanation of the application outcome.
Actors	Customer, bank executive
Overview	A customer submits an online application for a loan and is given a denied result. The customer is provided with: (i) an automated explanation. (ii) a consultation with a bank executive to discuss the result.
Query data	salary = \$35K, net worth = \$75K, age = 55 years, length of credit history = 30 years, occupation = office executive, working hours per week = 25, education = diploma.
Use Case 3	AI developer accessing explanation for troubleshooting a reported error case and a quality engineer subsequently validating the system updates.
Actors	AI developer, quality engineer
Overview	An error is reported on a specific query and a developer updates the model during debugging. The developer accesses the explanation of the error case to verify the results. Finally, a quality engineer validates the system updates with another round of testing.
Query data	Synthetic query similar to the error case requiring troubleshooting.

We propose ten characteristics of privacy preserving XAI that aim to balance privacy, explainability and utility. The first six characteristics are derived from privacy attacks and unintentional leakage discussed in Section 4. The remaining four characteristics are focussed on addressing performance issues and ensuring regulatory compliance. The proposed characteristics are as follows:

6.1 Prevent training data identification

XAI tools should be designed such that they do not facilitate identification of individuals used in training the model. In Use Case 1, if the adversary has access to a specific individual's input details and retrieves the corresponding outputs including the outcome and explanations, no additional advantage should be provided through explanations in determining if the individual was used in training the bank's model. Thus, the explanations should be resilient to membership inference (Section 4.1.1).

6.2 Prevent sensitive data inference

XAI tools should be designed to prevent reconstruction or inference of sensitive attributes of individuals. In Use Case 1, if the adversary has access to the non-sensitive features of an individual and the outcome of the loan application but is unaware of any sensitive feature such as salary or age, the explanations provided should not aid in inferring these sensitive features of the individual. Thus, the explanations should be resilient to model inversion (Section 4.1.2).

6.3 Prevent reverse engineering of model

XAI tools should be designed to prevent reverse engineering of the model functionalities. In Use Case 1, the adversary, by querying the bank's model and by inspecting the explanations, should be unable to build a surrogate of the original model. Thus, the explanations should be resilient to model extraction (Section 4.1.3).

6.4 Prevent property inference

XAI tools should be designed to prevent the inferencing of aggregate properties of the training data. In Use Case 1, by using targeted queries on the bank's model, the adversary should be unable to exploit explanations in determining group properties such as the ratio of old and young customers in the training data or the distribution of wealthy and average income training participants. Thus, the explanations should be resilient to property inference (Section 4.1.4).

6.5 Prevent direct exposure

Explanations generated by XAI tools should not disclose personally identifiable and/or sensitive information to unauthorized individuals (Chen et al., 2022a). Certain explanation types, such as influence functions or native counterfactuals, reveal actual datapoints leading to unintended privacy violations (Berning et al., 2024; Shokri et al., 2020; 2021). In Use Case 2, when the customer seeks an explanation on his/her application outcome, the explanation might indicate the failure to meet respective thresholds of $40K for salary and 50 years for age. Revealing actual values of protected features would breach the customer's privacy when accessed by other actors, such as the bank executive during the customer's consultation. The customer may, however, subsequently provide consent to the executive to retrieve their personal and financial details from the bank's records for consultation.

6.6 Prevent indirect exposure

The content of the provided explanations should not indirectly expose personally identifiable and/or sensitive information through correlated or proxy features to unauthorized individuals. In Use Case 2, if the explanation discloses a non-sensitive attribute such as the length of credit history, to actors other than the customer, it could indirectly lead to exposure of a sensitive attribute, such as age, due to the strong correlation between the two attributes.

6.7 Access control of explanations

The content of explanations should be accessible only to authorized users (Blanco-Justicia et al., 2020; Kuppa & Le-Khac, 2021) with provided details based on need-to-know basis. In Use Case 2, the customer is authorized to access his/her own explanation as it pertains to their specific application outcome. The bank executive should be authorized to access the explanation only if a human intermediary is required to enhance the process of explanation for the customer. In Use Case 3, the AI developer and quality engineer should be permitted to access explanations and outcomes only for synthetic queries generated to simulate specific error cases rather than for real production data.

6.8 Upholding of explanation quality

The quality of explanations should not be compromised by the introduction of privacy preservation measures. Explanations must remain useful and meaningful (Shokri et al., 2021) to target stakeholders. In Use Cases 2 and 3, the details contained in the explanations to respective users should assist them in completing their tasks effectively and/or help them interpret the outcome of the AI system.

6.9 Acceptable run time

The run time of XAI methods, being an important evaluation metric (Bodria et al., 2023), should not deteriorate by introduction of privacy preservation measures. In Use Cases 2 and 3, the explanation recipients should see the outputs within an acceptable timeframe. A long turnaround time may lead to the explanations becoming ineffective for the task at hand.

6.10 Compliance with applicable AI/privacy regulations

XAI being an AI system, should comply with applicable AI and privacy regulations specfic to the jurisdiction in which it operates. For instance, if the XAI is deployed in Canada with Canadian residents as the target users, it must adhere to the provisions of the Artificial Intelligence and Data Act (AIDA, 2022). Users should be clearly informed of the XAI capabilities of the system including the types and content of explanations and the third parties with whom the explanations may be shared. Furthermore, appropriate consent must be obtained from users, as required by applicable laws and regulations.

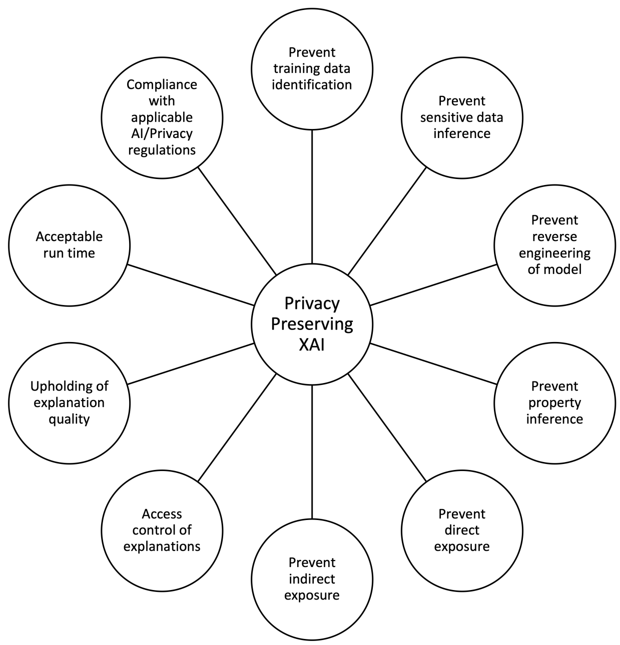

Figure 11: Proposed privacy preserving XAI characteristics.

7 Discussion

In this section, we summarize the results of our work, open issues, and challenges in the field. Additionally, we offer recommendations for future research directions to advance the development and deployment of privacy preserving XAI systems.

7.1 Summary and implications

The comprehensive review of existing literature facilitated a synthesis of current knowledge on the conflict between privacy and explainability, both being important pillars of Trustworthy AI. The analyzed studies demonstrate that the additional information provided in the form of explanations can benefit adversaries in launching privacy attacks in XAI. We identified and categorized certain types of privacy leakage due to malicious intent of adversaries as intentional causes. These include membership inference, model inversion and model extraction, all of which are have been demonstrated on explanations generated using different methods. These attacks pose a threat to the privacy of individuals whose data is contained in the training set, thus increasing the risk of identification of individuals or exposure of their sensitive information. Moreover, reconstruction of data from inference-time queries renders active XAI users vulnerable to similar privacy breaches (Zhao et al., 2023; 2021). The threat of model extraction through explanations, targets the confidentiality and intellectual property of model owners. While property inference can expose sensitive aggregates or group properties of the training data, our review found no evidence of such attacks targeting XAI systems specifically. Beyond privacy attacks, ML models exhibit inherent privacy vulnerabilities, such as memorization of training data or overfitting, which can lead to various inferencing attacks. These privacy problems are inherited in XAI systems, and we have categorized them as unintentional causes. Additionally, the explanation content can be at direct threat of privacy breaches by unauthorized users due to lack of access control, or through proxies and correlated fields. Such vectors further compound the privacy risks associated with deploying XAI technologies in sensitive domains.

Due to the growing concerns surrounding privacy risks of explainability, researchers have proposed defense mechanisms for privacy preservation with explanations. This review identifies that techniques, such as DP and anonymization, are extensively explored in this context, as evidenced by the number of studies that have employed these methods. However, there remains limited exploration of alternative approaches, including knowledge integration, cryptography, and perturbation, all of which present promising avenues for enhancing privacy preservation. Hence there is scope to utilize these underexplored techniques to achieve privacy in XAI systems. In distributed environments, Fed-XAI attempts to achieve explainability while preserving privacy of local data, yet its current implementations are insufficient to guarantee privacy in the generated explanations. Explanations produced by Fed-XAI may be vulnerable to malicious attacks, as they could inadvertently provide a backdoor to sensitive information. The integration of cryptographic techniques into FL deters the server from observing clients' true model parameters and opens avenues for injection of malicious activity, further impacting the security of the system (Guo et al., 2022).

The investigation of privacy risks and preservation methods in XAI has led to the identification of key characteristics that privacy preserving explanations must exhibit. In addition to being resilient to privacy attacks and preventing both direct and indirect exposure of sensitive information, explanations should satisfy performance and utility constraints. Given that explanations may contain potentially identifiable data and may be subject to legal and regulatory frameworks, they are required to comply with the applicable AI and privacy laws within the jurisdiction. This article identifies and highlights a gap in the research of methods within the field of XAI, i.e., explainability methods should be designed considering privacy as a system requirement. The findings of this paper can be utilized by researchers to understand state-of-the-art privacy attacks and corresponding preservation methods. Practitioners can leverage these insights to enhance their understanding of the privacy risks associated with XAI and identify potential solutions to mitigate those risks across various XAI methods.

7.2 Open issues, challenges, and recommendations for future work

Based on the privacy risks and mitigation methods surveyed, several open issues and challenges have been identified that require further attention. These challenges underscore the complexity of balancing privacy with the need for explainability in AI systems. In particular, the following issues remain critical:

7.2.1 Improving usability of XAI methods

End-users are integral and inseparable component of XAI as they directly engage and draw insights from the content generated by these systems. While current XAI methods are predominantly model-centric, focussing on model development, evaluation and audit processes (Kaplan et al., 2024), there is an increasing need for a shift towards a user-centric approach. This transition would prioritize providing need-to-know information to end-users based on their specific roles within the system. Explanations should be designed to meet the diverse informational needs of users, integrating user-centric design principles in a privacy preserving manner (Ali et al., 2023; Kaplan et al., 2024). This approach should aim to deliver explanations in a format that is accessible and meaningful, taking into consideration the varying levels of expertise and requirements of different user groups. Furthermore, to enhance the effectiveness of explanations, appropriate tools, such as interactivity and visualization, should be used to enhance the process of explanation and deepen users' understanding (Bo et al., 2024). In addition, application of the 3C-principle of context, content, and consent (Brunotte et al., 2023) can improve the usability of XAI tools by satisfying their requirements and expectations.

7.2.2 Development of privacy metrics for XAI

Though privacy is a vital requirement of XAI, currently there is a lack of suitable metrics to evaluate the privacy of explanations. Existing XAI literature provides evaluation metrics such as sufficiency (Dasgupta et al., 2022), consistency (Dasgupta et al., 2022) and sensitivity (Yeh et al., 2019) to assess aspects such as faithfulness and robustness (Hedström et al., 2023). However, these metrics do not provide a means for quantitative evaluation of privacy. Developing a privacy metric would provide developers and users an understanding of the privacy trade-offs with different methods.

7.2.3 Balancing trade-off in privacy, explainability and utility

The introduction of privacy enhancing technologies often result in adverse effects on utility (Ezzeddine et al., 2024; Harder et al., 2020), such as on the model's accuracy and the quality of explanations. For instance, the perturbation of classifier weights of support vector machines for privacy preservation is observed to deteriorate the classification accuracy and credibility of counterfactuals (Mochaourab et al., 2023). Similarly, the use of differential privacy in neural network models is found to lower its accuracy (Blanco-Justicia et al., 2023) and explanation quality (Liu et al., 2024). Generalization techniques for anonymization also introduce similar trade-off, adversely affecting the quality of explanations (Berning et al., 2024).

To address these issues, determining the appropriate trade-off between privacy, explainability and utility can help to achieve the balance between these properties. The use of compatibility matrix (Abbasi et al., 2024) or hyperparameters such as the privacy budget, ϵ, in differential privacy, is useful in tuning the desired level of these properties. Similar tuning mechanisms should be made available in other privacy preserving approaches to achieve the required trade-off. Metrics, such as trade-off score (Abbasi et al., 2024), could be useful to quantify and monitor the balance of these properties enabling researchers and practitioners to adjust the parameters based on the trade-offs involved.

7.2.4 Examining and improving trade-off in different privacy preserving methods for XAI

Different privacy preservation methods applicable to XAI are discussed in Section 5. Analysing the privacy-explainability-utility trade-off of these methods will help to identify the most effective solutions and also highlight their limitations. While techniques such as differential privacy and anonymization have been mainly explored in XAI systems, other underutilized techniques such as use of knowledge integration and cryptographic protocols, could provide alternative approaches. Distributed privacy enhancing solutions, such as Fed-XAI, warrant further investigations to determine strategies to mitigate possible privacy leakages from XAI components. By systematically examining and comparing these various privacy preserving techniques, researchers can identify best practices and design hybrid approaches to effectively balance different properties.

7.2.5 Development of XAI methods that are privacy preserving by design

As emphasized by Hoepman (2014), privacy is a core property of computer systems and requires addressing from system design phase, rather than treated as an add-on. In this context, the characteristics of privacy preserving XAI outlined in this review, can aid researchers and developers in building algorithms that are privacy-enhanced by design (Bozorgpanah & Torra, 2024). Furthermore, there is growing interest in neuro-symbolic approaches (Hitzler et al., 2022) and semantic technologies (Seeliger et al., 2019) as potential solutions as explainable-by-design strategies. Researchers and developers should continue to investigate how these techniques can be leveraged to build XAI systems that priortize privacy.

7.2.6 Privacy preserving XAI for Gen-AI and LLMs

XAI research has mainly focused on discriminative models that produce decision boundaries, and there has been limited work on developing explainability methods for Gen-AI and LLMs (Schneider, 2024; Sun et al., 2022; Weisz et al., 2023). Due to the complex structure and vast number of parameters in these models, traditional explainability methods become impractical to them (Zhao et al., 2024). These models have privacy issues, such as memorization of training data that escalates as the models become larger (Carlini et al., 2021). In addition, downstream private datasets used for in-context learning in LLMs, are found to be susceptible to membership inference (Duan et al., 2023). To mitigate some of these risks, methods such as retrieval-augmented generation (RAG) are being explored for fine tuning of outputs by augmenting external data sources (Zeng et al., 2024).

XAI plays a vital role in fostering trustworthiness (Wang & Ding, 2024) and ensuring ethical applications of these models (Luo & Specia, 2024). However, as explainability is introduced in Gen-AI and LLMs, it is necessary to ensure that it does not exacerbate the inherent privacy issues in these systems or create new vulnerabilities. A privacy analysis of explainability methods in the early stages of development and the use of privacy attacks for auditing (Carlini et al., 2021), will boost the development of privacy-enhanced systems. Thus, with the growing accessiblity and widespread use of Gen-AI and LLMs, developing appropriate user-centric privacy preserving explainability techniques is an important avenue for further research.

7.2.7 Evaluation of privacy preserving XAI characteristics

The characteristics of privacy preserving XAI that we propose, aims to highlight the desirable qualities that XAI should exhibit to protect privacy while producing useful explanations to the target users. In further work, we will evaluate current XAI methods in light of these proposed characteristics to determine gaps in the methods and inform strategies for improvement. We will also enhance current methods so that the generated explanations better align with the proposed characteristics. We aim to improve the applicability of the characteristics through the evaluation of XAI methods.

7.2.8 Comparative study of privacy risks of different XAI categories

Existing XAI methods are found to belong to different categories such as interpretable, example-based, knowledge-based, and feature-based (Section 2.3). Each category is found to have its own unique challenges, such as interpretable models are transparent but suffer low accuracy compared to complex black-box models (Blanco-Justicia et al., 2020; El Zein et al., 2024; Gunning & Aha, 2019). Example-based methods that use instances as explanations, such as influence functions and native counterfactuals, cause direct exposure of training data (Shokri et al., 2021; Veugen et al., 2022). Among feature-based methods, backpropagation are found to be more susceptible to privacy risks compared to perturbation-based (Shokri et al., 2021). A comparative study of different approaches can identify approaches that are more resilient to privacy risks compared to others. Knowledge-based methods have the capacity of segregating features from explanations which may be helpful in safeguarding the privacy of training data and model confidentiality. An analysis of the privacy risks of different approaches will help to determine the suitable approach that new methods should adopt.

8 Conclusion

XAI is an active field of research and a crucial pillar of Trustworthy AI. It aims to bring logical explanations, a fundamental property of all computer systems, to black-box AI models. Explainability of models is essential to secure user trust in automated outcomes, especially in critical domains where such outcomes have high impact on the lives of individuals. Though explainability has emerged as a gold standard for Trustworthy AI, previous works have highlighted potential privacy risks of introducing transparency to black boxes. To the best of our knowledge, there is a lack of detailed review on the tension between privacy and explainability. In this article, we have focused on this gap and conducted a scoping review to elicit details on the privacy risks posed by XAI and the corresponding solutions for privacy preservation in XAI. Our review draws attention to the intentional and unintentional misuse of explanation interfaces and the pressing need for developing XAI that is privacy preserving. In addition to reviewing the privacy risks and the progress achieved by researchers in achieving privacy improvement in XAI systems, we propose the characteristics of privacy preserving XAI, to assist AI engineers and researchers in understanding the requirements of XAI that achieves privacy with utility. We base these characteristics on the identified risks, the encountered performance issues, and the expected regulatory compliance. The characteristics can be utilized for designing new explainability methods and for evaluation of existing methods. Finally, we conclude the article by identifying the open issues and challenges in the field and provide recommendations for future work. Among the directions identified, developing privacy metrics, creating privacy preserving explanations for generative models and balancing the trade-off of privacy, utility and explainability in existing and new XAI methods, will determine its success as a foundation pillar of Trustworthy AI.

References

Wisam Abbasi, Paolo Mori, and Andrea Saracino. Further Insights: Balancing Privacy, Explainability, and Utility in Machine Learning-based Tabular Data Analysis. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, pp. 1–10, Vienna Austria, July 2024. ACM. ISBN 979-8-4007-1718-5. doi: 10.1145/3664476.3670901. URL https://dl.acm.org/doi/10.1145/3664476.3670901.

AIDA. An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts, November 2022. URL https://www.justice.gc.ca/eng/csj-sjc/pl/charter-charte/c27_1.html.

Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99:101805, November 2023. ISSN 15662535. doi: 10.1016/j.inffus.2023.101805. URL https://linkinghub.elsevier.com/retrieve/pii/S1566253523001148.

Laith Alzubaidi, Aiman Al-Sabaawi, Jinshuai Bai, Ammar Dukhan, Ahmed H. Alkenani, Ahmed Al-Asadi, Haider A. Alwzwazy, Mohamed Manoufali, Mohammed A. Fadhel, A. S. Albahri, Catarina Moreira, Chun Ouyang, Jinglan Zhang, Jose Santamaría, Asma Salhi, Freek Hollman, Ashish Gupta, Ye Duan, Timon Rabczuk, Amin Abbosh, and Yuantong Gu. Towards Risk-Free Trustworthy Artificial Intelligence: Significance and Requirements. *International Journal of Intelligent Systems*, 2023:1–41, October 2023. ISSN 1098-111X, 0884-8173. doi: 10.1155/2023/4459198. URL https://www.hindawi.com/journals/ijis/2023/4459198/.

Shuai An and Yang Cao. Counterfactual Explanation at Will, with Zero Privacy Leakage. *Proceedings of the ACM on Management of Data*, 2(3):1–29, May 2024. ISSN 2836-6573. doi: 10.1145/3654933. URL https://dl.acm.org/doi/10.1145/3654933.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *6th International Conference on Learning Representations, ICLR 2018*, Vancouver, BC, Canada, April 2018. OpenReview.net. doi: 10.3929/ethz-b-000249929. URL https://openreview.net/forum?id=Sy21R9JAW. arXiv:1711.06104 [cs, stat].

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2019.12.012.

Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. Model extraction from counterfactual explanations, September 2020. URL http://arxiv.org/abs/2009.01884. arXiv:2009.01884 [cs, stat].

Incheol Baek and Yon Dohn Chung. Differentially private and explainable boosting machine with enhanced utility. *Neurocomputing*, 607:128424, November 2024. ISSN 09252312. doi: 10.1016/j.neucom.2024.128424. URL https://linkinghub.elsevier.com/retrieve/pii/S0925231224011950.

J.M. Benitez, J.L. Castro, and I. Requena. Are artificial neural networks black boxes? *IEEE Transactions on Neural Networks*, 8(5):1156–1164, September 1997. ISSN 1045-9227, 1941-0093. doi: 10.1109/72.623216. URL https://ieeexplore.ieee.org/document/623216/.

Sjoerd Berning, Vincent Dunning, Dayana Spagnuelo, Thijs Veugen, and Jasper Van Der Waa. The Trade-off Between Privacy & Quality for Counterfactual Explanations. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, pp. 1–9, Vienna Austria, July 2024. ACM. ISBN 979-8-4007-1718-5. doi: 10.1145/3664476.3670897. URL https://dl.acm.org/doi/10.1145/3664476.3670897.

Umang Bhatt, Adrian Weller, and José M. F. Moura. Evaluating and Aggregating Feature-based Model Explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 3016–3022, Yokohama, Japan, July 2020. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-6-5. doi: 10.24963/ijcai.2020/417. URL https://www.ijcai.org/proceedings/2020/417.

Or Biran and Kathleen McKeown. Human-Centric Justification of Machine Learning Predictions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 1461–1467, Melbourne, Australia, August 2017. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-0-3. doi: 10.24963/ijcai.2017/202. URL https://www.ijcai.org/proceedings/2017/202.

Alberto Blanco-Justicia, Josep Domingo-Ferrer, Sergio Martínez, and David Sánchez. Machine learning explainability via microaggregation and shallow decision trees. *Knowledge-Based Systems*, 194:105532, April 2020. ISSN 09507051. doi: 10.1016/j.knosys.2020.105532. URL https://linkinghub.elsevier.com/retrieve/pii/S0950705120300368.

Alberto Blanco-Justicia, David Sánchez, Josep Domingo-Ferrer, and Krishnamurty Muralidhar. A Critical Review on the Use (and Misuse) of Differential Privacy in Machine Learning. *ACM Computing Surveys*, 55(8):1–16, August 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3547139. URL https://dl.acm.org/doi/10.1145/3547139.

Jessica Y Bo, Pan Hao, and Brian Y Lim. Incremental XAI: Memorable Understanding of AI with Incremental Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–17, Honolulu HI USA, May 2024. ACM. ISBN 979-8-4007-0330-0. doi: 10.1145/3613904.3642689. URL https://dl.acm.org/doi/10.1145/3613904.3642689.

Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5):1719–1778, September 2023. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-023-00933-9. URL https://link.springer.com/10.1007/s10618-023-00933-9.

March Boedihardjo, Thomas Strohmer, and Roman Vershynin. Privacy of Synthetic Data: A Statistical Framework. *IEEE Transactions on Information Theory*, 69(1):520–527, January 2023. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2022.3216793. URL https://ieeexplore.ieee.org/document/9927488/.

Aso Bozorgpanah and Vicenç Torra. Explainable machine learning models with privacy. *Progress in Artificial Intelligence*, 13(1):31–50, March 2024. ISSN 2192-6352, 2192-6360. doi: 10.1007/s13748-024-00315-2. URL https://link.springer.com/10.1007/s13748-024-00315-2.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Wasja Brunotte, Larissa Chazette, and Kai Korte. Can Explanations Support Privacy Awareness? A Research Roadmap. In *2021 IEEE 29th International Requirements Engineering Conference Workshops*

(REW), pp. 176–180, Notre Dame, IN, USA, September 2021. IEEE. ISBN 978-1-6654-1898-0. doi: 10.1109/REW53955.2021.00032. URL https://ieeexplore.ieee.org/document/9582383/.

Wasja Brunotte, Jakob Droste, and Kurt Schneider. Context, Content, Consent - How to Design User-Centered Privacy Explanations (S). pp. 86–89, July 2023. doi: 10.18293/SEKE2023-032. URL http://ksiresearchorg.ipage.com/seke/seke23paper/paper032.pdf.

José Luis Corcuera Bárcena, Mattia Daole, Pietro Ducange, Francesco Marcelloni, Alessandro Renda, Fabrizio Ruffini, and Alessio Schiavo. Fed-XAI: Federated Learning of Explainable Artificial Intelligence Models. In *CEUR Workshop Proceedings (CEUR-WS.org)*, pp. 104–117, Italy, November 2022.

Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. A survey on XAI and natural language explanations. *Information Processing & Management*, 60(1):103111, January 2023. ISSN 03064573. doi: 10.1016/j.ipm.2022.103111. URL https://linkinghub.elsevier.com/retrieve/pii/S0306457322002126.

Filipe Campos, Liliana Petrychenko, Luís F Teixeira, and Wilson Silva. Latent Diffusion Models for Privacy-preserving Medical Case-based Explanations. In *1st Workshop on Explainable Artificial Intelligence for the Medical Domain, EXPLIMED 2024*, volume 3831, Santiago de Compostela; Spain, 2024. CEUR-WS.

Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Claudio Stanzione. Explainable Artificial Intelligence in CyberSecurity: A Survey. *IEEE Access*, 10:93575–93600, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3204171. URL https://ieeexplore.ieee.org/document/9877919/.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, August 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership Inference Attacks From First Principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, San Francisco, CA, USA, May 2022. IEEE. ISBN 978-1-6654-1316-9. doi: 10.1109/SP46214.2022.9833649. URL https://ieeexplore.ieee.org/document/9833649/.

Larissa Chazette, Wasja Brunotte, and Timo Speith. Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*, pp. 197–208, Notre Dame, IN, USA, September 2021. IEEE. ISBN 978-1-6654-2856-9. doi: 10.1109/RE51729.2021.00025. URL https://ieeexplore.ieee.org/document/9604587/.

Chien-Lun Chen, Leana Golubchik, and Ranjan Pal. Achieving Transparency Report Privacy in Linear Time. *Journal of Data and Information Quality*, 14(2):1–56, June 2022a. ISSN 1936-1955, 1936-1963. doi: 10.1145/3460001. URL https://dl.acm.org/doi/10.1145/3460001.

Peng Chen, Xin Du, Zhihui Lu, Jie Wu, and Patrick C.K. Hung. EVFL: An explainable vertical federated learning for data-oriented Artificial Intelligence systems. *Journal of Systems Architecture*, 126:102474, May 2022b. ISSN 13837621. doi: 10.1016/j.sysarc.2022.102474. URL https://linkinghub.elsevier.com/retrieve/pii/S1383762122000583.

Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-Only Membership Inference Attacks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1964–1974. PMLR, July 2021. URL https://proceedings.mlr.press/v139/choquette-choo21a.html.

Roger Clarke. Internet privacy concerns confirm the case for intervention. *Communications of the ACM*, 42(2):60–67, February 1999. ISSN 0001-0782, 1557-7317. doi: 10.1145/293411.293475. URL https://dl.acm.org/doi/10.1145/293411.293475.

Gilad Cohen and Raja Giryes. Membership Inference Attack Using Self Influence Functions. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4880–4889, Waikoloa, HI, USA, January 2024. IEEE. ISBN 979-8-3503-1892-0. doi: 10.1109/WACV57701.2024.00482. URL https://ieeexplore.ieee.org/document/10484386/.

José Luis Corcuera Bárcena, Pietro Ducange, Francesco Marcelloni, Giovanni Nardini, Alessandro Noferi, Alessandro Renda, Fabrizio Ruffini, Alessio Schiavo, Giovanni Stea, and Antonio Virdis. Enabling federated learning of explainable AI models within beyond-5G/6G networks. *Computer Communications*, 210:356–375, October 2023. ISSN 01403664. doi: 10.1016/j.comcom.2023.07.039. URL https://linkinghub.elsevier.com/retrieve/pii/S0140366423002724.

Covidence. Covidence systematic review software. URL https://www.covidence.org.

Mark Craven and Jude W Shavlik. Extracting Tree-Structured Representations of Trained Networks. In *In Proceedings of the 8th International Conference on Neural Information Processing Systems (NIPS'95)*, pp. 24–30, Cambridge, MA, USA, November 1995. MIT Press.

James Curzon, Tracy Ann Kosa, Rajen Akalu, and Khalil El-Khatib. Privacy and Artificial Intelligence. *IEEE Transactions on Artificial Intelligence*, 2(2):96–108, April 2021. ISSN 2691-4581. doi: 10.1109/TAI.2021.3088084. URL https://ieeexplore.ieee.org/document/9450036/.

Mattia Daole, Pietro Ducange, Francesco Marcelloni, and Alessandro Renda. Trustworthy AI in Heterogeneous Settings: Federated Learning of Explainable Classifiers. In *2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–9, Yokohama, Japan, June 2024. IEEE. ISBN 979-8-3503-1954-5. doi: 10.1109/FUZZ-IEEE60900.2024.10612109. URL https://ieeexplore.ieee.org/document/10612109/.

Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Framework for Evaluating Faithfulness of Local Explanations. In *International Conference on Machine Learning*, pp. 4794–4815. PMLR, June 2022.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 598–617, San Jose, CA, May 2016. IEEE. ISBN 978-1-5090-0824-7. doi: 10.1109/SP.2016.42. URL http://ieeexplore.ieee.org/document/7546525/.

Ashley Deeks. The Judicial Demand For Explainable Artificial Intelligence. *Columbia Law Review*, 119(7): 1829–1850, 2019. URL https://www.jstor.org/stable/26810851.

Tribikram Dhar, Nilanjan Dey, Surekha Borra, and R. Simon Sherratt. Challenges of Deep Learning in Medical Image Analysis—Improving Explainability and Trust. *IEEE Transactions on Technology and Society*, 4(1):68–75, March 2023. ISSN 2637-6415. doi: 10.1109/TTS.2023.3234203. URL https://ieeexplore.ieee.org/document/10005626/.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. NIPS'18, pp. 590–601, Montreal, QC, Canada, 2018. Curran Associates Inc. doi: 10.5555/3326943.3326998.

Antreas Dionysiou, Vassilis Vassiliades, and Elias Athanasopoulos. Exploring Model Inversion Attacks in the Black-box Setting. *Proceedings on Privacy Enhancing Technologies*, 2023(1):190–206, January 2023. ISSN 2299-0984. doi: 10.56553/popets-2023-0012. URL https://petsymposium.org/popets/2023/popets-2023-0012.php.

Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning, March 2017. URL http://arxiv.org/abs/1702.08608. arXiv:1702.08608 [cs, stat].

Nathan Dowlin, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy. pp. 201–210, New York City, NY, USA, 2016. PMLR.

Haonan Duan, Adam Dziedzic, Mohammad Yaghini, Nicolas Papernot, and Franziska Boenisch. On the Privacy Risk of In-context Learning. In *61st Annual Meeting Of The Association For Computational Linguistics*, July 2023.

Vasisht Duddu and Antoine Boutet. Inferring Sensitive Attributes from Model Explanations. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*, pp. 416–425, New York, NY, USA, October 2022. Association for Computing Machinery. doi: 10.1145/3511808.3557362. URL https://dl.acm.org/doi/abs/10.1145/3511808.3557362.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, 55(9):1–33, September 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3561048. URL https://dl.acm.org/doi/10.1145/3561048.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In Serge Vaudenay (ed.), *Advances in Cryptology - EUROCRYPT 2006*, volume 4004, pp. 486–503. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-34546-6 978-3-540-34547-3. doi: 10.1007/11761679_29. URL http://link.springer.com/10.1007/11761679_29. Series Title: Lecture Notes in Computer Science.

Soumia Zohra El Mestari, Gabriele Lenzini, and Huseyin Demirci. Preserving data privacy in machine learning systems. *Computers & Security*, 137:103605, February 2024. ISSN 01674048. doi: 10.1016/j.cose.2023.103605. URL https://linkinghub.elsevier.com/retrieve/pii/S0167404823005151.

Yamane El Zein, Mathieu Lemay, and Kévin Huguenin. PrivaTree: Collaborative Privacy-Preserving Training of Decision Trees on Biomedical Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 21(1):1–13, January 2024. ISSN 1545-5963, 1557-9964, 2374-0043. doi: 10.1109/TCBB.2023.3286274. URL https://ieeexplore.ieee.org/document/10153643/.

Engineering Village. Engineering village: Search & discovery platform. URL https://www.engineeringvillage.com/.

Fatima Ezzeddine, Mirna Saad, Omran Ayoub, Davide Andreoletti, Martin Gjoreski, Ihab Sbeity, Marc Langheinrich, and Silvia Giordano. Differential Privacy for Anomaly Detection: Analyzing the Trade-Off Between Privacy and Explainability. In Luca Longo, Sebastian Lapuschkin, and Christin Seifert (eds.), *Explainable Artificial Intelligence*, volume 2155, pp. 294–318. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-63799-5 978-3-031-63800-8. doi: 10.1007/978-3-031-63800-8_15. URL https://link.springer.com/10.1007/978-3-031-63800-8_15. Series Title: Communications in Computer and Information Science.

Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Probabilistic Dataset Reconstruction from Interpretable Models. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 1–17, Toronto, ON, Canada, April 2024. IEEE. ISBN 979-8-3503-4950-4. doi: 10.1109/SaTML59370.2024.00009. URL https://ieeexplore.ieee.org/document/10516646/.

Jelena Fiosina. Interpretable Privacy-Preserving Collaborative Deep Learning for Taxi Trip Duration Forecasting. In Cornel Klein, Matthias Jarke, Markus Helfert, Karsten Berns, and Oleg Gusikhin (eds.), *Smart Cities, Green Technologies, and Intelligent Transport Systems*, volume 1612, pp. 392–411. Springer International Publishing, Cham, 2022. ISBN 978-3-031-17097-3 978-3-031-17098-0. doi: 10.1007/978-3-031-17098-0_20. URL https://link.springer.com/10.1007/978-3-031-17098-0_20. Series Title: Communications in Computer and Information Science.

Sam Fletcher and Md. Zahidul Islam. Decision Tree Classification with Differential Privacy: A Survey. *ACM Computing Surveys*, 52(4):1–33, July 2020. ISSN 0360-0300, 1557-7341. doi: 10.1145/3337064. URL https://dl.acm.org/doi/10.1145/3337064.

Ruth C. Fong and Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, Venice, October

2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.371. URL http://ieeexplore.ieee.org/document/8237633/.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333, Denver Colorado USA, October 2015. ACM. ISBN 978-1-4503-3832-5. doi: 10.1145/2810103.2813677. URL https://dl.acm.org/doi/10.1145/2810103.2813677.

Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 619–633, Toronto Canada, October 2018. ACM. ISBN 978-1-4503-5693-0. doi: 10.1145/3243734.3243834. URL https://dl.acm.org/doi/10.1145/3243734.3243834.

Alex Gaudio, Asim Smailagic, Christos Faloutsos, Shreshta Mohan, Elvin Johnson, Yuhao Liu, Pedro Costa, and Aurélio Campilho. DeepFixCX: Explainable privacy-preserving image compression for medical image analysis. *WIREs Data Mining and Knowledge Discovery*, March 2023. ISSN 1942-4787, 1942-4795. doi: 10.1002/widm.1495. URL https://onlinelibrary.wiley.com/doi/10.1002/widm.1495.

GDPR. Art. 22 GDPR, April 2016. URL https://gdpr-info.eu/art-22-gdpr/.

Sofie Goethals, Kenneth Sörensen, and David Martens. The Privacy Issue of Counterfactual Explanations: Explanation Linkage Attacks. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–24, October 2023. ISSN 2157-6904, 2157-6912. doi: 10.1145/3608482. URL https://dl.acm.org/doi/10.1145/3608482.

Abigail Goldsteen, Gilad Ezov, and Ariel Farkash. Reducing Risk of Model Inversion Using Privacy-Guided Training, June 2020. URL http://arxiv.org/abs/2006.15877. arXiv:2006.15877 [cs].

Alejandro Guerra-Manzanares, L. Julian Lechuga Lopez, Michail Maniatakos, and Farah E. Shamout. Privacy-Preserving Machine Learning for Healthcare: Open Challenges and Future Perspectives. In Hao Chen and Luyang Luo (eds.), *Trustworthy Machine Learning for Healthcare*, volume 13932, pp. 25–40. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-39538-3 978-3-031-39539-0. doi: 10.1007/978-3-031-39539-0_3. URL https://link.springer.com/10.1007/978-3-031-39539-0_3. Series Title: Lecture Notes in Computer Science.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, April 2022. ISSN 1384-5810, 1573-756X. doi: 10.1007/s10618-022-00831-6. URL https://link.springer.com/10.1007/s10618-022-00831-6.

David Gunning and David W Aha. DARPA's Explainable Artificial Intelligence Program. *AI Magazine*, 40 (2):44–58, June 2019. doi: 10.1609/aimag.v40i2.2850.

Jingjing Guo, Haiyang Li, Feiran Huang, Zhiquan Liu, Yanguo Peng, Xinghua Li, Jianfeng Ma, Varun G. Menon, and Konstantin Kostromitin Igorevich. ADFL: A Poisoning Attack Defense Framework for Horizontal Federated Learning. *IEEE Transactions on Industrial Informatics*, 18(10):6526–6536, October 2022. ISSN 1551-3203, 1941-0050. doi: 10.1109/TII.2022.3156645. URL https://ieeexplore.ieee.org/document/9735274/.

Jingjing Guo, Zhiquan Liu, Siyi Tian, Feiran Huang, Jiaxing Li, Xinghua Li, Kostromitin Konstantin Igorevich, and Jianfeng Ma. TFL-DT: A Trust Evaluation Scheme for Federated Learning in Digital Twin for Mobile Networks. *IEEE Journal on Selected Areas in Communications*, 41(11):3548–3560, November 2023. ISSN 0733-8716, 1558-0008. doi: 10.1109/JSAC.2023.3310094. URL https://ieeexplore.ieee.org/document/10234616/.

Frederik Harder, Matthias Bauer, and Mijung Park. Interpretable and Differentially Private Predictions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4083–4090, April 2020. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v34i04.5827. URL https://ojs.aaai.org/index.php/AAAI/article/view/5827.

Anna Hedström, Leander Weber, Dilyara Bareeva, Daniel Krakowczyk, Franz Motzkus, Wojciech Samek, and Sebastian Lapuschkin. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

Pascal Hitzler, Federico Bianchi, Monireh Ebrahimi, and Md Kamruzzaman Sarker. Neural-symbolic integration and the Semantic Web. *Semantic Web*, 11(1):3–11, January 2020. ISSN 22104968, 15700844. doi: 10.3233/SW-190368. URL https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/SW-190368.

Pascal Hitzler, Aaron Eberhart, Monireh Ebrahimi, Md Kamruzzaman Sarker, and Lu Zhou. Neuro-symbolic approaches in artificial intelligence. *National Science Review*, 9(6):nwac035, June 2022. ISSN 2095-5138, 2053-714X. doi: 10.1093/nsr/nwac035. URL https://academic.oup.com/nsr/article/doi/10.1093/nsr/nwac035/6542460.

Jaap-Henk Hoepman. Privacy Design Strategies. In Nora Cuppens-Boulahia, Frédéric Cuppens, Sushil Jajodia, Anas Abou El Kalam, and Thierry Sans (eds.), *ICT Systems Security and Privacy Protection*, volume 428, pp. 446–459. Springer Berlin Heidelberg, Berlin, Heidelberg, 2014. ISBN 978-3-642-55414-8 978-3-642-55415-5. doi: 10.1007/978-3-642-55415-5_38. URL http://link.springer.com/10.1007/978-3-642-55415-5_38. Series Title: IFIP Advances in Information and Communication Technology.

Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, Glasgow Scotland Uk, May 2019. ACM. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300809. URL https://dl.acm.org/doi/10.1145/3290605.3300809.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership Inference Attacks on Machine Learning: A Survey. *ACM Computing Surveys*, 54(11s):1–37, January 2022. ISSN 0360-0300, 1557-7341. doi: 10.1145/3523273. URL https://dl.acm.org/doi/10.1145/3523273.

ICO. Explaining decisions made with AI, 2020. URL https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-ai/.

Eleni Ilkou and Maria Koutraki. Symbolic Vs Sub-symbolic AI Methods: Friends or Enemies? In *Proceedings of the CIKM 2020 Workshops*, volume 2699, October 2020.

Marija Jegorova, Chaitanya Kaul, Charlie Mayor, Alison Q. O'Neil, Alexander Weir, Roderick Murray-Smith, and Sotirios A. Tsaftaris. Survey: Leakage and Privacy at Inference Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2022. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2022.3229593. URL https://ieeexplore.ieee.org/document/9987657/.

Hoyong Jeong, Suyoung Lee, Sung Ju Hwang, and Sooel Son. Learning to Generate Inversion-Resistant Model Explanations. New Orleans, 2022.

Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 259–274, London United Kingdom, November 2019. ACM. ISBN 978-1-4503-6747-9. doi: 10.1145/3319535.3363201. URL https://dl.acm.org/doi/10.1145/3319535.3363201.

Yan Jia, Chaitanya Kaul, Tom Lawton, Roderick Murray-Smith, and Ibrahim Habli. Prediction of weaning from mechanical ventilation using Convolutional Neural Networks. *Artificial Intelligence in Medicine*, 117:102087, July 2021. ISSN 09333657. doi: 10.1016/j.artmed.2021.102087. URL https://linkinghub.elsevier.com/retrieve/pii/S0933365721000804.

José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, October 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00236-4. URL https://www.nature.com/articles/s42256-020-00236-4.

Elvin Johnson, Shreshta Mohan, Alex Gaudio, Asim Smailagic, Christos Faloutsos, and Aurelio Campilho. HeartSpot: Privatized and Explainable Data Compression for Cardiomegaly Detection. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 01–04, Ioannina, Greece, September 2022. IEEE. ISBN 978-1-6654-8791-7. doi: 10.1109/BHI56158.2022.9926777. URL https://ieeexplore.ieee.org/document/9926777/.

Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viegas, and Michael Terry. XRAI: Better Attributions Through Regions. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4947–4956, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-7281-4803-8. doi: 10.1109/ICCV.2019.00505. URL https://ieeexplore.ieee.org/document/9008576/.

Sinan Kaplan, Hannu Uusitalo, and Lasse Lensu. A unified and practical user-centric framework for explainable artificial intelligence. *Knowledge-Based Systems*, 283:111107, January 2024. ISSN 09507051. doi: 10.1016/j.knosys.2023.111107. URL https://linkinghub.elsevier.com/retrieve/pii/S0950705123008572.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Computing Surveys*, 55(5):1–29, May 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3527848. URL https://dl.acm.org/doi/10.1145/3527848.

Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems*, volume 29 of *NIPS 2016*, 2016. ISBN 978-1-5108-3881-9.

Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. In *Proceedings of Machine Learning Research*, volume 70, pp. 1885–1894, July 2017. URL https://proceedings.mlr.press/v70/koh17a.

Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated Learning: Strategies for Improving Communication Efficiency. Barcelona, Spain, 2016. URL http://arxiv.org/abs/1610.05492. arXiv:1610.05492 [cs].

Aditya Kuppa and Nhien-An Le-Khac. Adversarial XAI Methods in Cybersecurity. *IEEE Transactions on Information Forensics and Security*, 16:4924–4938, 2021. ISSN 1556-6013, 1556-6021. doi: 10.1109/TIFS.2021.3117075. URL https://ieeexplore.ieee.org/document/9555622/.

Freddy Lecue. On the role of knowledge graphs in explainable AI. *Semantic Web*, 11(1):41–51, 2020. doi: 10.3233/SW-190374.

De Li, Jinyan Wang, Zhou Tan, Xianxian Li, and Yuhang Hu. Differential Privacy Preservation in Interpretable Feedforward-Designed Convolutional Neural Networks. In *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 631–638, Guangzhou, China, December 2020a. IEEE. ISBN 978-1-6654-0392-4. doi: 10.1109/TrustCom50675.2020.00089. URL https://ieeexplore.ieee.org/document/9343157/.

Xiao-Hui Li, Caleb Chen Cao, Yuhan Shi, Wei Bai, Han Gao, Luyu Qiu, Cong Wang, Yuanyuan Gao, Shenjia Zhang, Xun Xue, and Lei Chen. A Survey of Data-driven and Knowledge-aware eXplainable AI. *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2020b. ISSN 1041-4347, 1558-2191, 2326-3865. doi: 10.1109/TKDE.2020.2983930. URL https://ieeexplore.ieee.org/document/9050829/.

LiMin Fu. Rule generation from neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(8):1114–1124, August 1994. ISSN 00189472. doi: 10.1109/21.299696. URL http://ieeexplore.ieee.org/document/299696/.

Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Computing Surveys*, 54(2):1–36, March 2022a. ISSN 0360-0300, 1557-7341. doi: 10.1145/3436755. URL `https://dl.acm.org/doi/10.1145/3436755`.

Fan Liu, Zhiyong Cheng, Huilin Chen, Yinwei Wei, Liqiang Nie, and Mohan Kankanhalli. Privacy-Preserving Synthetic Data Generation for Recommendation Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1379–1389, Madrid Spain, July 2022b. ACM. ISBN 978-1-4503-8732-3. doi: 10.1145/3477495.3532044. URL `https://dl.acm.org/doi/10.1145/3477495.3532044`.

Han Liu, Yuhao Wu, Zhiyuan Yu, and Ning Zhang. Please Tell Me More: Privacy Impact of Explainability through the Lens of Membership Inference Attack. pp. 119–119, San Francisco, CA, USA, 2024. doi: 10.1109/SP54263.2024.00120.

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Long Beach, California, USA, December 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4. doi: 10.5555/3295222.3295230.

Haoyan Luo and Lucia Specia. From Understanding to Utilization: A Survey on Explainability for Large Language Models, February 2024. URL `http://arxiv.org/abs/2401.12874`. arXiv:2401.12874 [cs].

Xinjian Luo, Yangfan Jiang, and Xiaokui Xiao. Feature Inference Attack on Shapley Values. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2233–2247, Los Angeles CA USA, November 2022. ACM. ISBN 978-1-4503-9450-5. doi: 10.1145/3548606.3560573. URL `https://dl.acm.org/doi/10.1145/3548606.3560573`.

Raúl López-Blanco, Ricardo S. Alonso, Angélica González-Arrieta, Pablo Chamoso, and Javier Prieto. Federated Learning of Explainable Artificial Intelligence (FED-XAI): A Review. In Sascha Ossowski, Pawel Sitek, Cesar Analide, Goreti Marreiros, Pablo Chamoso, and Sara Rodríguez (eds.), *Distributed Computing and Artificial Intelligence, 20th International Conference*, volume 740, pp. 318–326. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-38332-8 978-3-031-38333-5. doi: 10.1007/978-3-031-38333-5_32. URL `https://link.springer.com/10.1007/978-3-031-38333-5_32`. Series Title: Lecture Notes in Networks and Systems.

Yao Ma, Xurong Zhai, Dan Yu, Yuli Yang, Xingyu Wei, and Yongle Chen. Label-Only Membership Inference Attack Based on Model Explanation. *Neural Processing Letters*, 56(5):236, September 2024. ISSN 1573-773X. doi: 10.1007/s11063-024-11682-1. URL `https://link.springer.com/10.1007/s11063-024-11682-1`.

R. Machlev, L. Heistrene, M. Perl, K.Y. Levy, J. Belikov, S. Mannor, and Y. Levron. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, 9:100169, August 2022. ISSN 26665468. doi: 10.1016/j.egyai.2022.100169. URL `https://linkinghub.elsevier.com/retrieve/pii/S2666546822000246`.

Saeed Mahloujifar, Esha Ghosh, and Melissa Chase. Property Inference from Poisoning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1120–1137, San Francisco, CA, USA, May 2022. IEEE. ISBN 978-1-6654-1316-9. doi: 10.1109/SP46214.2022.9833623. URL `https://ieeexplore.ieee.org/document/9833623/`.

Abdul Majeed and Sungchang Lee. Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access*, 9:8512–8545, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2020.3045700. URL `https://ieeexplore.ieee.org/document/9298747/`.

Sascha Marton, Stefan Lüdtke, Christian Bartelt, Andrej Tschalzev, and Heiner Stuckenschmidt. Explaining neural networks without access to training data. *Machine Learning*, January 2024. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-023-06428-4. URL `https://link.springer.com/10.1007/s10994-023-06428-4`.

John A. McDermid, Yan Jia, Zoe Porter, and Ibrahim Habli. Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2207):20200363, October 2021. ISSN 1364-503X, 1471-2962. doi: 10.1098/rsta.2020.0363. URL https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0363.

Claire McKay Bowen and Simson Garfinkel. The Philosophy of Differential Privacy. *Notices of the American Mathematical Society*, 68(10):1, November 2021. ISSN 0002-9920, 1088-9477. doi: 10.1090/noti2363. URL https://www.ams.org/notices/202110/rnoti-p1727.pdf.

Mohammad I. Merhi. An Assessment of the Barriers Impacting Responsible Artificial Intelligence. *Information Systems Frontiers*, April 2022. ISSN 1387-3326, 1572-9419. doi: 10.1007/s10796-022-10276-3. URL https://link.springer.com/10.1007/s10796-022-10276-3.

Bertalan Meskó and Eric J. Topol. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine*, 6(1):120, July 2023. ISSN 2398-6352. doi: 10.1038/s41746-023-00873-0. URL https://www.nature.com/articles/s41746-023-00873-0.

Claudia R. Milaré, André C. P. De L. F. De Carvalho, and Maria C. Monard. An Approach To Explain Neural Networks Using Symbolic Algorithms. *International Journal of Computational Intelligence and Applications*, 02(04):365–376, December 2002. ISSN 1469-0268, 1757-5885. doi: 10.1142/S1469026802000695. URL https://www.worldscientific.com/doi/abs/10.1142/S1469026802000695.

Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. Model Reconstruction from Model Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 1–9, Atlanta GA USA, January 2019. ACM. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287562. URL https://dl.acm.org/doi/10.1145/3287560.3287562.

Aditi Mishra, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, and Chris Bryan. PromptAid: Prompt Exploration, Perturbation, Testing and Iteration using Visual Analytics for Large Language Models, April 2023. URL http://arxiv.org/abs/2304.01964. arXiv:2304.01964 [cs].

Takayuki Miura, Toshiki Shibahara, and Naoto Yanai. MEGEX: Data-Free Model Extraction Attack Against Gradient-Based Explainable AI. In *Proceedings of the 2nd ACM Workshop on Secure and Trustworthy Deep Learning Systems*, pp. 56–66, Singapore Singapore, July 2024. ACM. ISBN 979-8-4007-0691-2. doi: 10.1145/3665451.3665533. URL https://dl.acm.org/doi/10.1145/3665451.3665533.

Rami Mochaourab, Sugandh Sinha, Stanley Greenstein, and Panagiotis Papapetrou. Demonstrator on Counterfactual Explanations for Differentially Private Support Vector Machines. In Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas (eds.), *Machine Learning and Knowledge Discovery in Databases*, volume 13718, pp. 662–666. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-26421-4 978-3-031-26422-1. doi: 10.1007/978-3-031-26422-1_52. URL https://link.springer.com/10.1007/978-3-031-26422-1_52. Series Title: Lecture Notes in Computer Science.

Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4): 1–45, December 2021. ISSN 2160-6455, 2160-6463. doi: 10.1145/3387166. URL https://dl.acm.org/doi/10.1145/3387166.

M. Molhoek and J. Van Laanen. Secure Counterfactual Explanations in a Two-party Setting. In *2024 27th International Conference on Information Fusion (FUSION)*, pp. 1–10, Venice, Italy, July 2024. IEEE. ISBN 978-1-7377497-6-9. doi: 10.23919/FUSION59988.2024.10706413. URL https://ieeexplore.ieee.org/document/10706413/.

Christoph Molnar. *Interpretable Machine Learning*. March 2023. ISBN 979-8-4114-6333-0.

Helena Montenegro and Jaime S. Cardoso. Anonymizing medical case-based explanations through disentanglement. *Medical Image Analysis*, 95:103209, July 2024. ISSN 13618415. doi: 10.1016/j.media.2024.103209. URL https://linkinghub.elsevier.com/retrieve/pii/S1361841524001348.

Helena Montenegro, Wilson Silva, and Jaime S. Cardoso. Privacy-Preserving Generative Adversarial Network for Case-Based Explainability in Medical Image Analysis. *IEEE Access*, 9:148037–148047, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3124844. URL `https://ieeexplore.ieee.org/document/9598877/`.

Viraaji Mothukuri, Reza M. Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. A survey on security and privacy of federated learning. *Future Generation Computer Systems*, 115:619–640, February 2021. ISSN 0167739X. doi: 10.1016/j.future.2020.10.007. URL `https://linkinghub.elsevier.com/retrieve/pii/S0167739X20329848`.

Deepa Muralidhar, Rafik Belloum, Kathia Marçal De Oliveira, and Ashwin Ashok. Elements that Influence Transparency in Artificial Intelligent Systems - A Survey. In José Abdelnour Nocera, Marta Kristín Lárusdóttir, Helen Petrie, Antonio Piccinno, and Marco Winckler (eds.), *Human-Computer Interaction – INTERACT 2023*, volume 14142, pp. 349–358. Springer Nature Switzerland, Cham, 2023. ISBN 978-3-031-42279-9 978-3-031-42280-5. doi: 10.1007/978-3-031-42280-5_21. URL `https://link.springer.com/10.1007/978-3-031-42280-5_21`. Series Title: Lecture Notes in Computer Science.

Z. Müftüoğlu, M. A. Kızrak, and T. Yıldırım. Privacy-Preserving Mechanisms with Explainability in Assistive AI Technologies. In George A. Tsihrintzis, Maria Virvou, Anna Esposito, and Lakhmi C. Jain (eds.), *Advances in Assistive Technologies*, volume 28, pp. 287–309. Springer International Publishing, Cham, 2022. ISBN 978-3-030-87131-4 978-3-030-87132-1. doi: 10.1007/978-3-030-87132-1_13. URL `https://link.springer.com/10.1007/978-3-030-87132-1_13`. Series Title: Learning and Analytics in Intelligent Systems.

Francesca Naretto, Anna Monreale, and Fosca Giannotti. Evaluating the Privacy Exposure of Interpretable Global Explainers. pp. 13–19, Atlanta, GA, USA, 2022. IEEE Computer Society. doi: 10.1109/CogMI56440.2022.00012.

Mohamed Nassar, Khaled Salah, Muhammad Habib ur Rehman, and Davor Svetinovic. Blockchain for explainable and trustworthy artificial intelligence. *WIREs Data Mining and Knowledge Discovery*, 10(1), January 2020. ISSN 1942-4787, 1942-4795. doi: 10.1002/widm.1340. URL `https://onlinelibrary.wiley.com/doi/10.1002/widm.1340`.

Truc Nguyen, Phung Lai, Hai Phan, and My T. Thai. XRand: Differentially Private Defense against Explanation-Guided Attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10):11873–11881, June 2023. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v37i10.26401. URL `https://ojs.aaai.org/index.php/AAAI/article/view/26401`.

Harsha Nori, Rich Caruana, Zhiqi Bu, Judy Hanwen Shen, and Janardhan Kulkarni. Accuracy, Interpretability, and Differential Privacy via Explainable Boosting. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8227–8237, 2021. URL `https://proceedings.mlr.press/v139/nori21a.html`.

Neel Patel, Reza Shokri, and Yair Zick. Model Explanations with Differential Privacy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1895–1904, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533235. URL `https://dl.acm.org/doi/10.1145/3531146.3533235`.

Martin Pawelczyk, Himabindu Lakkaraju, and Seth Neel. On the Privacy Risks of Algorithmic Recourse. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 9680–9696, Valencia, Spain, 2023. PMLR. URL `https://proceedings.mlr.press/v206/pawelczyk23a.html`.

Seyedeh Neelufar Payrovnaziri, Zhaoyi Chen, Pablo Rengifo-Moreno, Tim Miller, Jiang Bian, Jonathan H Chen, Xiuwen Liu, and Zhe He. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the American Medical Informatics Association*, 27 (7):1173–1185, July 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa053. URL `https://academic.oup.com/jamia/article/27/7/1173/5838471`.

Nikolaos Pitropakis, Emmanouil Panaousis, Thanassis Giannetsos, Eleftherios Anastasiadis, and George Loukas. A taxonomy and survey of attacks against machine learning. *Computer Science Review*, 34: 100199, November 2019. ISSN 15740137. doi: 10.1016/j.cosrev.2019.100199. URL `https://linkinghub.elsevier.com/retrieve/pii/S1574013718303289`.

Andrés Páez. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, 29(3):441–459, September 2019. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-019-09502-w. URL `http://link.springer.com/10.1007/s11023-019-09502-w`.

Enayat Rajabi and Kobra Etminani. Knowledge-graph-based explainable AI: A systematic review. *Journal of Information Science*, pp. 016555152211128, September 2022. ISSN 0165-5515, 1741-6485. doi: 10.1177/01655515221112844. URL `http://journals.sagepub.com/doi/10.1177/01655515221112844`.

Atul Rawal, James McCoy, Danda B. Rawat, Brian M. Sadler, and Robert St. Amant. Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives. *IEEE Transactions on Artificial Intelligence*, 3(6):852–866, December 2022. ISSN 2691-4581. doi: 10.1109/TAI.2021.3133846. URL `https://ieeexplore.ieee.org/document/9645355/`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, San Francisco California USA, August 2016. ACM. doi: 10.1145/2939672.2939778. URL `https://dl.acm.org/doi/10.1145/2939672.2939778`.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v32i1.11491. URL `https://ojs.aaai.org/index.php/AAAI/article/view/11491`.

Jože M. Rožanec, Blaž Fortuna, and Dunja Mladenić. Knowledge graph-based rich and confidentiality preserving Explainable Artificial Intelligence (XAI). *Information Fusion*, 81:91–102, May 2022. ISSN 15662535. doi: 10.1016/j.inffus.2021.11.015. URL `https://linkinghub.elsevier.com/retrieve/pii/S1566253521002414`.

Conrad Sanderson, David Douglas, and Qinghua Lu. Implementing Responsible AI: Tensions and Trade-Offs Between Ethics Aspects. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, Gold Coast, Australia, June 2023. IEEE. ISBN 978-1-6654-8867-9. doi: 10.1109/IJCNN54540.2023.10191274. URL `https://ieeexplore.ieee.org/document/10191274/`.

Johannes Schneider. Explainable Generative AI (GenXAI): A Survey, Conceptualization, and Research Agenda, April 2024. URL `http://arxiv.org/abs/2404.09554`. arXiv:2404.09554 [cs].

Arne Seeliger, Matthias Pfaff, and Helmut Krcmar. Semantic Web Technologies for Explainable Machine Learning Models: A Literature Review. *PROFILES/SEMEX@ ISWC*, 2465:1–16, 2019. URL `https://api.semanticscholar.org/CorpusID:204832199`.

Giorgio Severi, Jim Meyer, Alina Oprea, and Scott Coull. Explanation-Guided Backdoor Poisoning Attacks Against Malware Classifiers. In *30th USENIX Security Symposium (USENIX Security 21)*, 2021.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, San Jose, CA, USA, May 2017. IEEE. ISBN 978-1-5090-5533-3. doi: 10.1109/SP.2017.41. URL `http://ieeexplore.ieee.org/document/7958568/`.

Reza Shokri, Martin Strobel, and Yair Zick. Exploiting Transparency Measures for Membership Inference: a Cautionary Tale. In *The AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI)*, volume 13, New York, USA, 2020. AAAI.

Reza Shokri, Martin Strobel, and Yair Zick. On the Privacy Risks of Model Explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 231–241, Virtual Event USA, July 2021. ACM. ISBN 978-1-4503-8473-5. doi: 10.1145/3461702.3462533. URL https://dl.acm.org/doi/10.1145/3461702.3462533.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3145–3153, Sydney, NSW, Australia, 2017a. JMLR.org. doi: 10.5555/3305890.3306006.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences, April 2017b. URL http://arxiv.org/abs/1605.01713. arXiv:1605.01713 [cs].

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Workshop at International Conference on Learning Representations*, April 2014. arXiv:1312.6034 [cs].

Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine Learning Models that Remember Too Much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 587–601, Dallas Texas USA, October 2017. ACM. ISBN 978-1-4503-4946-8. doi: 10.1145/3133956.3134077. URL https://dl.acm.org/doi/10.1145/3133956.3134077.

Christoforos N. Spartalis, Theodoros Semertzidis, and Petros Daras. Balancing XAI with Privacy and Security Considerations. In Sokratis Katsikas, Habtamu Abie, Silvio Ranise, Luca Verderame, Enrico Cambiaso, Rita Ugarelli, Isabel Praça, Wenjuan Li, Weizhi Meng, Steven Furnell, Basel Katt, Sandeep Pirbhulal, Ankur Shukla, Michele Ianni, Mila Dalla Preda, Kim-Kwang Raymond Choo, Miguel Pupo Correia, Abhishta Abhishta, Giovanni Sileno, Mina Alishahi, Harsha Kalutarage, and Naoto Yanai (eds.), *Computer Security. ESORICS 2023 International Workshops*, volume 14399, pp. 111–124. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-54128-5 978-3-031-54129-2. doi: 10.1007/978-3-031-54129-2_7. URL https://link.springer.com/10.1007/978-3-031-54129-2_7. Series Title: Lecture Notes in Computer Science.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net. San Diego, CA, USA, 2015. ICLR. arXiv:1412.6806 [cs].

Martin Strobel and Reza Shokri. Data Privacy and Trustworthy Machine Learning. *IEEE Security & Privacy*, 20(5):44–49, September 2022. ISSN 1540-7993, 1558-4046. doi: 10.1109/MSEC.2022.3178187. URL https://ieeexplore.ieee.org/document/9802763/.

Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D. Weisz. Investigating Explainability of Generative AI for Code through Scenario-based Design. In *27th International Conference on Intelligent User Interfaces*, pp. 212–228, Helsinki Finland, March 2022. ACM. ISBN 978-1-4503-9144-3. doi: 10.1145/3490099.3511119. URL https://dl.acm.org/doi/10.1145/3490099.3511119.

Mukund Sundararajan and Amir Najmi. The Many Shapley Values for Model Explanation. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9269–9278, Virtual, September 2020. PMLR.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pp. 3319–3328, Sydney, NSW, Australia, 2017. JMLR.org. doi: 10.5555/3305890.3306024.

Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):571–588, 2002a. ISSN 0218-4885. Publisher: World Scientific.

Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002b. ISSN 0218-4885. Publisher: World Scientific.

Elham Tabassi. AI Risk Management Framework: AI RMF (1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, 2023. URL https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

Vagan Terziyan and Oleksandra Vitko. Explainable AI for Industry 4.0: Semantic Representation of Deep Learning Models. *Procedia Computer Science*, 200:216–226, 2022. ISSN 18770509. doi: 10.1016/j.procs.2022.01.220. URL https://linkinghub.elsevier.com/retrieve/pii/S1877050922002290.

Ilaria Tiddi and Stefan Schlobach. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, 302:103627, January 2022. ISSN 00043702. doi: 10.1016/j.artint.2021.103627. URL https://linkinghub.elsevier.com/retrieve/pii/S0004370221001788.

Ryotaro Toma and Hiroaki Kikuchi. Combinations of AI Models and XAI Metrics Vulnerable to Record Reconstruction Risk. In *Lecture Notes in Computer Science*, pp. 329–343. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-69650-3 978-3-031-69651-0. doi: 10.1007/978-3-031-69651-0_22. URL https://link.springer.com/10.1007/978-3-031-69651-0_22. ISSN: 0302-9743, 1611-3349.

D.E.D. Torres and C.M.S. Rocco. Extracting trees from trained SVM models using a TREPAN based approach. In *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, pp. 6 pp., Rio de Janeiro, Brazil, 2005. IEEE. ISBN 978-0-7695-2457-3. doi: 10.1109/ICHIS.2005.41. URL http://ieeexplore.ieee.org/document/1587773/.

Anh-Tu Tran, The-Dung Luong, and Van-Nam Huynh. A comprehensive survey and taxonomy on privacy-preserving deep learning. *Neurocomputing*, 576:127345, April 2024. ISSN 09252312. doi: 10.1016/j.neucom.2024.127345. URL https://linkinghub.elsevier.com/retrieve/pii/S0925231224001164.

Andrea C. Tricco, Erin Lillie, Wasifa Zarin, Kelly K. O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah D.J. Peters, Tanya Horsley, Laura Weeks, Susanne Hempel, Elie A. Akl, Christine Chang, Jessie McGowan, Lesley Stewart, Lisa Hartling, Adrian Aldcroft, Michael G. Wilson, Chantelle Garritty, Simon Lewin, Christina M. Godfrey, Marilyn T. Macdonald, Etienne V. Langlois, Karla Soares-Weiser, Jo Moriarty, Tammy Clifford, Özge Tunçalp, and Sharon E. Straus. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, 169(7):467–473, October 2018. ISSN 0003-4819, 1539-3704. doi: 10.7326/M18-0850. URL https://www.acpjournals.org/doi/10.7326/M18-0850.

Cristina Trocin, Patrick Mikalef, Zacharoula Papamitsiou, and Kieran Conboy. Responsible AI for Digital Health: a Synthesis and a Research Agenda. *Information Systems Frontiers*, June 2021. ISSN 1387-3326, 1572-9419. doi: 10.1007/s10796-021-10146-4. URL https://link.springer.com/10.1007/s10796-021-10146-4.

Jonathan Ullman and Salil Vadhan. PCPs and the Hardness of Generating Private Synthetic Data. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, and Yuval Ishai (eds.), *Theory of Cryptography*, volume 6597, pp. 400–416. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-19570-9 978-3-642-19571-6. doi: 10.1007/978-3-642-19571-6_24. URL http://link.springer.com/10.1007/978-3-642-19571-6_24. Series Title: Lecture Notes in Computer Science.

Michael Veale, Reuben Binns, and Lilian Edwards. Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180083, November 2018. ISSN 1364-503X, 1471-2962. doi: http://dx.doi.org/10.1098/rsta.2018.0083. URL https://royalsocietypublishing.org/doi/10.1098/rsta.2018.0080.

Thijs Veugen, Bart Kamphorst, and Michiel Marcus. Privacy-Preserving Contrastive Explanations with Local Foil Trees. In Shlomi Dolev, Jonathan Katz, and Amnon Meisels (eds.), *Cyber Security, Cryptology, and Machine Learning*, volume 13301, pp. 88–98. Springer International Publishing, Cham, 2022. ISBN 978-3-031-07688-6 978-3-031-07689-3. doi: 10.1007/978-3-031-07689-3_7. URL https://link.springer.com/10.1007/978-3-031-07689-3_7. Series Title: Lecture Notes in Computer Science.

Vy Vo, Trung Le, Van Nguyen, He Zhao, Edwin V. Bonilla, Gholamreza Haffari, and Dinh Phung. Feature-based Learning for Diverse and Privacy-Preserving Counterfactual Explanations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2211–2222, Long Beach CA USA, August 2023. ACM. ISBN 979-8-4007-0103-0. doi: 10.1145/3580305.3599343. URL https://dl.acm.org/doi/10.1145/3580305.3599343.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal*, 2017. ISSN 1556-5068. doi: 10.2139/ssrn.3063289. URL https://www.ssrn.com/abstract=3063289.

Guan Wang, Charlie Xiaoqian Dang, and Ziye Zhou. Measure Contribution of Participants in Federated Learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 2597–2604, Los Angeles, CA, USA, December 2019. IEEE. ISBN 978-1-7281-0858-2. doi: 10.1109/BigData47090.2019.9006179. URL https://ieeexplore.ieee.org/document/9006179/.

Ping Wang and Heng Ding. The rationality of explanation or human capacity? Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance. *Information Processing & Management*, 61(4):103732, July 2024. ISSN 03064573. doi: 10.1016/j.ipm.2024.103732. URL https://linkinghub.elsevier.com/retrieve/pii/S030645732400092X.

Yongjie Wang, Hangwei Qian, and Chunyan Miao. DualCF: Efficient Model Extraction Attack from Counterfactual Explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1318–1329, Seoul Republic of Korea, June 2022. ACM. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533188. URL https://dl.acm.org/doi/10.1145/3531146.3533188.

Samuel D. Warren and Louis D. Brandeis. The Right to Privacy. *Harvard Law Review*, IV(5), December 1890.

Justin D Weisz, Michael Muller, Jessica He, and Stephanie Houde. Toward General Design Principles for Generative AI Applications. 2023.

Alan F. Westin. *Privacy and Freedom*. Atheneum, New York, 1967.

Michael Winikoff and Julija Sardelic. Artificial Intelligence and the Right to Explanation as a Human Right. *IEEE Internet Computing*, 25(2):116–120, March 2021. ISSN 1089-7801, 1941-0131. doi: 10.1109/MIC.2020.3045821. URL https://ieeexplore.ieee.org/document/9420081/.

Yuncheng Wu, Shaofeng Cai, Xiaokui Xiao, Gang Chen, and Beng Chin Ooi. Privacy preserving vertical federated learning for tree-based models. *Proceedings of the VLDB Endowment*, 13(12):2090–2103, August 2020. ISSN 2150-8097. doi: 10.14778/3407790.3407811. URL https://dl.acm.org/doi/10.14778/3407790.3407811.

Yuncheng Wu, Naili Xing, Gang Chen, Tien Tuan Anh Dinh, Zhaojing Luo, Beng Chin Ooi, Xiaokui Xiao, and Meihui Zhang. Falcon: A Privacy-Preserving and Interpretable Vertical Federated Learning System. *Proceedings of the VLDB Endowment*, 16(10):2471–2484, June 2023. ISSN 2150-8097. doi: 10.14778/3603581.3603588. URL https://dl.acm.org/doi/10.14778/3603581.3603588.

Anli Yan, Ruitao Hou, Xiaozhang Liu, Hongyang Yan, Teng Huang, and Xianmin Wang. Towards explainable model extraction attacks. *International Journal of Intelligent Systems*, 37(11):9936–9956, November 2022. ISSN 0884-8173, 1098-111X. doi: 10.1002/int.23022. URL https://onlinelibrary.wiley.com/doi/10.1002/int.23022.

Anli Yan, Ruitao Hou, Hongyang Yan, and Xiaozhang Liu. Explanation-based data-free model extraction attacks. *World Wide Web*, 26(5):3081–3092, September 2023a. ISSN 1386-145X, 1573-1413. doi: 10.1007/s11280-023-01150-6. URL https://link.springer.com/10.1007/s11280-023-01150-6.

Anli Yan, Teng Huang, Lishan Ke, Xiaozhang Liu, Qi Chen, and Changyu Dong. Explanation leaks: Explanation-guided model extraction attacks. *Information Sciences*, 632:269–284, June 2023b. ISSN 00200255. doi: 10.1016/j.ins.2023.03.020. URL https://linkinghub.elsevier.com/retrieve/pii/S002002552300316X.

Guang Yang, Qinghao Ye, and Jun Xia. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, 77:29–52, January 2022. ISSN 15662535. doi: 10.1016/j.inffus.2021.07.016. URL https://linkinghub.elsevier.com/retrieve/pii/S1566253521001597.

Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 225–240, London United Kingdom, November 2019. ACM. ISBN 978-1-4503-6747-9. doi: 10.1145/3319535.3354261. URL https://dl.acm.org/doi/10.1145/3319535.3354261.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Sai Suggala, David I Inouye, and Pradeep Ravikumar. On the (In)fidelity and Sensitivity of Explanations. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 10967–10978, December 2019.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, Oxford, July 2018. IEEE. ISBN 978-1-5386-6680-7. doi: 10.1109/CSF.2018.00027. URL https://ieeexplore.ieee.org/document/8429311/.

Xuefei Yin, Yanming Zhu, and Jiankun Hu. A Comprehensive Survey of Privacy-preserving Federated Learning: A Taxonomy, Review, and Future Directions. *ACM Computing Surveys*, 54(6):1–36, July 2022. ISSN 0360-0300, 1557-7341. doi: 10.1145/3460427. URL https://dl.acm.org/doi/10.1145/3460427.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-Cost High-Power Membership Inference Attacks. 2024. URL https://openreview.net/forum?id=sT7UJh5CTc.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, and Jiliang Tang. The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG), February 2024. URL http://arxiv.org/abs/2402.16893. arXiv:2402.16893 [cs].

Xiaoyu Zhang, Chao Chen, Yi Xie, Xiaofeng Chen, Jun Zhang, and Yang Xiang. A survey on privacy inference attacks and defenses in cloud-based Deep Neural Network. *Computer Standards & Interfaces*, 83:103672, January 2023. ISSN 09205489. doi: 10.1016/j.csi.2022.103672. URL https://linkinghub.elsevier.com/retrieve/pii/S0920548922000435.

Yifei Zhang, Dun Zeng, Jinglong Luo, Xinyu Fu, Guanzhong Chen, Zenglin Xu, and Irwin King. A Survey of Trustworthy Federated Learning: Issues, Solutions, and Challenges. *ACM Transactions on Intelligent Systems and Technology*, 15(6):1–47, December 2024. ISSN 2157-6904, 2157-6912. doi: 10.1145/3678181. URL https://dl.acm.org/doi/10.1145/3678181.

Zijiao Zhang, Chong Wu, Shiyou Qu, and Xiaofang Chen. An explainable artificial intelligence approach for financial distress prediction. *Information Processing & Management*, 59(4):102988, July 2022. ISSN 03064573. doi: 10.1016/j.ipm.2022.102988. URL https://linkinghub.elsevier.com/retrieve/pii/S0306457322001030.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for Large Language Models: A Survey. *ACM Transactions on Intelligent Systems and Technology*, pp. 3639372, January 2024. ISSN 2157-6904, 2157-6912. doi: 10.1145/3639372. URL https://dl.acm.org/doi/10.1145/3639372.

Jiaqi Zhao, Hui Zhu, Fengwei Wang, Rongxing Lu, and Hui Li. Efficient and privacy-preserving tree-based inference via additive homomorphic encryption. *Information Sciences*, 650:119480, December 2023. ISSN 00200255. doi: 10.1016/j.ins.2023.119480. URL https://linkinghub.elsevier.com/retrieve/pii/S0020025523010654.

Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Lim. Exploiting Explanations for Model Inversion Attacks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 662–672, Montreal, QC, Canada, October 2021. IEEE. ISBN 978-1-6654-2812-5. doi: 10.1109/ICCV48922.2021.00072. URL `https://ieeexplore.ieee.org/document/9709977/`.

Xiubin Zhu, Dan Wang, Witold Pedrycz, and Zhiwu Li. Horizontal Federated Learning of Takagi–Sugeno Fuzzy Rule-Based Models. *IEEE Transactions on Fuzzy Systems*, 30(9):3537–3547, September 2022. ISSN 1063-6706, 1941-0034. doi: 10.1109/TFUZZ.2021.3118733. URL `https://ieeexplore.ieee.org/document/9565342/`.

Ana Šarčević, Damir Pintar, Mihaela Vranić, and Agneza Krajna. Cybersecurity Knowledge Extraction Using XAI. *Applied Sciences*, 12(17):8669, August 2022. ISSN 2076-3417. doi: 10.3390/app12178669. URL `https://www.mdpi.com/2076-3417/12/17/8669`.