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

Sonal Allana*

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

Mohan Kankanhalli

School of Computing, National University of Singapore, mohan@comp.nus.edu.sg

Rozita Dara

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

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.

CCS CONCEPTS • **Computing methodologies~Artificial intelligence**; • **Security and privacy**; • **General and reference~Document types~Surveys and overviews**;

Additional Keywords and Phrases: XAI, explainability, privacy, attacks, defenses, characteristics

---

* Corresponding author.

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 [111] and introduces problems such as lack of fairness when data is biased, or irrelevant results when data is incomplete or outdated [173]. 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 is primarily a trial-and-error approach, where high accuracy is targeted by tweaking the hyperparameters such as the learning rate, epochs, batch sizes or architecture of the system, such as the number or type 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 [162], is often out of reach of AI systems. Explanations are crucial in high-risk applications [117] in domains such as healthcare [43, 50, 191], finance [50, 200], defense [50], justice [42], energy and power [105] where the impact on human life and well-being is significant [88, 109, 128] and the inability to do so deters their successful implementation [128, 153, 191].

Trustworthy AI strives to mitigate risks arising from AI systems due to possible harms from their data-driven nature. Trustworthiness is based on foundation principles of reliability, validity, robustness, privacy, explainability, fairness [5, 165] and other fundamental qualities that 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 [30]. There are numerous benefits of including explanations in AI models. Besides aiding data scientists in getting a better understanding of the data [76] and performing required data cleansing [32], explanations can help developers in detecting errors in input and determining features that can be modified to change the outcome [41]. When multiple models are available with similar accuracy, an explanation method can help to choose between models [44]. Interpretable models can enable knowledge discovery by detecting knowledge or patterns that were missed by uninterpretable ones [89]. Since humans remain an important component in the decision-making process as end-users and consumers of automated decisions [167], explanations can give them an understanding of the model outcome especially when they are adversely affected by the decisions [4]. It can also facilitate privacy awareness in end-users [23], enabling them to make right choices for their personal data and aid regulators and compliance officers to understand the compliance of models [109] with applicable regulations. With generative AI (Gen-AI) and large language models (LLMs) entering mainstream, explanations constitute an important design principle [183] in enabling a better mental model for users [160] and in communicating its capabilities and limitations to them [183]. It can also support users in effective prompt engineering to determine the words that impact the output of a model [115] and in verification of generated content to mitigate the problem of hallucinations [146].

1.2 Challenges for privacy in explainability

In many application domains of AI such as healthcare, finance and justice, training models on sensitive personal information is inevitable for usefulness of these systems [176]. For instance, a lung cancer detection model would require training on chest X-ray images, which constitutes personal information of patients. Similarly, a loan evaluation model of a bank, would require training on the financial health of customers, which is also personal to 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 [156] which can be exploited by adversaries for extraction of personal information. Gen-AI models create new content from large datasets in text, audio and image formats [160] which could potentially contain sensitive personal information [112]. LLMs are also found to leak the privacy of the data used in prompts during in-context learning [48, 166]. Due to such privacy risks involved, when personal data is used in training, testing, or inferencing on AI models, they become subject to data regulation and privacy acts [78].

Explainability is a foundation principle of Trustworthy AI, however, introducing explanations in AI systems is found to create a tension with the privacy of the system. Explanation interfaces are found to give adversaries an additional attack surface [49, 99] to mine the information contained in the model. Researchers have demonstrated privacy attacks targeting explanations to retrieve information about membership in the training set [99, 126, 151], build surrogates of the target model [3, 181, 190], infer sensitive attributes of individuals [49, 103] and reconstruct the complete training set [151]. These attacks are 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 [114] and hence valuable and confidential to organizations [185] 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 [103, 133, 190]. Due to these concerns of the privacy vulnerabilities of explanations, necessary privacy preservation measures are required in XAI systems [3, 151, 203].

1.3 Main contributions

Previous research has identified that the privacy issues in explainability are insufficiently studied [99, 103, 126] 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.
- *Exploration of applicable privacy preservation methods in XAI.* We determine the privacy preserving 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 a checklist for achieving the tradeoff 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 our 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 explanators to answer RQ1. In Section 5, we elaborate the use of privacy preserving methods on explanators 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 that enhance understanding and trust of end-users [68]. 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" [68]. The study established users' preference for systems with explanations over systems that provided only decisions. Ribeiro et al. [140] refer to explanations of predictions as qualitative artifacts that provide the relationship between an input instance and the output prediction.
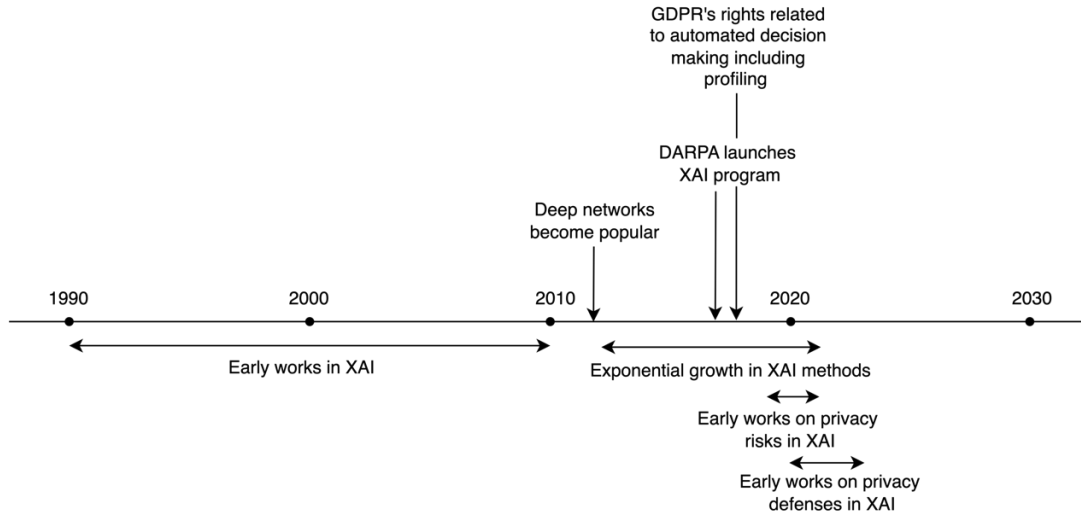

Figure 1: Important milestones and inception of privacy attacks/defenses in XAI.

2.2 Evolution of XAI and corresponding privacy issues

The field of explainability can be traced to the early 1990s, driven by the lack of clarity on the outputs of black-box models. Early works [12, 37, 96, 113, 170] have used different approaches for extracting rules from these systems. The rise of deep

learning and the improvement in the accuracy of black-box systems, propelled complex uninterpretable systems into mainstream usage. However, their use in critical domains remains a challenge due to their lack of transparency. Regulations, such as the GDPR's rights related to automated decision-making including profiling [63], contributed to bringing the urgent need for explanations and hence the field of XAI into the forefront and resulted in an exponential growth in proposed explainability methods. However, the introduction of transparency through XAI methods also introduced privacy leakage through explanation interfaces described in early works [114, 150, 151] in the form of privacy attacks on the training data and the underlying model. Correspondingly, researchers have also explored defenses and pioneering works in this field [71, 133] have proposed different strategies for generating privacy preserved explanations. Currently the privacy issues in XAI are far from being resolved, with new attacks being discovered and researchers proposing corresponding mitigation measures to alleviate the risks. Figure 1 summarizes the main milestones in the field of XAI and indicates the timeline of discovery of privacy issues and proposed defenses.

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 [71]. 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 [50]. 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 [8]. 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 [8, 120, 138] are some examples of interpretable models. According to Arrieta et al. [8], 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. A major setback to their successful adoption as explainable-by-design methods, is their lower accuracy [16, 53, 68] 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 [15]. For models such as decision trees, the understandability deteriorates as the complexity, i.e., the depth of the tree increases [145]. Nonetheless, due to their intrinsically transparent architecture, interpretable models are often used as surrogates for black-box models [109]. The use of multiple surrogate models facilitates the availability of different types of explanations [50] 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 [191].

2.3.2 Example-based methods

These methods use examples, i.e., data instances as samples to explain the model [109]. The instances may be from the training set or generated by the method [84, 95]. These methods are also referred to as record-based [150], instance-based [65] or case-based [122] methods in literature. They can complement feature-based methods to aid understandability of the end users [83] and also improve the interpretability of complex distributions [89]. They are intuitive and natural in their ability to provide explanations to humans [84]. Some methods in this category are anchors [141], contrastive explanations [44], counterfactuals [178], influence functions [90] and prototypes and criticisms [89].

Example-based explanations, though easily interpretable by end-users, can cause breach of privacy when datapoints are revealed as explanations [151, 176]. 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 [105].

2.3.3 Knowledge-based methods

These methods utilize knowledge representation techniques in machine learning (ML) models [168]. The integration of background knowledge [74] facilitates the inclusion of context [93, 132], thus increasing the trustworthiness of explanations. Neuro-symbolic [74] or in-between methods [79] is an emerging field that explores integration of symbolic AI approaches from knowledge representation and reasoning with subsymbolic or connectionist based approaches [73]. A neuro-symbolic hybrid can enable the integration of the resilience of neural approaches with the interpretability of symbolic approaches. Semantic web technologies offer tools for semantic interpretation and reasoning from knowledge bases [147]. Knowledge graphs, expressed using Resource Description Framework, and ontologies, expressed using Web Ontology Language, are the common tools that can be deployed for explainability. Knowledge graphs have applicability as pre-model and post-model XAI for extracting features and relations from data as well as for inferencing and reasoning [137]. The field of semantic web technologies in explainability is attractive because of its potential of creating knowledge-rich explanations without adversely affecting the performance of the system [147].

2.3.4 Feature-based methods

These explanation methods score or measure the effect of individual input features on the output of the model [8, 14, 50, 159]. They are also referred to as feature importance [109], feature relevance [8] or attribution-based [99] 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 [161]. Two important categories of feature-based methods identified in literature are perturbation and backpropagation-based methods [7, 109].

- *Perturbation-based methods* remove, alter, or mask an input feature or set of features and observe the difference with the original output [109]. Some perturbation-based methods are LIME [140], permutation feature importance [22], SHAP [101] and MASK [59].
- *Backpropagation-based methods* compute input attributions in forward and backward passes of the network [7]. The use of the gradient of the output with the respective input features [109, 159] 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 [86, 120]. Some examples of backpropagation-based methods are gradient [155], gradient x input [154], guided backpropagation [158] and integrated gradients [162].

Compared to backpropagation, perturbation-based methods require running the model with different sets of input, hence they are slower [7] and increasing the number of features increases the performance time [86]. 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 [86]. Though feature-based explanations are popular and used by many Machine Learning as a Service (MLaaS) platforms [103], the explanations though useful to researchers, are found to be difficult to understand by end-users [176].

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	[141]
	Contrastive explanations		Model-agnostic	[44]
	Counterfactuals		Model-agnostic	[178]
	Influence functions		Model-agnostic	[90]
	Prototypes and criticisms		Model-agnostic	[89]
Knowledge-based	Semantic web technologies		Model-agnostic	[147]
	Neuro-symbolic approaches		Model-specific	[74]
Feature-based	Perturbation-based	LIME	Model-agnostic	[140]
		Permutation Feature Importance	Model-agnostic	[22]
		SHAP	Model-agnostic	[101]
		MASK	Model-agnostic	[59]
	Backpropagation-based	Gradient	Model-specific	[155]
		Gradient x Input	Model-specific	[154]
		Guided Backpropagation	Model-specific	[158]
		Integrated Gradients	Model-specific	[162]

2.4 Related reviews on XAI

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 fills this gap and is unique in comparison 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. [8]. Mohseni et al. [118] executed an interdisciplinary survey and provided a framework for XAI design and evaluation methods. Dwivedi et al. [50] covered a wide breadth of explanation algorithms, programming frameworks and software toolkits for XAI development. Ali et al. [4] surveyed explainability through the lens of trustworthiness and elaborated on evaluation metrics, packages and XAI datasets. Bodria et al. [19] categorized explanation methods and benchmarked popular methods using quantitative metrics. Muralidar et al. [125] reviewed transparency elements from HCI in the context of explanations while Cambria et al. [25] explored the presentation methods and usage of natural language with XAI. In addition to these broad surveys on explainability, researchers have also surveyed specific application domains such as healthcare [135, 191], cybersecurity [27] and energy and power systems [105]. Reviews on specific XAI categories and methods also exist, such as counterfactuals [67], data-driven knowledge-aware XAI systems [95], knowledge-graph based XAI [137, 168] and semantic web technologies for explanations [147]. The use of XAI with federated learning, termed as Federated XAI (Fed-

XAI), was reviewed by López-Blanco et al. [100]. H. Zhao et al. [201] surveyed explainability for transformer-based language models and categorized methods based on training paradigms.

The review in this article differs from the above reviews in that it explores the risks to privacy arising from including explainability in AI systems. Further, we review strategies used by researchers in mitigating the privacy leakage in XAI. We use an established scoping review strategy guided by well-defined research questions. Our taxonomy of XAI, privacy risks and corresponding mitigation methods are distilled from the widely accepted understanding of the privacy and XAI community in existing literature. Our review methodology facilitates the answering of the identified research questions from the extracted 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) [172]. In this section, we elaborate the process followed and the identified research trends.

3.1 Literature Selection and Extraction

We followed a 4-step process comprising of identification, screening, eligibility, and extraction as depicted in Figure 2. In the first step, namely, identification, Elsevier Engineering Village [54] search platform was used and the search was conducted on Compendex and Inspec databases, which indexes publications from major 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 [114, 150, 151] 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.
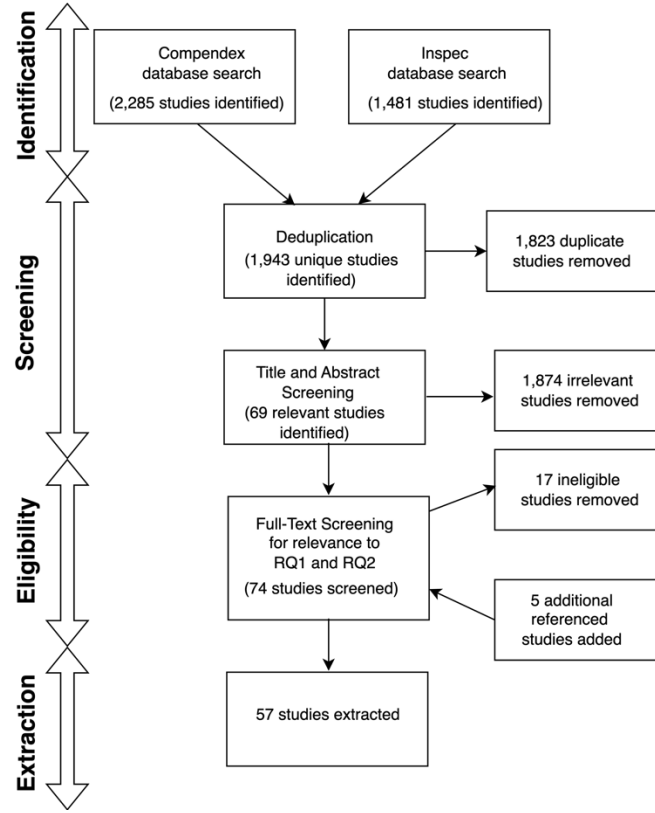
Figure 2: Scoping review process as per PRISMA-ScR.

The search results comprising of 3,766 studies were exported from Engineering Village and imported into Covidence [205] review management software. During the import process, the software eliminated duplicate studies retrieved from both databases, retaining only unique records. After deduplication, 1,943 studies were forwarded for screening. A researcher screened the title and abstract to determine relevance to RQ1 or RQ2 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, those that addressed only security issues in explainability, privacy issues in ML, or survey papers were eliminated. This resulted in removal of 17 studies. During this stage, 5 relevant studies that did not appear in the original search results, were found 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). As observed, there is an increase in reported privacy risks in XAI methods in the period considered. Correspondingly, there is also an increase in the number of studies on the use of various privacy preservation methods in XAI as observed from Figure 3 (b). Differential privacy and anonymization are the popular privacy preservation methods used with XAI methods. Among the privacy risks identified, 3 attacks, namely, membership inference, model inversion and model extraction, have similar

proportion of studies as seen in Figure 3 (c). One privacy attack, namely, property inference, is not tested in XAI systems. Figure 3 (d) shows the categories of XAI targeted by different attacks. As observed, feature-based and example-based XAI are mainly targeted for privacy attacks in comparison to interpretable methods. To the best of our knowledge, no attacks are reported on knowledge-based methods.
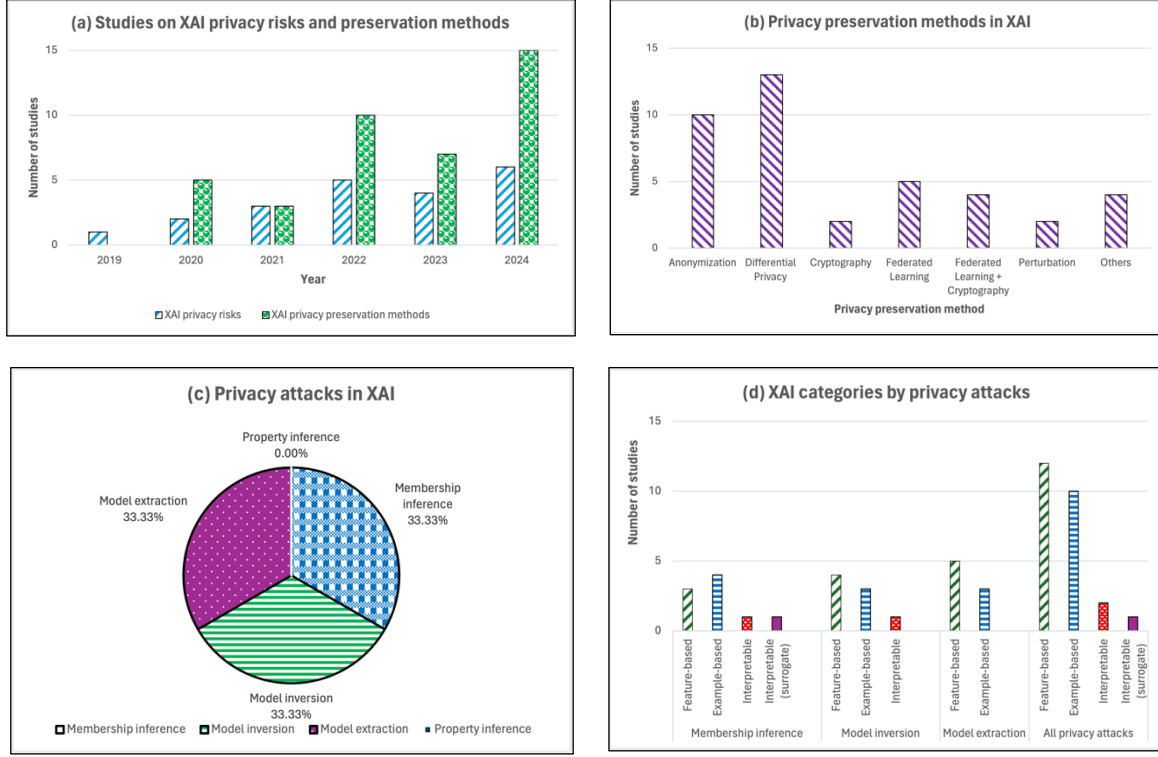

Figure 3: Research trends identified from extracted studies.

4 PRIVACY RISKS IN XAI

Traditionally privacy is referred to as the "right to be left alone" [182] 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" [184]. 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 [38]. Clarke 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" [34]. In this article, we refer to the last definition of privacy.

During the training process, ML models feed on large amounts of data. More data is provided to the models, in the form of queries, during inference time. If a model uses or creates personal data, it falls within the scope of data protection regulations [78]. Consider an example of a model that makes decisions on the credit card application of an individual. The

model would typically require query data in the form of personal details such as annual salary, age, credit rating and other sensitive fields to determine an approval or rejection. In this case, the outcome of the application is a decision made on the individual and is also considered as personal data. Usage of personal data in the training or inferencing process of models, make them subject to privacy regulations. In XAI, the output is accompanied by an explanation of the result, also referred to as transparency report, and provides further insights into the decision-making process of the model [151]. When the explanation contains or leaks personal data, it constitutes a privacy risk [203].

Trustworthy AI is built on explainability as one of its ingrained principles. Explainability assists in gaining insights into the functionalities of a black-box AI system [165], however, the relationship between privacy and explainability has contrasting aspects. Explainability aids privacy in several ways such as, in creating privacy awareness in users [23], in ascertaining that privacy of a system is achieved [46, 124], and in determining correlations with identifiable data for removal [76]. On the downside, explanations can reveal sensitive information contained in models and training data [71, 138, 203] thus leading to privacy risks [92]. Thus, there are conflicting outcomes [30, 66, 130, 144, 157] of including explainability as a non-functional requirement in AI systems. In the following subsections, we answer RQ1 by discussing the threat model of XAI to elaborate the assets, actors, and the type of privacy risks, AI systems with explanations are vulnerable to.

4.1 System Assets

To understand the privacy risks in XAI, we first focus on the assets (Figure 4) that need privacy protection. Assuming that the model has been trained on personal data and that the model also needs personal data in the form of queries for generation of outputs and explanations, the following potential artifacts are vulnerable to privacy breaches:

1. training data
2. query data
3. model outputs (prediction or classification or recommendation in discriminative models or generated content in generative models [160])
4. explanation of outputs (i.e., transparency reports)
5. model architecture and hyperparameters

The first four assets above, can be intuitively understood as containing personal data to different degrees. If the model is not trained on personal data and does not output any personally identifiable or sensitive information, it falls outside the scope of privacy risks [78]. The fifth asset, i.e., model architecture and hyperparameters, is susceptible to loss of confidentiality. Organizations spend manpower, time, and effort in collecting and curating data and in training and deploying their models. Hence it holds commercial value [198] and an exposure would breach intellectual property [16, 139, 198], proprietary value [114] and confidentiality [105, 176]. Since models can be argued as sources of data, allowing data to be estimated from them with different degrees of accuracy [175], a breach of the model can lead to breach of training and query data, thus leading to privacy risk.
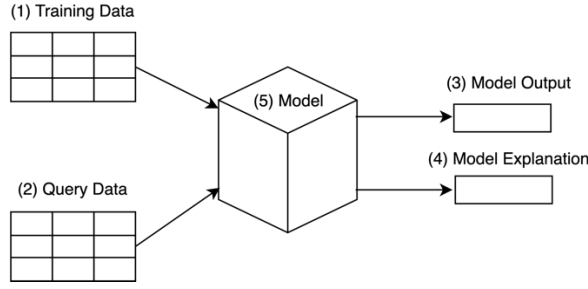
Figure 4: Assets that pose privacy risks in XAI.

4.2 System Actors

The AI lifecycle [129, 149] (Figure 5) consists of a data acquisition phase where pertinent datasets are explored and preprocessed as needed. This is followed by model development where the appropriate model is chosen and hyperparameters decided. The model is trained on the datasets, and its performance evaluated on criteria such as accuracy and training time. Finally, the model is deployed in the real-world setting. Most AI models require monitoring and appropriate tuning while in this phase. Models eventually become obsolete and require appropriate retirement at the end of the cycle. XAI, being an AI system, follows the phases of the AI lifecycle. Data curators, developers, software architects and quality engineers [30] are the main actors involved in the initial phases involving development and evaluation. Once the system is deployed in production, it is accessible by various end-users through APIs or GUIs as needed. While in production, valid actors such as developers and quality engineers may need to see transparency reports when end-users challenge results and during calibration of AI systems [10].
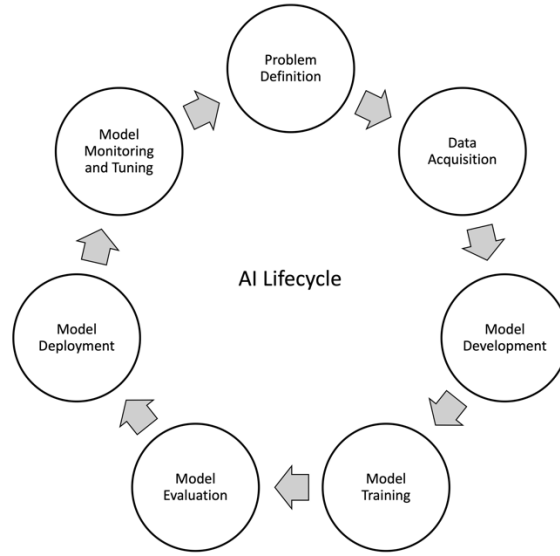

Figure 5: Different phases of AI lifecycle.

During different phases, the system can become accessible to adversaries. These may secure different levels of access to an XAI. In white-box access, adversaries possess information about the model internals such as architecture and hyperparameters [97, 198]. In black-box access, adversaries access the input/output of the system [97] without any knowledge of the training process or model internals [142]. Any intermediate access between white-box and black-box is referred to as gray-box [80]. While unintentional privacy leaks may occur during the different lifecycle phases or during the course of an explanation [138], adversaries are associated with intentional privacy breaches. They may act as passive observers [80] and use the model outputs for launching privacy attacks, or they may actively interfere in the training process of the model [127]. Thus, XAI systems require adequate protection mechanisms from both intentional and unintentional privacy leakage (Figure 6).
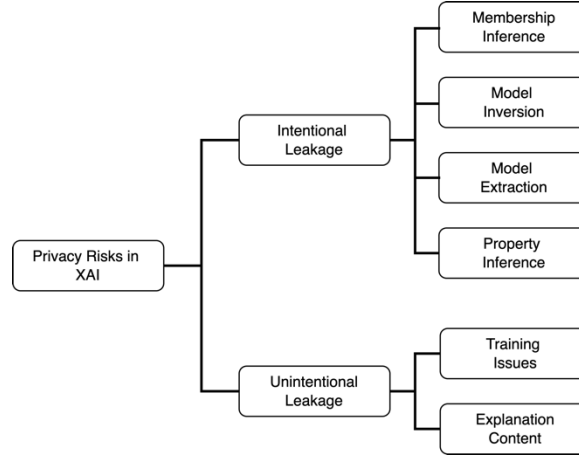
Figure 6: Categorization of privacy risks in XAI.

4.3 Intentional Privacy Leakage

In this subsection, we review intentional privacy risks in the form of privacy attacks by adversaries. Since XAI systems are essentially AI systems with an additional asset, i.e., explanations, they are vulnerable to malicious attacks applicable to AI models. Previous research has determined security attacks targeting AI, such as evasion and poisoning [136]. However, in our review we differentiate from these security attacks and focus on privacy attacks that target personal data of individuals or model confidentiality. Below we summarize these attacks in the XAI context where model explanations further aid [203] the identification or exposure of personal information of individuals or the intellectual property of the model owner. Table 2 summarizes these risks and the works detailing them.

4.3.1 Membership Inference

This is a privacy risk of identification of an individual in the training set of a model [152, 196] (Figure 7). In this attack, the adversary can have black-box or white-box access to the model [175] after it has been deployed in production. This risk is significant since datasets used in critical domains, may reveal sensitive information about individuals used in training the model [152]. An example of membership inference is when an individual's personal details, such as age, gender, medical conditions, and other attributes, are known to an adversary and is used to determine if he/she was part of the training data of a disease detection model, thus indicating the individual may have the disease with a high chance [77]. Overfitting of the target model is identified as a main cause of membership inference [82, 194].
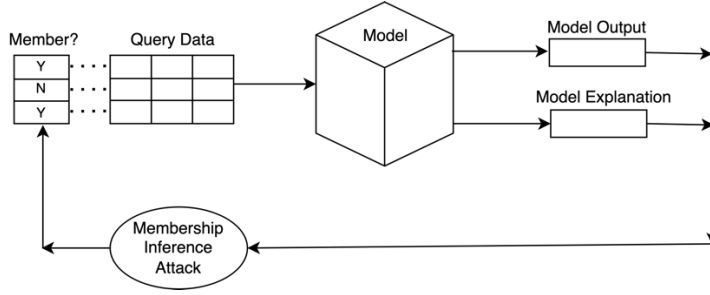
Figure 7: Membership inference exploiting explanations.

The seminal work on membership inference on feature-based and example-based XAI systems was demonstrated by Shokri et al. [150, 151]. 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. H. Liu et al. [99] proposed 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. [104] used Shapley value explanations to improve label-only membership inference, that employs hard prediction labels instead of confidence scores for determining membership [33]. Explanations were used to improve neighbourhood sampling thus reducing the number of queries.

In the example-based category, Shokri et al. [150, 151] considered influence functions on logistic regression models. Since the explanations from influence functions consist of actual datapoints, the study observed that attackers could obtain certainty about membership, thus leading to stronger attacks. Cohen and Giryes [35] considered self-influence functions instead, that show the influence of a datapoint on its own prediction. The demonstrated attack required white-box access to the target neural network parameters, activations, and gradients. The choice of a good threshold range for self-influence scores for members was crucial for this attack and was achieved by maximizing the balanced accuracy on the training set.

Kuppa and Le-Khac [92] used a different type of example-based explanation, namely, counterfactuals, for membership inference. The authors trained shadow models using counterfactual samples and auxiliary datasets. The use of a threshold, on the difference in predictions of the attack and target models, was employed to determine membership. Pawelczyk et al. [134] 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 [28] that improved the attack.

Interpretable models using decision trees, and surrogate models created using Trepan algorithm [37], were evaluated for membership inference by Naretto et al. [126]. The authors also studied the effect of overfitting on the attack. The study found the success of membership inference higher on both interpretable and surrogate models in comparison to black-box models. The attack was also more successful on surrogates of overfitted models in comparison to well-regularized models.

Membership inference in machine learning models has been explored extensively in existing literature [77] and attacks have exploited confidence scores and predictions [152]. However, the above attacks that exploit explanations, suggest that XAI interfaces provide a new avenue for adversaries to launch this attack. So far, the attack has targeted feature-based, example-based, and interpretable (including surrogates) XAI methods. Factors such as dataset type [151], dimension [134,

151], model architecture [151] and overfitting [134], influence the effectiveness of this attack. Some demonstrated attacks are also possible without knowledge of the training dataset or target architectures [99].

Interpretable models are often suggested as surrogates for explaining black-box models [109], however as demonstrated by these attacks, this layer of interpretability can introduce a backdoor to the target system and lead to privacy leaks [126]. In the example-based category, influence functions display training datapoints that are influential for a point of interest and are a direct breach of privacy of the training data [151]. They can also lead to exposure of outliers due to their distinct characteristics and higher influence on the training process [151]. Among feature-based methods, those using perturbations are comparatively resilient to membership inference due to use of out-of-distribution points, however, this can also impact explanation fidelity [151]. Feature-based methods with better explanation quality are also found to be susceptible to higher leakage [99] suggesting a conflict between privacy and utility.

4.3.2 Model Inversion

This type of privacy risk is found to be of two types, namely, data reconstruction and training class inference [198]. In this attack, the adversary can have black-box or white-box access to the model [60, 175] after it has been deployed in production. In data reconstruction (Figure 8), individuals' data used in training or querying [203] the model is recovered partly or completely and constitutes a risk of exposure [45]. Attribute inference is a type of data reconstruction that can determine the values of certain attributes, generally those sensitive to individuals [65, 194] 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 [45, 192, 198].
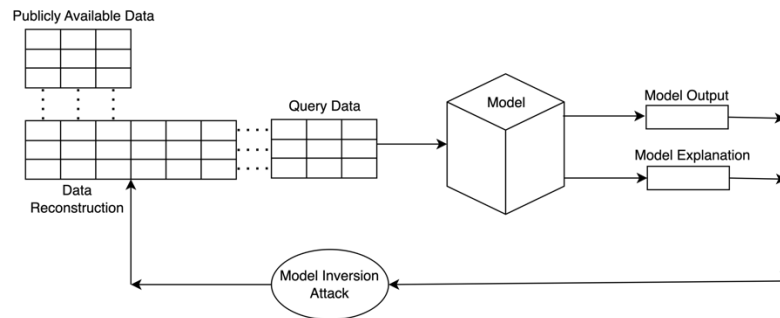

Figure 8: Model inversion exploiting explanations.

In XAI, model inversion attacks are reported on example-based, feature-based, and interpretable systems. Shokri et al. [150, 151] 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 points. Goethals et al. [64] 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 X. Zhao et al. [203] leading to inadvertent exposure. XAI systems that provided class-specific multiple explanations were prone to more leakage. The authors also used attention transfer to demonstrate the risk for non-explanation models. Other works, such as Duddu and

Boutet [49] and X. Luo et al. [103], have focused on attribute inference on 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 and Kikuchi [169] further showed that the efficacy of their proposed attack was dependent on the combination of black-box architecture and XAI method, with linear models using Shapley values, being more vulnerable to attribute inference.

Ferry et al. [56] 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 [49, 203] and constitute a privacy risk indicating the contradiction between the need for explanations in Trustworthy AI and protecting privacy [203]. Data reconstruction attacks impact active users of AI systems rather than training data as in membership inference, putting end-users at risk [203] and thus having a higher impact. In certain proposed model inversion attacks, sensitive attributes can be retrieved from models trained on non-sensitive attributes [49] while other proposed attacks demonstrate higher leakage from more important attributes [103] and recovery of entire training datasets [151]. In addition, the above works highlight the misuse of XAI techniques even for models that do not provide explanations [203].

There is also a conflict of privacy seen with the utility of the system. For instance, synthetic counterfactuals created by perturbing actual samples can provide resilience to inversion in comparison to native counterfactuals, however, their use is found to affect the plausibility and runtime of explanations [64], suggesting degrading utility. Similarly, for improving understandability of end-users, multiple diverse explanations are usually recommended [177], however they are also found to contribute to leakage of privacy, suggesting that explanation APIs should be restricted [203] thus adversely affecting the utility to end-users.

4.3.3 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 9). 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 [80]. Since the extracted models can further leak personal data through membership inference and model inversion [156], 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 [3, 116]. In a typical model extraction attack, the adversary has black-box access to a victim model in production and uses an unlabeled dataset to query it, thus generating labels to build an attack dataset [190]. The labeled attack dataset is then used for training the cloned model. In data-free model extraction, the attacker does not need to collect the attack dataset and instead generative models can be used to build it, making it convenient for adversaries to run this attack when the input data is hard to find [116].
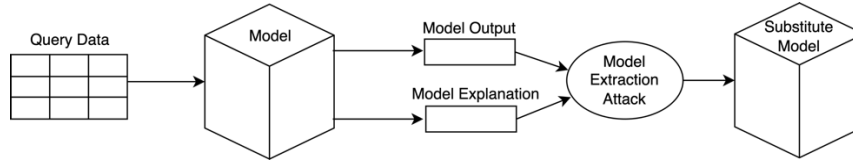
Figure 9: Model extraction exploiting explanations.

Model extraction using explanations is proposed on 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. [114]. The use of explanations improved the attack by reducing the number of queries in comparison to using model predictions. Miura et al. [116] also leveraged gradient-based explanations but used data-free approach to train generative models for creating the attack dataset. Explanations were found to improve the quality of the generative model, and the accuracy of the cloned model improved with increase in the diversity of the generated attack dataset. The approach of Yan, Hou, et al. [189] was also based on data-free extraction, however the authors used explanation loss to guide the generative model. Accuracy of the cloned model was improved by matching the victim model's predictions and explanations. A model agnostic extraction technique applicable to both gradient and perturbation-based XAI was proposed by Yan, Huang, et al. [190]. Explanations were found to provide auxiliary information to adversaries, leading to efficient attacks with improved query budgets, compared to those using predictions. The attack could also be applied to non-explanation models and achieved accuracy equivalent to those of explanation models. In another work, Yan et al. [188] used a multitask learning architecture to extract both classification and explanation tasks of the victim model.

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. [3], 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 Y. Wang et al. [181]. The method used the original counterfactual explanation with its counterfactual as training pairs, to extract additional datapoints to train the cloned model. Kuppa and Le-Khac [92] proposed iterative querying of the victim model to capture the training data distribution and utilization of distillation loss to transfer knowledge from the victim to the cloned model. The attack was found to be successful due to the optimization of various properties such as diversity, proximity, feasibility, and sparsity in the counterfactual method used.

The misuse of XAI techniques for cloning models with high accuracy and fidelity, is a threat to the confidentiality of the model owner. The above attacks demonstrate the exploitation of XAI techniques to facilitate extraction of victim models efficiently with reduced number of queries in comparison to extraction attacks based solely on predictions [114, 116]. The reduction in the number of queries to conduct this attack by using explanation interfaces without access to predictions, is a significant gain for attackers, especially in pay-by-query models. Certain attacks are also possible with partial knowledge of the data distribution [3] or even when the attack dataset has no overlap with the training dataset [188]. 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 have improved accuracy of the attack [116]. The diversity of the generated input datasets in such attacks is also found to improve the accuracy of the cloned models [116].

In addition to the direct threat to explanation models, XAI techniques can be misused for extraction of non-explanation models [190]. The use of diverse explanations, that is useful to build trust of users in the explanations, can lead to further

leakage of privacy [3]. The optimization of counterfactuals to satisfy various properties to improve explanation quality, can provide more information to adversaries about class-specific decision boundaries thus aiding the attack [92] and leading to the conflict of explainability and privacy with utility.

4.3.4 Property Inference

This type of risk reveals properties from the training data such as global statistics or aggregates [106], which model owners did not intend on sharing [61] (Figure 10). The inferred property need not be a feature in the training data or correlated to any feature, for instance, deducing the ratio of different genders in the training set [126] or the aggregate employee sentiments through internal company emails used to train a spam classifier [106]. Thus property inferencing can lead to exposure of sensitive information and constitutes a privacy risk. Though this privacy risk is a known issue in AI models, to the best of our knowledge, no attacks are reported so far using explanations.
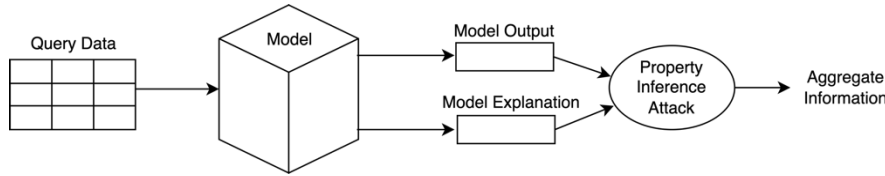

Figure 10: Property inference exploiting explanations.

Table 2: Studies on intentional privacy leakage in XAI systems.

Privacy Risk	XAI Category	XAI Method	Study
Membership inference	Interpretable	Decision tree	[126]
	Interpretable (surrogate)[a]	Trepan	[126]
	Example-based	Influence functions	[150, 151]
		Counterfactuals	[92, 134]
		Self-influence functions	[35]
	Feature-based	Gradient, integrated gradients, guided backpropagation, LRP, LIME, SmoothGrad	[150, 151]
		Integrated gradients, SmoothGrad, VarGrad, Grad-CAM++, SHAP, LIME	[99]
		Shapley values	[104]
Model inversion	Interpretable	Decision tree, rule list	[56]
	Example-based	Influence functions	[150, 151]
		Counterfactuals (native)	[64]
	Feature-based	Gradient, gradient x input, class activation maps (CAM), Grad-CAM, LRP	[203]
		Integrated gradients, DeepLIFT, GradientSHAP, SmoothGrad	[49]
		Shapley values	[103, 169]
Model extraction	Example-based	Counterfactuals	[3, 92, 181]
	Feature-based	Gradient	[114]
		Gradient, Grad-CAM, MASK	[188]
		Gradient, Grad-CAM, MASK, LIME	[190]
		Grad-CAM, LIME	[189]

Privacy Risk	XAI Category	XAI Method	Study
		Gradient, integrated gradients, SmoothGrad	[116]
Property inference	Not reported	Not reported	-

^a Interpretable (surrogate) systems use an interpretable model as a surrogate for explaining the black-box.

4.4 Unintentional Privacy Leakage

In this subsection, we discuss unintentional privacy leakage in XAI that occurs without malicious intent [80]. Some leakage can occur due to the natural mechanisms of the training process while some can be caused through the content of explanations.

4.4.1 Training issues

Training issues such as, overfitting and memorization, identified in AI models can lead to privacy leakage. Overfitting occurs when the model performs better on the training data in comparison to test data [80] and is found to aid membership and attribute inference attacks [194]. Memorization refers to the model remembering subsets of training data [156] and occurs during training before overfitting begins [80]. It can cause leakage when data owners deploy models with code written by other parties, such as in MLaaS, allowing sensitive information to be leaked from training data [156].

4.4.2 Explanation content

The content of transparency reports 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. [88] provide another example of unintentional leakage through contrastive explanations that can lead to inference of sensitive details of individuals whose partial attributes are known. Interpretable models used as surrogates, can reveal properties of the training data or additional information about the black-box than was intended to share [16]. In addition to this type of direct exposure, the explanation may be exposed to unintended users due to lack of appropriate access control [92]. For example, during troubleshooting of error cases, a developer or quality engineer may inadvertently access sensitive personal information in the explanation. Explanation content containing proxies or correlated fields can also lead to inferring of sensitive fields.

5 PRIVACY PRESERVATION METHODS IN XAI

To mitigate the privacy risks discussed in Section 4, many works have highlighted the need for privacy preserving XAI techniques [3, 151, 203]. In response to these calls for increasing safety in explanations by stemming privacy leakage, researchers have begun efforts in exploring algorithms that provide transparency reports in a privacy preserving manner. Many works have leveraged existing techniques from privacy preserving ML (PPML) and have studied their applicability to explanations. In this section, we consolidate the solutions proposed in literature for improvement of XAI methods in response to RQ2. We categorize these solutions under the main approaches in PPML. Table 3 summarizes these approaches and methods, and Figure 11 consolidates the proposed privacy preservation methods for specific privacy attacks.
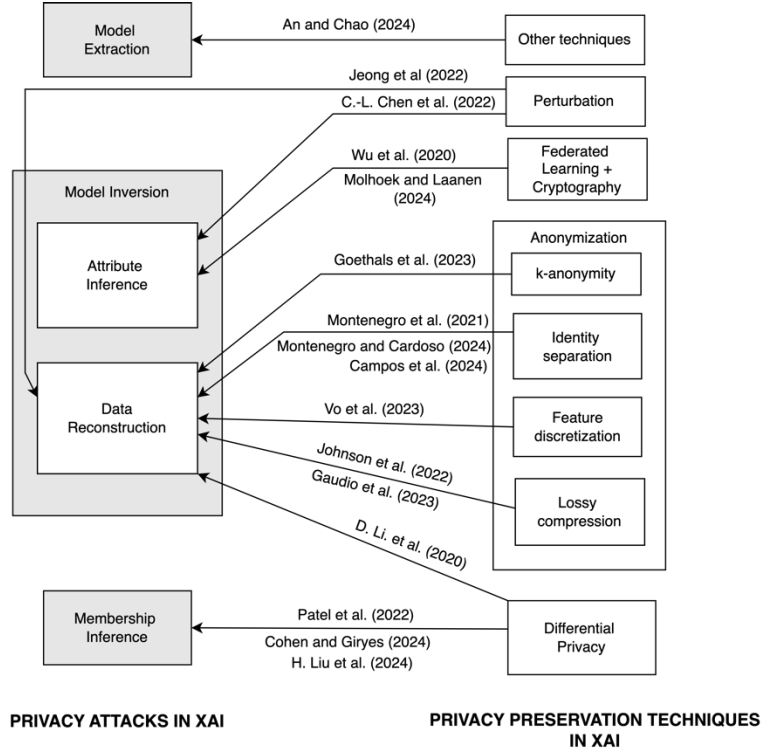
Figure 11: Proposed privacy preservation methods for specific privacy attacks in XAI.

5.1 Differential privacy

Differential privacy (DP) [51] is a well-known technique that provides a quantifiable definition of privacy and the incremental privacy loss from publishing confidential data [110]. A mechanism is differentially private if it can hide the participation of any single individual in a dataset [71]. This can be achieved by using noise and is typically associated with an adverse effect on the accuracy of the system [71]. By adjusting the privacy budget, ϵ, from 0 to ∞, practitioners can manage this trade-off between privacy and accuracy [110]. Because of its strong privacy guarantee, early works in privacy preserving explanations have adopted DP, using different strategies, to protect 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 [133] of the model. The technique can be applied on the explanation generation algorithm [133], the training process of the target model [35, 99] or on the training data [21, 55].

Decision trees are popular due to their simplicity and inherent interpretable qualities, however they are prone to privacy leakage [56, 126]. Various algorithms for building decision trees based on DP guarantees are proposed in previous works [58] with different trade-offs in privacy and utility. The interpretability of private trees and hence their usefulness to XAI users, depend on factors such as the privacy budget per query, tree depth, pruning strategies and termination criteria [58]. Nori et al. [131] privatized another type of interpretable model, namely, Explainable Boosting Machines (EBM), using DP for prevention of privacy leakage of training data. The privatized system offered good accuracy at low privacy budgets, allowed rectification of errors introduced by noise, removal of bias and enforcing constraints such as monotonicity. Baek

and Chung [9] further enhanced the utilization of privacy budget in these models to improve accuracy, by optimization of gradient error and pruning of non-important features.

An interpretable model using differentially private locally linear maps with Gaussian mechanism per output class was proposed by Harder et al. [71]. 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 SmoothGrad. However, increasing the number of such maps per class dropped accuracy due to the distribution of privacy budget over additional parameters. A different interpretable model, in the form of feedforward-designed convolutional neural network (FF-CNN), was made privacy preserving by D. Li et al. [94] using DP guarantees on subspace approximation with adjusted bias (Saab). The use of DP was found to be 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. [133] introduced a differentially private approach for loss calculation in the explanation algorithm. The work further proposed an adaptive method of reusing previous explanations for prudent usage of the privacy budget. Nguyen et al. [130] employed local DP to restrict adversaries from learning the top influential features through aggregated scores in feature-based XAI. Though, this was proposed as a solution to a backdoor security attack exploiting explanations [148], the random perturbation of influential features under local DP guarantees, was observed to preserve the privacy of those features while maintaining explanation fidelity. Bozorgpanah and Torra [21] also applied local DP to mask the training dataset and determine its impact on privacy and utility of feature-based explanations. An irregularity metric was defined to measure the feature distortion due to privatization of the original dataset and the change in explanation values. The use of additive noise on the training dataset was observed to introduce irregularities, reducing the utility of the explanations. Ezzeddine et al. [55] added calibrated noise to training datasets and evaluated the impact on SHAP explanations using various distance metrics. The change in SHAP values in the privatized systems was found to be correlated to the privacy budget and data dependent. Abbasi et al. [1] explored the use of DP for generation of synthetic data for training of different types of model architectures and used similarity score to track the change in explanations and utility loss to evaluate drop in accuracy, thus quantifying the triad of privacy, utility and explainability.

Besides these works, some researchers have also explored the use of DP in the training process of the target model in feature-based [99] and example-based [35, 117] XAI, with mitigation observed against membership inference for high privacy budgets [35, 99]. DP noise enables the regularization of target models [131] and its mathematical guarantee enables quantification of privacy, making it an attractive privacy enhancing technology. In the context of XAI, in addition to its application in mitigating membership inference from explanations [35, 99, 133], it is also found to mitigate reconstruction of sensitive inputs [94]. However, the improvement in mitigation of attacks at high privacy budgets and hence degrading accuracy [35, 99] 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 [99, 133]. This impact can be more pronounced for minority groups [133]. The technique is also ineffective against attribute inference attacks when there are existing strong correlations between different attributes [31].

Algorithms such as DPSaab [94], are observed to be better at accuracy and privacy trade-off than others. Strategies such as, reusing previously generated differentially private explanations, can be employed to utilize the privacy budget effectively [133]. Methods using local DP, are observed to achieve good faithfulness of explanations with privacy [130], thus demonstrating that it is possible to balance multiple desirable properties. Hence careful use of this technique is required in a way that privacy can be achieved while retaining 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. Popular methods in this category include homomorphic encryption, secret sharing, and secure multi-party computation [195]. In XAI, this technique has been limitedly explored in interpretable and example-based systems. It has also been used in conjunction with other privacy preserving techniques such as federated learning [53, 119, 186, 187].

For interpretable tree-based models, J. Zhao et al. [202] proposed an additive homomorphic scheme for model owners and query users, to push 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 producing accuracy comparable to non-private inference. In the example-based category, Veugen et al. [176] 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, to create an explanation for a query data point. In addition, a synthetic data point from the foil leaf was provided to the end-user to enhance explainability.

Cryptographic methods, such as homomorphic encryption, add to the computational complexity [97] of the system. The use of encryption also makes the model less transparent to data scientists for correction of errors, inspection of data, addition of features or fine tuning of the model [47], thus adversely affecting interpretability. Hence it is essential to implement cryptographic protocols in XAI system components in the right use cases to enhance other privacy preservation techniques or when other techniques are infeasible or costly.

5.3 Anonymization

Anonymization is a technique of transformation of data [107] to obscure the distinctive features of records, thus protecting their privacy. The process is associated with the removal or modification of direct and quasi-identifiers [107], that can uniquely identify individuals. Various methods of anonymization exist, such as k-anonymity, l-diversity, and t-closeness [195]. In XAI, different anonymization techniques have been proposed in example-based, feature-based, and interpretable methods. The use of techniques such as disentangled representation learning and lossy compression on sensitive images such as medical data, are found to generate privatized explainable-by-design images.

A dataset is considered k-anonymous if every record is indistinguishable from k-1 other records [163] thus providing a measure of the risk of re-identification of records. K-anonymity can be achieved using methods such as suppression and generalization [164]. Though traditionally this technique is applied to target datasets for protection, Goethals et al. [64] 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 found to have lower information loss and higher validity; outperforming counterfactuals generated from k-anonymized datasets. Berning et al. [13] further analyzed k-anonymous counterfactuals and determined its effectiveness only in dense areas of the dataset. The study found its real privacy level disproportionate to the value of k. According to Vo et al. [177], the generation of k-anonymous counterfactuals requires querying the explainer for a large number of counterfactuals and is computationally expensive. The authors suggest a different strategy of privatizing diverse counterfactuals through discretization of continuous features, a technique which is closely related to generalization in privacy preserving data mining, and effective against linkage attacks.

K-anonymity's application on the complete training dataset for feature-based XAI was implemented by Bozorgpanah and Torra [21] using microaggregation. The study found alignment in the explanations generated from non-private and private datasets, with few irregularities, indicating that utility was preserved after privatization. Blanco-Justicia et al. [16]

also applied microaggregation to build local tree-based surrogate explanations from clusters around a point to be explained. K-anonymity was enforced by restricting the cluster size and the usage of shallow trees in the method enabled comprehensibility.

A line of work has focused on anonymization of images in the medical domain, where privacy protection of patients' visual data is crucial. The main objective of such techniques is data transformation through removal of identifying features while retaining explanatory evidence. Strategies such as the use of autoencoders for disentanglement of identifiable and medical characteristics [121] and the use of Siamese network for increasing identity distance between original and privatized images [122] for generating case-based explanations are proposed. Campos et al. [26] proposed the generation of synthetic images using latent diffusion models and removing those similar to training data, thus creating a collection of synthetic images to use as explanations. In addition to these techniques, the use of lossy compression by pixel sampling, is found to remove identification information while being post-hoc explainable and without accessing the original images [62, 85]. The compression also has added advantage of reducing the size of images to a fraction of the original size, thus making medical training datasets smaller [62].

For a critical domain, such as healthcare, anonymizing patients' images can assist in protecting identifiable data. However, the applied techniques should preserve image quality for utility to medical practitioners and patients [26]. Image anonymization techniques, unlike DP, have unproven guarantees while despite DP's guarantees, it is unable to scale beyond low resolution images [26]. Lossy compression provides a fast and effective way of privatizing images, with the benefits of achieving privacy and explainability while reducing training dataset sizes and enabling data sharing with multiple parties in non-private settings [62]. Thus, this method can be explored as a suitable alternative for generating explanations from sensitive image data.

Anonymization techniques protect privacy of individuals by preventing re-identification and linkage attacks [177]. When using k-anonymity, the choice of k that gives both accuracy and acceptable privacy risk level is critical [21]. Higher values of k yield higher privacy but explainability may be adversely affected [13, 16]. The real level of privacy may also not scale with increasing values of k [13]. Hence the choice of k that gives the right balance of privacy, explainability and utility is important. In addition, k-anonymity has downsides, such as its dependence on data characteristics, susceptibility to homogeneity attack [13], and its vulnerability to privacy leakage when background knowledge is available or diversity is lacking in the private attributes [64]. Other techniques such as l-diversity and t-closeness may overcome these challenges but are yet to be explored with explanations.

The generation of synthetic data for privacy preserving data analysis is explored in previous non-XAI works [20, 98]. 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 [174]. When models are trained on such data, the explanations through XAI tools are expected to be inherently privacy preserving, hence this can also be a promising future approach to safeguarding privacy in explainable systems.

5.4 Perturbation

Perturbation of sensitive data is a well-known technique in privacy preserving data publishing [171, 195]. When explanations contain sensitive information, obfuscating the contents can prevent direct exposure. Perturbation techniques can also be applied to stem indirect leakage from inferencing of sensitive attributes through explanations.

C.-L. Chen et al. [31] 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 control 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 or fidelity. Jeong et al. [81] proposed a defense framework against black-box model inversion attacks on saliency map explanations from image models. The framework comprised of two-player minimax game between inversion and noise injector networks, in which the inversion network attempts to reconstruct images from saliency maps and the noise injector perturbs explanations to counter the inversion. The use of multiple evaluation metrics to differentiate between the original and reconstructed images, was used to quantify the privacy of the defense mechanism.

For 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 [82], is found ineffective in mitigating membership inference through explanations [99]. Perturbation of explanations is also found insufficient in defending against data-free model extraction based on explanations [189]. However, the strategy has shown promising results in countering model inversion. The use of perturbation techniques at the explanation interface is attractive due to its ease of implementation, requiring no retraining of the model [81]. However, because the addition of large magnitude noise can degrade the usefulness of explanations [81], perturbations should be carefully calibrated to minimize any adverse impact on explanation quality.

5.5 Federated Learning

Among the privacy enhancing techniques available in PPML, Federated learning (FL) is an architectural solution [52] 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 [91]. It thus enables collaborative learning while keeping users' private data at the source [66] and mitigates the privacy risk of multiple parties sharing their sensitive data with other parties [53] or a centralized server [204]. In horizontal federated learning (HFL), local datasets have the same feature space but different samples while in vertical federated learning (VFL), the feature space is different but there are overlaps in samples [57].

To provide both privacy and explainability in Trustworthy AI, the combination of FL and XAI, i.e., Fed-XAI is suggested [11, 36, 100] and refers to the federated learning of XAI models. Many approaches of Fed-XAI using HFL and VFL are proposed in literature. Fiosina [57] used a HFL approach for forecasting taxi trip duration and applied feature-based explainability methods post-hoc. P. Chen et al. [32] 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 [39], Takagi-Sugeno [204] and Takagi-Sugeno-Kang [36] fuzzy rule-based models. In these setups, interpretability is achieved using underlying explainable-by-design [36] models.

Though FL aids privacy by default, it is prone to reconstruction and inferencing attacks [123, 199]. The sharing of gradients and model parameters, communication mechanism and aggregation process can lead to leakage of privacy of the participating clients [199]. Hence researchers have proposed integration of other privacy preserving techniques, such as cryptography, with FL methods. In one such work, Molhoek and Laanen [119] 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. [53] proposed a HFL structure using decision tree models, wherein a global decision tree was collaboratively trained by participants and aggregation of intermediate results used additive secret-sharing. A VFL technique, Falcon [187], utilized a hybrid approach consisting of partially homomorphic encryption (PHE) and additive secret sharing for exchange of intermediate computations. It used logistic/linear regression and multi-layer perceptron as local models. Another setup,

Pivot [186], proposed as part of Falcon, used classification and regression tree-based models. Threshold partially homomorphic encryption (TPHE) and additive secret sharing was used to protect privacy of intermediate exchanges. Though these works successfully integrate cryptographic techniques with FL, research has also found that the use of these methods in FL makes it difficult for the centralized server to differentiate the true model parameters, and this can lead to backdoor attacks by adversaries [69]. Hence appropriate defense frameworks, such as trust evaluation schemes [70], 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 [204] located at client systems. It reduces the footprint of user data in the network [123] by keeping data at the source and avoiding transmission and storage of sensitive information in a centralized location when multiple parties are involved [179]. Despite its benefits, in its current form FL faces challenges for its risk-free adoption [123] including ensuring privacy constraints, merging of local XAI models and dealing with large data streaming that can lead to concept drifts [11]. 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 efforts in this area is necessary to stem inadvertent privacy leakage through explanations.

5.6 Other techniques

In addition to the main privacy preservation methods from PPML, certain non-standard techniques are also explored in certain types of XAI. Some of these methods include limiting access to training data, obscuring features, or providing an abstraction of the target models.

A client-centric, data-driven approach of generating counterfactuals was proposed by An and Cao [6] 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 properties such as diversity and succinctness. Marton et al. [108] described a data-free approach of distilling the function represented by neural networks into interpretable models. The method used synthetic data to train a set of neural networks and extracted their 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. [143] 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 and Vitko [167] described an approach to build semantic 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 could be used in a decentralized setup for collaborative decision making, without individual parties sharing private local data.

The above works demonstrate the use of data-free and knowledge-driven techniques of XAI to build privacy- by-design systems. These approaches are applicable for protection of training data and the confidentiality of the model. The approach of disconnecting features from the model, enables the creation of an abstraction layer [143] for generation of explanations, thus 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)	[58]
	DP locally linear maps	Training data	Interpretable (locally linear maps)	[71]
	DPSaab	Training data	Interpretable (FF-CNN)	[94]
	DP-EBM	Training data	Interpretable (EBM)	[9, 131]
	DP explanation generation	Training and query data	Feature-based methods using local linear approximations (LIME, etc.)	[133]
	Local DP	Training data	Feature-based methods that aggregate scores (SHAP, etc.)	[130]
	DP trained SVM	Training data	Example-based (counterfactuals)	[117]
	DP-SGD	Training data	Feature-based (Grad-CAM)	[99]
	DP-RMSProp	Training data	Example-based (self-influence functions)	[35]
	Local DP	Training data	Feature-based (TreeSHAP)	[21]
	Local DP	Training data	Feature-based (SHAP)	[55]
	DP-WGAN (Wasserstein GAN)	Training data	Various XAI methods from DALEX framework[a]	[1]
Cryptography	Privacy preserving foil trees	Training data, model	Example-based (contrastive explanations)	[176]
	Additive homomorphic encryption	Query data, inference results, model	Interpretable (tree-based models)	[202]
Anonymization	Microaggregation (MDAV)	Training data, model	Interpretable (decision trees)	[16]
	Privacy preserving generative model	Training data	Example-based (case-based)	[122]
	HeartSpot (lossy compression)	Training data	Feature-based (saliency maps)	[85]
	Discretization of features (generalization)	Training data	Example-based (counterfactuals)	[177]
	CF-K (k-anonymity of counterfactuals)	Training data	Example-based (native counterfactuals)	[13, 64]
	DeepFixCX (lossy compression)	Training data	Feature-based (saliency maps)	[62]
	Microaggregation (MDAV)	Training data	Feature-based (TreeSHAP)	[1]
	Disentangled representation learning	Training data	Example-based (case-based)	[121]
	Latent diffusion models	Training data	Example-based (case-based)	[26]
Perturbation	GNIME	Training and query data	Feature-based (saliency maps)	[81]
	Linear-Time Optimal Privacy Scheme	Training and query data	Various XAI methods (interpretable surrogates, feature-based, etc.)	[31]

Privacy Preservation Category	Privacy Preserving Algorithm	Protected Asset	XAI Category (Method)	Study
Federated Learning	Pivot (VFL, additive secret sharing, TPHE)	Training data	Tree-based models (transparent)	[186]
	HFL	Training data	Feature-based methods (DeepLIFT, integrated gradients, LIME, etc.)	[57]
	HFL	Training data	Interpretable (Takagi-Sugeno,Takagi–Sugeno–Kang, fuzzy rule-based classifier)	[36, 39, 204]
	VFL	Training data	Counterfactuals	[32]
	Falcon (VFL, additive secret sharing, PHE)	Explanations and training data	Feature-based (LIME)	[187]
	PrivaTree (HFL, additive secret sharing)	Training data	Decision trees (transparent)	[53]
	VFL, SMC, Synthetic data	Query data	Example-based (counterfactuals)	[119]
Other techniques	Semantic XAI	Training data	Interpretable (decision trees, semantic rules)	[167]
	Semantic technologies (knowledge graphs, ontologies)	Model	Feature-based (LIME, etc.)	[143]
	Guarded counterfactuals	Training data, model	Example-based (counterfactuals)	[6]
	Interpretation-Nets	Training data	Interpretable (decision trees)	[108]

ª DALEX framework is available on https://github.com/modeloriented/dalex.

6 PRIVACY PRESERVING XAI CHARACTERISTICS

In the previous sections, we have seen the privacy risks in XAI due to intentional and unintentional causes. We have also reviewed applicable privacy preserving methods to protect the additional attack surface that explanations expose, in addition to the existing privacy related vulnerabilities in AI systems. In this section, based on the answers to RQ1 and RQ2, we attempt to answer RQ3 and identify desirable characteristics in XAI for mitigation of the identified risks. These characteristics shed light on the properties of privacy preserving XAI by considering the vulnerable assets that need protection and the users involved during the AI lifecycle. The characteristics intend to provide researchers and practitioners, a checklist for validation of existing privacy preserving XAI methods to aid in the design of new methods that target privacy by default. By reflecting these qualities, XAI can aspire to achieve the triad of privacy, explainability and utility.

We explain the characteristics (Figure 12) by considering three use cases outlined in Table 4. To facilitate understanding, a simplified example of an online loan application system that makes automated decisions using an AI model with XAI capabilities, is considered. The system uses 7 input features with salary, net worth, and age, being protected features that require privacy preservation. The use cases describe the following scenarios:

- Use Case 1 deals with intentional privacy leakage through an adversary.
- Use Case 2 involves interaction of a layman end-user, i.e., a bank's customer, who 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 of the applicant salary being < 40K and age > 50 years.

- Use Case 3 considers the interaction of technical support, in the roles of 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
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 tester validates the system updates with another round of testing.
Query data	Synthetic query similar to the error case requiring troubleshooting.

We propose 10 characteristics that XAI should demonstrate. The first 6 are derived from privacy attacks and unintentional leakage reviewed in Section 4. The remaining 4 are derived from performance issues and regulatory compliance. The proposed characteristics are as follows:

6.1.1 Prevent training data identification:

XAI tools should not aid 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 4.3.1).

6.1.2 Prevent sensitive data inference:

XAI tools should not aid reconstruction or inference of sensitive attributes of individuals. In Use Case 1, if the adversary knows 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 4.3.2).

6.1.3 Prevent reverse engineering of model:

XAI tools should not aid in reverse engineering 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 4.3.3).

6.1.4 Prevent property inference:

XAI tools should not aid 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 participants in the training data or the ratio of wealthy and average income training participants. Thus, the explanations should be resilient to property inference (Section 4.3.4).

6.1.5 Prevent direct exposure:

Explanations generated by XAI tools should not reveal personally identifiable and/or sensitive information to unauthorized individuals [31]. Certain explanation types, such as influence functions or native counterfactuals, reveal actual datapoints leading to direct exposure [13, 150, 151]. In Use Case 2, when the customer seeks an explanation on his/her application outcome, the explanation could indicate the failure to meet respective thresholds of $40K for salary and 50 years for age. Revealing the actual values of the protected features, would breach the privacy of the customer when it is 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 the consultation.

6.1.6 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 reveals the length of credit history, which is a non-sensitive attribute, to actors other than the customer, it can lead to breach of a sensitive attribute, i.e., age, due to the close correlation between the two attributes.

6.1.7 Access control of explanations:

The content of explanations should be accessed only by authorized users [16, 92]. The included details may vary from user to user on need-to-know basis. In Use Case 2, the customer is authorized to access his/her own explanation. 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 for synthetic queries generated to simulate specific error cases without accessing those for real production data.

6.1.8 Upholding of explanation quality:

The quality of explanations should not deteriorate by introduction of privacy preservation measures. Explanations are required to be useful [151] 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 an AI system.

6.1.9 Acceptable run time:

The run time of XAI methods, being an important evaluation metric [19], 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.

6.1.10 Compliance with applicable AI/privacy regulations:

XAI being an AI system, should comply with applicable AI and privacy regulations in respective jurisdictions. For instance, if the XAI is deployed in Canada with Canadian residents as the target users, it is expected to comply with the Artificial Intelligence and Data Act [2]. Users should be informed of the XAI capabilities of the system including the type and content of explanations and the parties with whom the explanations will be shared. Appropriate consent should be taken as required.
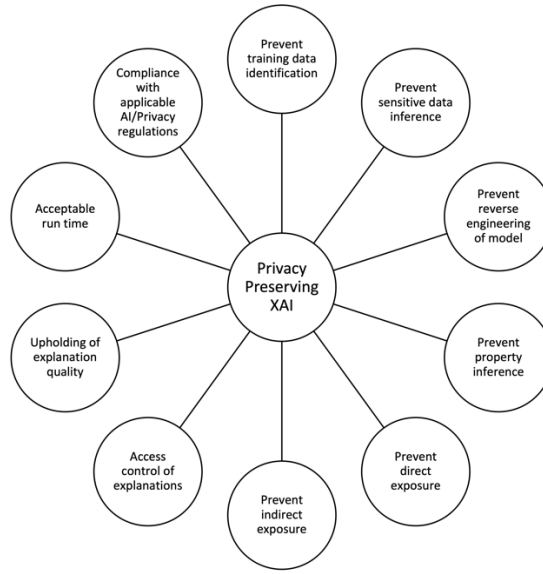

Figure 12: Proposed privacy preserving XAI characteristics.

7 DISCUSSION

In this section, we summarize the results of our work, its implications, the open issues, and challenges in the field. We also make recommendations for future research.

7.1 Summary, results, and implications

The review of existing literature enabled us to synthesize knowledge on the conflict between privacy and explainability, both being important pillars of Trustworthy AI. The extracted studies demonstrate that additional information provided in the form of explanations can be an advantage to adversaries to launch privacy attacks in XAI. We have categorized certain types of privacy leakage due to malicious intent of adversaries as intentional causes. These are in the form of membership inference, model inversion and model extraction, demonstrated on explanations generated using different methods. These

attacks pose a threat to the privacy of individuals contained in the training set, making them vulnerable to identification or exposure of their sensitive information. In addition, due to data reconstruction, individuals whose data is used to query the model [202, 203] during inferencing become susceptible to exposure of sensitive information, thus affecting the active users of an XAI system. The threat of model extraction through explanations, targets the confidentiality and intellectual property of the model owner. Property inference can enable determination of sensitive aggregates or group properties of the training data. However, no property inference attacks are found on XAI systems from the reviewed studies. In addition to privacy attacks, ML models have inherent privacy issues such as memorization of training data or overfitting leading to various inferencing attacks. These privacy problems are inherited in XAI systems, and we have categorized them as unintentional causes. In addition to this type of leakage due to training issues, 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.

Due to the imminent privacy risks of explainability, researchers have started exploring defense mechanisms for privacy preservation with explanations. This review determined that methods such as DP and anonymization are widely explored in this context, deducing from the number of studies that have used these methods. Fewer studies have used knowledge integration, cryptography, or perturbation as strategies for privacy preservation. Hence there is scope to utilize these underexplored techniques to achieve privacy in XAI systems. In distributed environment, Fed-XAI attempts to achieve explainability while preserving privacy of local data, but in its current form it cannot guarantee privacy with XAI, and explanations may provide a backdoor to malicious attacks. Besides, the integration of cryptographic techniques into Fed-XAI can further impact the security of the system [69].

The study of privacy risks and preservation methods in XAI led us to determine characteristics that privacy preserving explanations should reflect. In addition to being resilient to privacy attacks and prevention of direct and indirect exposure of sensitive information, explanations should satisfy performance and utility constraints. Since explanations can contain potentially identifiable data and may be subject to regulations, they are required to comply with the applicable laws within the jurisdiction. This article thus identifies and highlights a significant 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 as follows:

- Researchers can use the information to explore state-of-the-art privacy attacks, preservation methods and research related to privacy and explainability.
- 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 countermeasures surveyed, we have identified the following open issues and challenges. We also recommend potential directions for future work in this area.

7.2.1 Improving usability of XAI methods

End-users are an important component of XAI and directly consume the content generated by the system. Current XAI methods are essentially model-centric and useful for model development and audit [87]. However, the methods should steer towards a user-centric approach to provide need-to-know information to end-users depending on their role in the overall system. Explanations should be provided by integrating user-centric design principles in a privacy preserving manner, in a format suitable for users' consumption to cater to their diverse needs [4, 87]. Appropriate tools, such as interactivity and visualization, should be employed to enhance the process of explanation and deepen users' understanding

[18]. The use of the 3C-principle of context, content, and consent [24] for creating effective privacy explanations can be utilized for improving usability of XAI tools.

7.2.2 Development of privacy metrics for XAI

Privacy is a vital desideratum to ensure the safety of XAI but currently there is a lack of suitable metrics to measure the level of privacy of explanations. Current XAI literature provides evaluation metrics such as sufficiency [40], consistency [40] and sensitivity [193] to evaluate aspects such as faithfulness and robustness [72], but a metric for quantitative evaluation of privacy is lacking. To quantify the privacy of an explainability algorithm, a privacy metric is desirable.

7.2.3 Balancing tradeoff in privacy, explainability and accuracy

An adverse effect on accuracy is generally observed when privacy enhancement technologies are introduced [55, 71] in an AI system. For instance, perturbation of classifier weights of support vector machines has adverse impact on the classification accuracy and credibility of counterfactuals [117]. The use of differential privacy in neural network models is found to lower its accuracy and hence its utility [17]. Its introduction into the training process of models is also found to lower explanation quality [99]. Similarly, the use of generalization techniques for anonymization are found to reduce the quality of explanations [13].

Explainability integrated through XAI techniques in non-interpretable models can be evaluated with quantitative metrics such as faithfulness, robustness, stability and run time [19]. Determining the appropriate tradeoff between privacy, explainability and accuracy is crucial for the successful application of XAI. The use of compatibility matrix [1] or hyperparameters, for instance, 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 tradeoff. Metrics, such as trade-off score [1], can enable to quantify the balance of these properties.

7.2.4 Examining and improving tradeoff in different privacy preserving methods for XAI

Different privacy preservation methods applicable to XAI are discussed in Section 5. An examination of the privacy-utility-explainability tradeoff of these methods will help to identify those that are more effective compared to others. It will also help to identify their shortcomings so that novel or hybrid methods can be developed. Apart from differential privacy and anonymization, which have been mainly explored in XAI systems, other underexplored techniques such as use of knowledge integration and cryptographic protocols should be examined and compared. Distributed privacy enhancing solutions, such as Fed-XAI, should be further investigated to determine strategies to stem the possible leakages from XAI components.

7.2.5 Development of XAI methods that are privacy preserving by design

According to Hoepman [75], privacy is a core property of computer systems and needs addressing from system design phase rather than treated as an add-on. The proposed characteristics of privacy preserving XAI, can aid researchers and developers in building algorithms that are privacy enhanced by design [21]. Recently there is growing interest in neuro-symbolic approaches [74] and semantic technologies [147], and they can be explored for their usage as explainable-by-design models.

7.2.6 Privacy preserving XAI for Gen-AI and LLMs

XAI research has mainly focused on discriminative models that produce decision boundaries, and limited work has been done on developing explainability methods for Gen-AI and LLMs [146, 160, 183]. Due to the complex structure and vast number of parameters in these models, traditional explainability methods become impractical to them [201]. Currently these models have privacy issues, such as memorization of training data, that can enable the extraction of sensitive information which escalates as the models become larger [29]. In addition, the downstream private dataset used for in-context learning in LLMs is found to be susceptible to membership inference [48]. Methods such as retrieval-augmented generation (RAG) are currently being explored to counter some of the challenges, which enables the fine tuning of outputs by augmenting external data sources and can potentially mitigate leakage of memorized training data [197].

XAI is a vital desideratum for trustworthiness [180] and ethical applications of these models [102] but the introduction of explainability should not exacerbate the inherent privacy issues or create new vulnerabilities. A privacy analysis of novel applicable methods in the early stages of development and auditing using existing attacks [29], will boost the development of privacy-enhanced systems. Thus, with the current popularity of Gen-AI and LLMs for the masses, developing appropriate user-centric privacy preserving explainability techniques specifically for these systems is an important avenue for further research.

7.2.7 Evaluation of privacy preserving XAI characteristics

The proposed characteristics of privacy preserving XAI (Section 6) describe the desirable qualities of XAI that protect privacy while producing useful explanations to the target users. In future, we will work on the evaluation of XAI methods on the proposed characteristics. We will also enhance XAI methods so that the generated explanations satisfy the 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 [16, 53, 68]. Example-based methods that use instances as explanations, such as influence functions and native counterfactuals, cause direct exposure of training data [151, 176]. Among feature-based methods, backpropagation are found to be more susceptible to privacy risks compared to perturbation-based [151]. The privacy risks of different approaches should be comparatively studied to determine approaches that are more resilient to 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 tradeoff of privacy, utility and explainability in existing and new XAI methods, will determine its success as a foundation pillar of Trustworthy AI.

ACKNOWLEDGMENTS

REFERENCES

[1] Wisam Abbasi, Paolo Mori, and Andrea Saracino. 2024. 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*, July 30, 2024. ACM, Vienna Austria, 1–10. https://doi.org/10.1145/3664476.3670901

[2] AIDA. 2022. 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. *Bill C-27*. Retrieved from https://www.justice.gc.ca/eng/csj-sjc/pl/charter-charte/c27_1.html

[3] Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. 2020. Model extraction from counterfactual explanations. Retrieved February 24, 2023 from http://arxiv.org/abs/2009.01884

[4] 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. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion* 99, (November 2023), 101805. https://doi.org/10.1016/j.inffus.2023.101805

[5] 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. 2023. Towards Risk-Free Trustworthy Artificial Intelligence: Significance and Requirements. *Int. J. Intell. Syst.* 2023, (October 2023), 1–41. https://doi.org/10.1155/2023/4459198

[6] Shuai An and Yang Cao. 2024. Counterfactual Explanation at Will, with Zero Privacy Leakage. *Proc. ACM Manag. Data* 2, 3 (May 2024), 1–29. https://doi.org/10.1145/3654933

[7] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *6th International Conference on Learning Representations, ICLR 2018*, April 30, 2018. OpenReview.net, Vancouver, BC, Canada. https://doi.org/10.3929/ethz-b-000249929

[8] 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. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* 58, (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[9] Incheol Baek and Yon Dohn Chung. 2024. Differentially private and explainable boosting machine with enhanced utility. *Neurocomputing* 607, (November 2024), 128424. https://doi.org/10.1016/j.neucom.2024.128424

[10] Nagadivya Balasubramaniam, Marjo Kauppinen, Antti Rannisto, Kari Hiekkanen, and Sari Kujala. 2023. Transparency and explainability of AI systems: From ethical guidelines to requirements. *Inf. Softw. Technol.* 159, (July 2023), 107197. https://doi.org/10.1016/j.infsof.2023.107197

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

[12] J.M. Benitez, J.L. Castro, and I. Requena. 1997. Are artificial neural networks black boxes? *IEEE Trans. Neural Netw.* 8, 5 (September 1997), 1156–1164. https://doi.org/10.1109/72.623216

[13] Sjoerd Berning, Vincent Dunning, Dayana Spagnuelo, Thijs Veugen, and Jasper Van Der Waa. 2024. The Trade-off Between Privacy & Quality for Counterfactual Explanations. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*, July 30, 2024. ACM, Vienna Austria, 1–9. https://doi.org/10.1145/3664476.3670897

[14] Umang Bhatt, Adrian Weller, and José M. F. Moura. 2020. Evaluating and Aggregating Feature-based Model Explanations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, July 2020. International Joint Conferences on Artificial Intelligence Organization, Yokohama, Japan, 3016–3022. https://doi.org/10.24963/ijcai.2020/417

[15] Or Biran and Kathleen McKeown. 2017. Human-Centric Justification of Machine Learning Predictions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, August 2017. International Joint Conferences on Artificial Intelligence Organization, Melbourne, Australia, 1461–1467. https://doi.org/10.24963/ijcai.2017/202

[16] Alberto Blanco-Justicia, Josep Domingo-Ferrer, Sergio Martínez, and David Sánchez. 2020. Machine learning explainability via microaggregation and shallow decision trees. *Knowl.-Based Syst.* 194, (April 2020), 105532. https://doi.org/10.1016/j.knosys.2020.105532

[17] Alberto Blanco-Justicia, David Sánchez, Josep Domingo-Ferrer, and Krishnamurty Muralidhar. 2023. A Critical Review on the Use (and Misuse) of Differential Privacy in Machine Learning. *ACM Comput. Surv.* 55, 8 (August 2023), 1–16. https://doi.org/10.1145/3547139

[18] Jessica Y Bo, Pan Hao, and Brian Y Lim. 2024. Incremental XAI: Memorable Understanding of AI with Incremental Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, May 11, 2024. ACM, Honolulu HI USA, 1–17. https://doi.org/10.1145/3613904.3642689

[19] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2023. Benchmarking and survey of explanation methods for black box models. *Data Min. Knowl. Discov.* 37, 5 (September 2023), 1719–1778. https://doi.org/10.1007/s10618-023-00933-9

[20] March Boedihardjo, Thomas Strohmer, and Roman Vershynin. 2023. Privacy of Synthetic Data: A Statistical Framework. *IEEE Trans. Inf. Theory* 69, 1 (January 2023), 520–527. https://doi.org/10.1109/TIT.2022.3216793

[21] Aso Bozorgpanah and Vicenç Torra. 2024. Explainable machine learning models with privacy. *Prog. Artif. Intell.* 13, 1 (March 2024), 31–50. https://doi.org/10.1007/s13748-024-00315-2

[22] Leo Breiman. 2001. Random forests. *Mach. Learn.* 45, (2001), 5–32.

[23] Wasja Brunotte, Larissa Chazette, and Kai Korte. 2021. Can Explanations Support Privacy Awareness? A Research Roadmap. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, September 2021. IEEE, Notre Dame, IN, USA, 176–180. https://doi.org/10.1109/REW53955.2021.00032

[24] Wasja Brunotte, Jakob Droste, and Kurt Schneider. 2023. Context, Content, Consent - How to Design User-Centered Privacy Explanations (S). July 01, 2023. 86–89. https://doi.org/10.18293/SEKE2023-032

[25] Erik Cambria, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. 2023. A survey on XAI and natural language explanations. *Inf. Process. Manag.* 60, 1 (January 2023), 103111. https://doi.org/10.1016/j.ipm.2022.103111

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

[27] Nicola Capuano, Giuseppe Fenza, Vincenzo Loia, and Claudio Stanzione. 2022. Explainable Artificial Intelligence in CyberSecurity: A Survey. *IEEE Access* 10, (2022), 93575–93600. https://doi.org/10.1109/ACCESS.2022.3204171

[28] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. 2022. Membership Inference Attacks From First Principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, May 2022. IEEE, San Francisco, CA, USA, 1897–1914. https://doi.org/10.1109/SP46214.2022.9833649

[29] 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. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, August 2021. USENIX Association, 2633–2650. Retrieved from https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

[30] Larissa Chazette, Wasja Brunotte, and Timo Speith. 2021. Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*, September 2021. IEEE, Notre Dame, IN, USA, 197–208. https://doi.org/10.1109/RE51729.2021.00025

[31] Chien-Lun Chen, Leana Golubchik, and Ranjan Pal. 2022. Achieving Transparency Report Privacy in Linear Time. *J. Data Inf. Qual.* 14, 2 (June 2022), 1–56. https://doi.org/10.1145/3460001

[32] Peng Chen, Xin Du, Zhihui Lu, Jie Wu, and Patrick C.K. Hung. 2022. EVFL: An explainable vertical federated learning for data-oriented Artificial Intelligence systems. *J. Syst. Archit.* 126, (May 2022), 102474. https://doi.org/10.1016/j.sysarc.2022.102474

[33] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. 2021. Label-Only Membership Inference Attacks. In *Proceedings of the 38th International Conference on Machine Learning* (*Proceedings of Machine Learning Research*), July 2021. PMLR, 1964-1974. Retrieved from https://proceedings.mlr.press/v139/choquette-choo21a.html

[34] Roger Clarke. 1999. Internet privacy concerns confirm the case for intervention. *Commun. ACM* 42, 2 (February 1999), 60–67. https://doi.org/10.1145/293411.293475

[35] Gilad Cohen and Raja Giryes. 2024. Membership Inference Attack Using Self Influence Functions. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 03, 2024. IEEE, Waikoloa, HI, USA, 4880–4889. https://doi.org/10.1109/WACV57701.2024.00482

[36] José Luis Corcuera Bárcena, Pietro Ducange, Francesco Marcelloni, Giovanni Nardini, Alessandro Noferi, Alessandro Renda, Fabrizio Ruffini, Alessio Schiavo, Giovanni Stea, and Antonio Virdis. 2023. Enabling federated learning of explainable AI models within beyond-5G/6G networks. *Comput. Commun.* 210, (October 2023), 356–375. https://doi.org/10.1016/j.comcom.2023.07.039

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

[38] James Curzon, Tracy Ann Kosa, Rajen Akalu, and Khalil El-Khatib. 2021. Privacy and Artificial Intelligence. *IEEE Trans. Artif. Intell.* 2, 2 (April 2021), 96–108. https://doi.org/10.1109/TAI.2021.3088084

[39] Mattia Daole, Pietro Ducange, Francesco Marcelloni, and Alessandro Renda. 2024. Trustworthy AI in Heterogeneous Settings: Federated Learning of Explainable Classifiers. In *2024 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, June 30, 2024. IEEE, Yokohama, Japan, 1–9. https://doi.org/10.1109/FUZZ-IEEE60900.2024.10612109

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

[41] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, May 2016. IEEE, San Jose, CA, 598–617. https://doi.org/10.1109/SP.2016.42

[42] Ashley Deeks. 2019. The Judicial Demand For Explainable Artificial Intelligence. *Columbia Law Rev.* 119, 7 (2019), 1829–1850.

[43] Tribikram Dhar, Nilanjan Dey, Surekha Borra, and R. Simon Sherratt. 2023. Challenges of Deep Learning in Medical Image Analysis—Improving Explainability and Trust. *IEEE Trans. Technol. Soc.* 4, 1 (March 2023), 68–75. https://doi.org/10.1109/TTS.2023.3234203

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

[45] Antreas Dionysiou, Vassilis Vassiliades, and Elias Athanasopoulos. 2023. Exploring Model Inversion Attacks in the Black-box Setting. *Proc. Priv. Enhancing Technol.* 2023, 1 (January 2023), 190–206. https://doi.org/10.56553/popets-2023-0012

[46] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. Retrieved February 28, 2023 from http://arxiv.org/abs/1702.08608

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

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

[49] Vasisht Duddu and Antoine Boutet. 2022. Inferring Sensitive Attributes from Model Explanations. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*, October 17, 2022. Association for Computing Machinery, New York, NY, USA, 416–425. https://doi.org/10.1145/3511808.3557362

[50] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. 2023. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.* 55, 9 (September 2023), 1–33. https://doi.org/10.1145/3561048

[51] Cynthia Dwork. 2006. Differential Privacy. In *Automata, Languages and Programming*, 2006. Springer Berlin Heidelberg, Berlin, Heidelberg, 1–12.

[52] Soumia Zohra El Mestari, Gabriele Lenzini, and Huseyin Demirci. 2024. Preserving data privacy in machine learning systems. *Comput. Secur.* 137, (February 2024), 103605. https://doi.org/10.1016/j.cose.2023.103605

[53] Yamane El Zein, Mathieu Lemay, and Kévin Huguenin. 2024. PrivaTree: Collaborative Privacy-Preserving Training of Decision Trees on Biomedical Data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 21, 1 (January 2024), 1–13. https://doi.org/10.1109/TCBB.2023.3286274

[54] Elsevier B.V. Engineering Village. Retrieved May 2, 2024 from https://www.engineeringvillage.com/

[55] Fatima Ezzeddine, Mirna Saad, Omran Ayoub, Davide Andreoletti, Martin Gjoreski, Ihab Sbeity, Marc Langheinrich, and Silvia Giordano. 2024. Differential Privacy for Anomaly Detection: Analyzing the Trade-Off Between Privacy and Explainability. In *Explainable Artificial Intelligence*, Luca Longo, Sebastian Lapuschkin and Christin Seifert (eds.). Springer Nature Switzerland, Cham, 294–318. https://doi.org/10.1007/978-3-031-63800-8_15

[56] Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. 2024. Probabilistic Dataset Reconstruction from Interpretable Models. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, April 09, 2024. IEEE, Toronto, ON, Canada, 1–17. https://doi.org/10.1109/SaTML59370.2024.00009

[57] Jelena Fiosina. 2022. Interpretable Privacy-Preserving Collaborative Deep Learning for Taxi Trip Duration Forecasting. In *Smart Cities, Green Technologies, and Intelligent Transport Systems*, Cornel Klein, Matthias Jarke, Markus Helfert, Karsten Berns and Oleg Gusikhin (eds.). Springer International Publishing, Cham, 392–411. https://doi.org/10.1007/978-3-031-17098-0_20

[58] Sam Fletcher and Md. Zahidul Islam. 2020. Decision Tree Classification with Differential Privacy: A Survey. *ACM Comput. Surv.* 52, 4 (July 2020), 1–33. https://doi.org/10.1145/3337064

[59] Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, October 2017. IEEE, Venice, 3449–3457. https://doi.org/10.1109/ICCV.2017.371

[60] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, October 12, 2015. ACM, Denver Colorado USA, 1322–1333. https://doi.org/10.1145/2810103.2813677

[61] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. 2018. 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*, October 15, 2018. ACM, Toronto Canada, 619–633. https://doi.org/10.1145/3243734.3243834

[62] Alex Gaudio, Asim Smailagic, Christos Faloutsos, Shreshta Mohan, Elvin Johnson, Yuhao Liu, Pedro Costa, and Aurélio Campilho. 2023. DeepFixCX: Explainable privacy-preserving image compression for medical image analysis. *WIREs Data Min. Knowl. Discov.* (March 2023). https://doi.org/10.1002/widm.1495

[63] GDPR. 2016. Art. 22 GDPR. *Art. 22 GDPR*. Retrieved from https://gdpr-info.eu/art-22-gdpr/

[64] Sofie Goethals, Kenneth Sörensen, and David Martens. 2023. The Privacy Issue of Counterfactual Explanations: Explanation Linkage Attacks. *ACM Trans. Intell. Syst. Technol.* 14, 5 (October 2023), 1–24. https://doi.org/10.1145/3608482

[65] Abigail Goldsteen, Gilad Ezov, and Ariel Farkash. 2020. Reducing Risk of Model Inversion Using Privacy-Guided Training. Retrieved April 5, 2023 from http://arxiv.org/abs/2006.15877

[66] Alejandro Guerra-Manzanares, L. Julian Lechuga Lopez, Michail Maniatakos, and Farah E. Shamout. 2023. Privacy-Preserving Machine Learning for Healthcare: Open Challenges and Future Perspectives. In *Trustworthy Machine Learning for Healthcare*, Hao Chen and Luyang Luo (eds.). Springer Nature Switzerland, Cham, 25–40. https://doi.org/10.1007/978-3-031-39539-0_3

[67] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Min. Knowl. Discov.* (April 2022). https://doi.org/10.1007/s10618-022-00831-6

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

[69] Jingjing Guo, Haiyang Li, Feiran Huang, Zhiquan Liu, Yanguo Peng, Xinghua Li, Jianfeng Ma, Varun G. Menon, and Konstantin Kostromitin Igorevich. 2022. ADFL: A Poisoning Attack Defense Framework for Horizontal Federated Learning. *IEEE Trans. Ind. Inform.* 18, 10 (October 2022), 6526–6536. https://doi.org/10.1109/TII.2022.3156645

[70] Jingjing Guo, Zhiquan Liu, Siyi Tian, Feiran Huang, Jiaxing Li, Xinghua Li, Kostromitin Konstantin Igorevich, and Jianfeng Ma. 2023. TFL-DT: A Trust Evaluation Scheme for Federated Learning in Digital Twin for Mobile Networks. *IEEE J. Sel. Areas Commun.* 41, 11 (November 2023), 3548–3560. https://doi.org/10.1109/JSAC.2023.3310094

[71] Frederik Harder, Matthias Bauer, and Mijung Park. 2020. Interpretable and Differentially Private Predictions. *Proc. AAAI Conf. Artif. Intell.* 34, 04 (April 2020), 4083–4090. https://doi.org/10.1609/aaai.v34i04.5827

[72] Anna Hedström, Leander Weber, Dilyara Bareeva, Daniel Krakowczyk, Franz Motzkus, Wojciech Samek, and Sebastian Lapuschkin. 2023. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *J. Mach. Learn. Res.* 24, 34 (2023), 1–11.

[73] Pascal Hitzler, Federico Bianchi, Monireh Ebrahimi, and Md Kamruzzaman Sarker. 2020. Neural-symbolic integration and the Semantic Web. *Semantic Web* 11, 1 (January 2020), 3–11. https://doi.org/10.3233/SW-190368

[74] Pascal Hitzler, Aaron Eberhart, Monireh Ebrahimi, Md Kamruzzaman Sarker, and Lu Zhou. 2022. Neuro-symbolic approaches in artificial intelligence. *Natl. Sci. Rev.* 9, 6 (June 2022), nwac035. https://doi.org/10.1093/nsr/nwac035

[75] Jaap-Henk Hoepman. 2014. Privacy Design Strategies. In *ICT Systems Security and Privacy Protection*, Nora Cuppens-Boulahia, Frédéric Cuppens, Sushil Jajodia, Anas Abou El Kalam and Thierry Sans (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 446–459. https://doi.org/10.1007/978-3-642-55415-5_38

[76] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. 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*, May 02, 2019. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300809

[77] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership Inference Attacks on Machine Learning: A Survey. *ACM Comput. Surv.* 54, 11s (January 2022), 1–37. https://doi.org/10.1145/3523273

[78] ICO. 2020. Explaining decisions made with AI. Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-ai/

[79] Eleni Ilkou and Maria Koutraki. 2020. Symbolic Vs Sub-symbolic AI Methods: Friends or Enemies? In *Proceedings of the CIKM 2020 Workshops*, October 19, 2020. .

[80] Marija Jegorova, Chaitanya Kaul, Charlie Mayor, Alison Q. O'Neil, Alexander Weir, Roderick Murray-Smith, and Sotirios A. Tsaftaris. 2022. Survey: Leakage and Privacy at Inference Time. *IEEE Trans. Pattern Anal. Mach. Intell.* (2022), 1–20. https://doi.org/10.1109/TPAMI.2022.3229593

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

[82] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. 2019. MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, November 06, 2019. ACM, London United Kingdom, 259–274. https://doi.org/10.1145/3319535.3363201

[83] Yan Jia, Chaitanya Kaul, Tom Lawton, Roderick Murray-Smith, and Ibrahim Habli. 2021. Prediction of weaning from mechanical ventilation using Convolutional Neural Networks. *Artif. Intell. Med.* 117, (July 2021), 102087. https://doi.org/10.1016/j.artmed.2021.102087

[84] José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. 2020. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* 2, 10 (October 2020), 573–584. https://doi.org/10.1038/s42256-020-00236-4

[85] Elvin Johnson, Shreshta Mohan, Alex Gaudio, Asim Smailagic, Christos Faloutsos, and Aurelio Campilho. 2022. HeartSpot: Privatized and Explainable Data Compression for Cardiomegaly Detection. In *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, September 27, 2022. IEEE, Ioannina, Greece, 01–04. https://doi.org/10.1109/BHI56158.2022.9926777

[86] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viegas, and Michael Terry. 2019. XRAI: Better Attributions Through Regions. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. IEEE, Seoul, Korea (South), 4947–4956. https://doi.org/10.1109/ICCV.2019.00505

[87] Sinan Kaplan, Hannu Uusitalo, and Lasse Lensu. 2024. A unified and practical user-centric framework for explainable artificial intelligence. *Knowl.-Based Syst.* 283, (January 2024), 111107. https://doi.org/10.1016/j.knosys.2023.111107

[88] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2023. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Comput. Surv.* 55, 5 (May 2023), 1–29. https://doi.org/10.1145/3527848

[89] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems* (*NIPS 2016*), 2016. .

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

[91] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated Learning: Strategies for Improving Communication Efficiency. 2016. Barcelona, Spain. Retrieved April 28, 2023 from http://arxiv.org/abs/1610.05492

[92] Aditya Kuppa and Nhien-An Le-Khac. 2021. Adversarial XAI Methods in Cybersecurity. *IEEE Trans. Inf. Forensics Secur.* 16, (2021), 4924–4938. https://doi.org/10.1109/TIFS.2021.3117075

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

[94] De Li, Jinyan Wang, Zhou Tan, Xianxian Li, and Yuhang Hu. 2020. 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)*, December 2020. IEEE, Guangzhou, China, 631–638. https://doi.org/10.1109/TrustCom50675.2020.00089

[95] 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. 2020. A Survey of Data-driven and Knowledge-aware eXplainable AI. *IEEE Trans. Knowl. Data Eng.* (2020), 1–1. https://doi.org/10.1109/TKDE.2020.2983930

[96] LiMin Fu. 1994. Rule generation from neural networks. *IEEE Trans. Syst. Man Cybern.* 24, 8 (August 1994), 1114–1124. https://doi.org/10.1109/21.299696

[97] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2022. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.* 54, 2 (March 2022), 1–36. https://doi.org/10.1145/3436755

[98] Fan Liu, Zhiyong Cheng, Huilin Chen, Yinwei Wei, Liqiang Nie, and Mohan Kankanhalli. 2022. Privacy-Preserving Synthetic Data Generation for Recommendation Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 06, 2022. ACM, Madrid Spain, 1379–1389. https://doi.org/10.1145/3477495.3532044

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

[100] Raúl López-Blanco, Ricardo S. Alonso, Angélica González-Arrieta, Pablo Chamoso, and Javier Prieto. 2023. Federated Learning of Explainable Artificial Intelligence (FED-XAI): A Review. In *Distributed Computing and Artificial Intelligence, 20th International Conference*, Sascha Ossowski, Pawel Sitek, Cesar Analide, Goreti Marreiros, Pablo Chamoso and Sara Rodríguez (eds.). Springer Nature Switzerland, Cham, 318–326. https://doi.org/10.1007/978-3-031-38333-5_32

[101] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (*NIPS'17*), December 04, 2017. Curran Associates Inc., Long Beach, California, USA, 4768–4777. https://doi.org/10.5555/3295222.3295230

[102] Haoyan Luo and Lucia Specia. 2024. From Understanding to Utilization: A Survey on Explainability for Large Language Models. Retrieved April 22, 2024 from http://arxiv.org/abs/2401.12874

[103] Xinjian Luo, Yangfan Jiang, and Xiaokui Xiao. 2022. Feature Inference Attack on Shapley Values. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, November 07, 2022. ACM, Los Angeles CA USA, 2233–2247. https://doi.org/10.1145/3548606.3560573

[104] Yao Ma, Xurong Zhai, Dan Yu, Yuli Yang, Xingyu Wei, and Yongle Chen. 2024. Label-Only Membership Inference Attack Based on Model Explanation. *Neural Process. Lett.* 56, 5 (September 2024), 236. https://doi.org/10.1007/s11063-024-11682-1

[105] R. Machlev, L. Heistrene, M. Perl, K.Y. Levy, J. Belikov, S. Mannor, and Y. Levron. 2022. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy AI* 9, (August 2022), 100169. https://doi.org/10.1016/j.egyai.2022.100169

[106] Saeed Mahloujifar, Esha Ghosh, and Melissa Chase. 2022. Property Inference from Poisoning. In *2022 IEEE Symposium on Security and Privacy (SP)*, May 2022. IEEE, San Francisco, CA, USA, 1120–1137. https://doi.org/10.1109/SP46214.2022.9833623

[107] Abdul Majeed and Sungchang Lee. 2021. Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. *IEEE Access* 9, (2021), 8512–8545. https://doi.org/10.1109/ACCESS.2020.3045700

[108] Sascha Marton, Stefan Lüdtke, Christian Bartelt, Andrej Tschalzev, and Heiner Stuckenschmidt. 2024. Explaining neural networks without access to training data. *Mach. Learn.* (January 2024). https://doi.org/10.1007/s10994-023-06428-4

[109] John A. McDermid, Yan Jia, Zoe Porter, and Ibrahim Habli. 2021. Artificial intelligence explainability: the technical and ethical dimensions. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* 379, 2207 (October 2021), 20200363. https://doi.org/10.1098/rsta.2020.0363

[110] Claire McKay Bowen and Simson Garfinkel. 2021. The Philosophy of Differential Privacy. *Not. Am. Math. Soc.* 68, 10 (November 2021), 1. https://doi.org/10.1090/noti2363

[111] Mohammad I. Merhi. 2022. An Assessment of the Barriers Impacting Responsible Artificial Intelligence. *Inf. Syst. Front.* (April 2022). https://doi.org/10.1007/s10796-022-10276-3

[112] Bertalan Meskó and Eric J. Topol. 2023. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *Npj Digit. Med.* 6, 1 (July 2023), 120. https://doi.org/10.1038/s41746-023-00873-0

[113] C.R. Milare, A.C.P.L.F. De Carvalho, and M.C. Monard. 2001. Extracting rules from neural networks using symbolic algorithms: preliminary results. In *Proceedings Fourth International Conference on Computational Intelligence and Multimedia Applications. ICCIMA 2001*, 2001. IEEE, Yokusika City, Japan, 384–388. https://doi.org/10.1109/ICCIMA.2001.970500

[114] Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. 2019. Model Reconstruction from Model Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, January 29, 2019. ACM, Atlanta GA USA, 1–9. https://doi.org/10.1145/3287560.3287562

[115] Aditi Mishra, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, and Chris Bryan. 2023. PromptAid: Prompt Exploration, Perturbation, Testing and Iteration using Visual Analytics for Large Language Models. Retrieved April 18, 2024 from http://arxiv.org/abs/2304.01964

[116] Takayuki Miura, Toshiki Shibahara, and Naoto Yanai. 2024. 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*, July 02, 2024. ACM, Singapore Singapore, 56–66. https://doi.org/10.1145/3665451.3665533

[117] Rami Mochaourab, Sugandh Sinha, Stanley Greenstein, and Panagiotis Papapetrou. 2023. Demonstrator on Counterfactual Explanations for Differentially Private Support Vector Machines. In *Machine Learning and Knowledge Discovery in Databases*, Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak and Grigorios Tsoumakas (eds.). Springer Nature Switzerland, Cham, 662–666. https://doi.org/10.1007/978-3-031-26422-1_52

[118] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* 11, 3–4 (December 2021), 1–45. https://doi.org/10.1145/3387166

[119] M. Molhoek and J. Van Laanen. 2024. Secure Counterfactual Explanations in a Two-party Setting. In *2024 27th International Conference on Information Fusion (FUSION)*, July 08, 2024. IEEE, Venice, Italy, 1–10. https://doi.org/10.23919/FUSION59988.2024.10706413

[120] Christoph Molnar. 2023. *Interpretable Machine Learning*.

[121] Helena Montenegro and Jaime S. Cardoso. 2024. Anonymizing medical case-based explanations through disentanglement. *Med. Image Anal.* 95, (July 2024), 103209. https://doi.org/10.1016/j.media.2024.103209

[122] Helena Montenegro, Wilson Silva, and Jaime S. Cardoso. 2021. Privacy-Preserving Generative Adversarial Network for Case-Based Explainability in Medical Image Analysis. *IEEE Access* 9, (2021), 148037–148047. https://doi.org/10.1109/ACCESS.2021.3124844

[123] Viraaji Mothukuri, Reza M. Parizi, Seyedamin Pouriyeh, Yan Huang, Ali Dehghantanha, and Gautam Srivastava. 2021. A survey on security and privacy of federated learning. *Future Gener. Comput. Syst.* 115, (February 2021), 619–640. https://doi.org/10.1016/j.future.2020.10.007

[124] Z. Müftüoğlu, M. A. Kızrak, and T. Yıldırım. 2022. Privacy-Preserving Mechanisms with Explainability in Assistive AI Technologies. In *Advances in Assistive Technologies*, George A. Tsihrintzis, Maria Virvou, Anna Esposito and Lakhmi C. Jain (eds.). Springer International Publishing, Cham, 287–309. https://doi.org/10.1007/978-3-030-87132-1_13

[125] Deepa Muralidhar, Rafik Belloum, Kathia Marçal De Oliveira, and Ashwin Ashok. 2023. Elements that Influence Transparency in Artificial Intelligent Systems - A Survey. In *Human-Computer Interaction – INTERACT 2023*, José Abdelnour Nocera, Marta Kristín Lárusdóttir, Helen Petrie, Antonio Piccinno and Marco Winckler (eds.). Springer Nature Switzerland, Cham, 349–358. https://doi.org/10.1007/978-3-031-42280-5_21

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

[127] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, May 2019. IEEE, San Francisco, CA, USA, 739–753. https://doi.org/10.1109/SP.2019.00065

[128] Mohamed Nassar, Khaled Salah, Muhammad Habib ur Rehman, and Davor Svetinovic. 2020. Blockchain for explainable and trustworthy artificial intelligence. *WIREs Data Min. Knowl. Discov.* 10, 1 (January 2020). https://doi.org/10.1002/widm.1340

[129] Madelena Y. Ng, Supriya Kapur, Katherine D. Blizinsky, and Tina Hernandez-Boussard. 2022. The AI life cycle: a holistic approach to creating ethical AI for health decisions. *Nat. Med.* 28, 11 (November 2022), 2247–2249. https://doi.org/10.1038/s41591-022-01993-y

[130] Truc Nguyen, Phung Lai, Hai Phan, and My T. Thai. 2023. XRand: Differentially Private Defense against Explanation-Guided Attacks. *Proc. AAAI Conf. Artif. Intell.* 37, 10 (June 2023), 11873–11881. https://doi.org/10.1609/aaai.v37i10.26401

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

[132] Andrés Páez. 2019. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds Mach.* 29, 3 (September 2019), 441–459. https://doi.org/10.1007/s11023-019-09502-w

[133] Neel Patel, Reza Shokri, and Yair Zick. 2022. Model Explanations with Differential Privacy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, June 21, 2022. ACM, Seoul Republic of Korea, 1895–1904. https://doi.org/10.1145/3531146.3533235

[134] Martin Pawelczyk, Himabindu Lakkaraju, and Seth Neel. 2023. On the Privacy Risks of Algorithmic Recourse. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* (*Proceedings of Machine Learning Research*), 2023. PMLR, Valencia, Spain, 9680–9696. Retrieved from https://proceedings.mlr.press/v206/pawelczyk23a.html

[135] Seyedeh Neelufar Payrovnaziri, Zhaoyi Chen, Pablo Rengifo-Moreno, Tim Miller, Jiang Bian, Jonathan H Chen, Xiuwen Liu, and Zhe He. 2020. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J. Am. Med. Inform. Assoc.* 27, 7 (July 2020), 1173–1185. https://doi.org/10.1093/jamia/ocaa053

[136] Nikolaos Pitropakis, Emmanouil Panaousis, Thanassis Giannetsos, Eleftherios Anastasiadis, and George Loukas. 2019. A taxonomy and survey of attacks against machine learning. *Comput. Sci. Rev.* 34, (November 2019), 100199. https://doi.org/10.1016/j.cosrev.2019.100199

[137] Enayat Rajabi and Kobra Etminani. 2022. Knowledge-graph-based explainable AI: A systematic review. *J. Inf. Sci.* (September 2022), 016555152211128. https://doi.org/10.1177/01655515221112844

[138] Atul Rawal, James McCoy, Danda B. Rawat, Brian M. Sadler, and Robert St. Amant. 2022. Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives. *IEEE Trans. Artif. Intell.* 3, 6 (December 2022), 852–866. https://doi.org/10.1109/TAI.2021.3133846

[139] M. Sadegh Riazi and Farinaz Koushanfar. 2018. Privacy-preserving deep learning and inference. In *Proceedings of the International Conference on Computer-Aided Design*, November 05, 2018. ACM, San Diego California, 1–4. https://doi.org/10.1145/3240765.3274560

[140] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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*, August 13, 2016. ACM, San Francisco California USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[141] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. *Proc. AAAI Conf. Artif. Intell.* 32, 1 (April 2018). https://doi.org/10.1609/aaai.v32i1.11491

[142] Maria Rigaki and Sebastian Garcia. 2023. A Survey of Privacy Attacks in Machine Learning. *ACM Comput. Surv.* 56, 4 (November 2023), 1–34. https://doi.org/10.1145/3624010

[143] Jože M. Rožanec, Blaž Fortuna, and Dunja Mladenić. 2022. Knowledge graph-based rich and confidentiality preserving Explainable Artificial Intelligence (XAI). *Inf. Fusion* 81, (May 2022), 91–102. https://doi.org/10.1016/j.inffus.2021.11.015

[144] Conrad Sanderson, David Douglas, and Qinghua Lu. 2023. Implementing Responsible AI: Tensions and Trade-Offs Between Ethics Aspects. In *2023 International Joint Conference on Neural Networks (IJCNN)*, June 18, 2023. IEEE, Gold Coast, Australia, 1–7. https://doi.org/10.1109/IJCNN54540.2023.10191274

[145] Ana Šarčević, Damir Pintar, Mihaela Vranić, and Agneza Krajna. 2022. Cybersecurity Knowledge Extraction Using XAI. *Appl. Sci.* 12, 17 (August 2022), 8669. https://doi.org/10.3390/app12178669

[146] Johannes Schneider. 2024. Explainable Generative AI (GenXAI): A Survey, Conceptualization, and Research Agenda. Retrieved April 18, 2024 from http://arxiv.org/abs/2404.09554

[147] Arne Seeliger, Matthias Pfaff, and Helmut Krcmar. 2019. Semantic Web Technologies for Explainable Machine Learning Models: A Literature Review. *PROFILESSEMEX ISWC* 2465, (2019), 1–16.

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

[149] Sakib Shahriar, Sonal Allana, Mehdi Hazrati Fard, and Rozita Dara. 2023. A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle. *IEEE Access* (2023), 1–1. https://doi.org/10.1109/ACCESS.2023.3287195

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

[151] Reza Shokri, Martin Strobel, and Yair Zick. 2021. On the Privacy Risks of Model Explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, July 21, 2021. ACM, Virtual Event USA, 231–241. https://doi.org/10.1145/3461702.3462533

[152] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, May 2017. IEEE, San Jose, CA, USA, 3–18. https://doi.org/10.1109/SP.2017.41

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

[154] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2017. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. Retrieved March 28, 2023 from http://arxiv.org/abs/1605.01713

[155] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Workshop at International Conference on Learning Representations*, April 19, 2014. .

[156] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. 2017. Machine Learning Models that Remember Too Much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, October 30, 2017. ACM, Dallas Texas USA, 587–601. https://doi.org/10.1145/3133956.3134077

[157] Christoforos N. Spartalis, Theodoros Semertzidis, and Petros Daras. 2024. Balancing XAI with Privacy and Security Considerations. In *Computer Security. ESORICS 2023 International Workshops*, 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.). Springer Nature Switzerland, Cham, 111–124. https://doi.org/10.1007/978-3-031-54129-2_7

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

[159] Martin Strobel and Reza Shokri. 2022. Data Privacy and Trustworthy Machine Learning. *IEEE Secur. Priv.* 20, 5 (September 2022), 44–49. https://doi.org/10.1109/MSEC.2022.3178187

[160] Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D. Weisz. 2022. Investigating Explainability of Generative AI for Code through Scenario-based Design. In *27th International Conference on Intelligent User Interfaces*, March 22, 2022. ACM, Helsinki Finland, 212–228. https://doi.org/10.1145/3490099.3511119

[161] Mukund Sundararajan and Amir Najmi. 2020. The Many Shapley Values for Model Explanation. In *Proceedings of the 37th International Conference on Machine Learning*, September 21, 2020. PMLR, Virtual, 9269–9278.

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

[163] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 05 (2002), 557–570.

[164] Latanya Sweeney. 2002. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 05 (2002), 571–588.

[165] Elham Tabassi. 2023. *AI Risk Management Framework: AI RMF (1.0)*. National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.AI.100-1

[166] Xinyu Tang, Richard Shin, Huseyin A. Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2024. Privacy-Preserving In-Context Learning with Differentially Private Few-Shot Generation. Retrieved April 22, 2024 from http://arxiv.org/abs/2309.11765

[167] Vagan Terziyan and Oleksandra Vitko. 2022. Explainable AI for Industry 4.0: Semantic Representation of Deep Learning Models. *Procedia Comput. Sci.* 200, (2022), 216–226. https://doi.org/10.1016/j.procs.2022.01.220

[168] Ilaria Tiddi and Stefan Schlobach. 2022. Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.* 302, (January 2022), 103627. https://doi.org/10.1016/j.artint.2021.103627

[169] Ryotaro Toma and Hiroaki Kikuchi. 2024. Combinations of AI Models and XAI Metrics Vulnerable to Record Reconstruction Risk. In *Privacy in Statistical Databases*, Josep Domingo-Ferrer and Melek Önen (eds.). Springer Nature Switzerland, Cham, 329–343. https://doi.org/10.1007/978-3-031-69651-0_22

[170] D.E.D. Torres and C.M.S. Rocco. 2005. Extracting trees from trained SVM models using a TREPAN based approach. In *Fifth International Conference on Hybrid Intelligent Systems (HIS'05)*, 2005. IEEE, Rio de Janeiro, Brazil, 6 pp. https://doi.org/10.1109/ICHIS.2005.41

[171] Anh-Tu Tran, The-Dung Luong, and Van-Nam Huynh. 2024. A comprehensive survey and taxonomy on privacy-preserving deep learning. *Neurocomputing* 576, (April 2024), 127345. https://doi.org/10.1016/j.neucom.2024.127345

[172] 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. 2018. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Ann. Intern. Med.* 169, 7 (October 2018), 467–473. https://doi.org/10.7326/M18-0850

[173] Cristina Trocin, Patrick Mikalef, Zacharoula Papamitsiou, and Kieran Conboy. 2021. Responsible AI for Digital Health: a Synthesis and a Research Agenda. *Inf. Syst. Front.* (June 2021). https://doi.org/10.1007/s10796-021-10146-4

[174] Jonathan Ullman and Salil Vadhan. 2011. PCPs and the Hardness of Generating Private Synthetic Data. In *Theory of Cryptography*, Yuval Ishai (ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 400–416. https://doi.org/10.1007/978-3-642-19571-6_24

[175] Michael Veale, Reuben Binns, and Lilian Edwards. 2018. Algorithms that remember: model inversion attacks and data protection law. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* 376, 2133 (November 2018), 20180083. http://dx.doi.org/10.1098/rsta.2018.0083

[176] Thijs Veugen, Bart Kamphorst, and Michiel Marcus. 2022. Privacy-Preserving Contrastive Explanations with Local Foil Trees. In *Cyber Security, Cryptology, and Machine Learning*, Shlomi Dolev, Jonathan Katz and Amnon Meisels (eds.). Springer International Publishing, Cham, 88–98. https://doi.org/10.1007/978-3-031-07689-3_7

[177] Vy Vo, Trung Le, Van Nguyen, He Zhao, Edwin V. Bonilla, Gholamreza Haffari, and Dinh Phung. 2023. Feature-based Learning for Diverse and Privacy-Preserving Counterfactual Explanations. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, August 06, 2023. ACM, Long Beach CA USA, 2211–2222. https://doi.org/10.1145/3580305.3599343

[178] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electron. J.* (2017). https://doi.org/10.2139/ssrn.3063289

[179] Guan Wang, Charlie Xiaoqian Dang, and Ziye Zhou. 2019. Measure Contribution of Participants in Federated Learning. In *2019 IEEE International Conference on Big Data (Big Data)*, December 2019. IEEE, Los Angeles, CA, USA, 2597–2604. https://doi.org/10.1109/BigData47090.2019.9006179

[180] Ping Wang and Heng Ding. 2024. The rationality of explanation or human capacity? Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance. *Inf. Process. Manag.* 61, 4 (July 2024), 103732. https://doi.org/10.1016/j.ipm.2024.103732

[181] Yongjie Wang, Hangwei Qian, and Chunyan Miao. 2022. DualCF: Efficient Model Extraction Attack from Counterfactual Explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, June 21, 2022. ACM, Seoul Republic of Korea, 1318–1329. https://doi.org/10.1145/3531146.3533188

[182] Samuel D. Warren and Louis D. Brandeis. 1890. The Right to Privacy. *Harv. Law Rev.* IV, 5 (December 1890).

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

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

[185] Michael Winikoff and Julija Sardelic. 2021. Artificial Intelligence and the Right to Explanation as a Human Right. *IEEE Internet Comput.* 25, 2 (March 2021), 116–120. https://doi.org/10.1109/MIC.2020.3045821

[186] Yuncheng Wu, Shaofeng Cai, Xiaokui Xiao, Gang Chen, and Beng Chin Ooi. 2020. Privacy preserving vertical federated learning for tree-based models. *Proc. VLDB Endow.* 13, 12 (August 2020), 2090–2103. https://doi.org/10.14778/3407790.3407811

[187] Yuncheng Wu, Naili Xing, Gang Chen, Tien Tuan Anh Dinh, Zhaojing Luo, Beng Chin Ooi, Xiaokui Xiao, and Meihui Zhang. 2023. Falcon: A Privacy-Preserving and Interpretable Vertical Federated Learning System. *Proc. VLDB Endow.* 16, 10 (June 2023), 2471–2484. https://doi.org/10.14778/3603581.3603588

[188] Anli Yan, Ruitao Hou, Xiaozhang Liu, Hongyang Yan, Teng Huang, and Xianmin Wang. 2022. Towards explainable model extraction attacks. *Int. J. Intell. Syst.* 37, 11 (November 2022), 9936–9956. https://doi.org/10.1002/int.23022

[189] Anli Yan, Ruitao Hou, Hongyang Yan, and Xiaozhang Liu. 2023. Explanation-based data-free model extraction attacks. *World Wide Web* 26, 5 (September 2023), 3081–3092. https://doi.org/10.1007/s11280-023-01150-6

[190] Anli Yan, Teng Huang, Lishan Ke, Xiaozhang Liu, Qi Chen, and Changyu Dong. 2023. Explanation leaks: Explanation-guided model extraction attacks. *Inf. Sci.* 632, (June 2023), 269–284. https://doi.org/10.1016/j.ins.2023.03.020

[191] Guang Yang, Qinghao Ye, and Jun Xia. 2022. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf. Fusion* 77, (January 2022), 29–52. https://doi.org/10.1016/j.inffus.2021.07.016

[192] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. 2019. Neural Network Inversion in Adversarial Setting via Background Knowledge Alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, November 06, 2019. ACM, London United Kingdom, 225–240. https://doi.org/10.1145/3319535.3354261

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

[194] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, July 2018. IEEE, Oxford, 268–282. https://doi.org/10.1109/CSF.2018.00027

[195] Xuefei Yin, Yanming Zhu, and Jiankun Hu. 2022. A Comprehensive Survey of Privacy-preserving Federated Learning: A Taxonomy, Review, and Future Directions. *ACM Comput. Surv.* 54, 6 (July 2022), 1–36. https://doi.org/10.1145/3460427

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

[197] Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, and Jiliang Tang. 2024. The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG). Retrieved April 22, 2024 from http://arxiv.org/abs/2402.16893

[198] Xiaoyu Zhang, Chao Chen, Yi Xie, Xiaofeng Chen, Jun Zhang, and Yang Xiang. 2023. A survey on privacy inference attacks and defenses in cloud-based Deep Neural Network. *Comput. Stand. Interfaces* 83, (January 2023), 103672. https://doi.org/10.1016/j.csi.2022.103672

[199] Yifei Zhang, Dun Zeng, Jinglong Luo, Xinyu Fu, Guanzhong Chen, Zenglin Xu, and Irwin King. 2024. A Survey of Trustworthy Federated Learning: Issues, Solutions, and Challenges. *ACM Trans. Intell. Syst. Technol.* 15, 6 (December 2024), 1–47. https://doi.org/10.1145/3678181

[200] Zijiao Zhang, Chong Wu, Shiyou Qu, and Xiaofang Chen. 2022. An explainable artificial intelligence approach for financial distress prediction. *Inf. Process. Manag.* 59, 4 (July 2022), 102988. https://doi.org/10.1016/j.ipm.2022.102988

[201] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for Large Language Models: A Survey. *ACM Trans. Intell. Syst. Technol.* (January 2024), 3639372. https://doi.org/10.1145/3639372

[202] Jiaqi Zhao, Hui Zhu, Fengwei Wang, Rongxing Lu, and Hui Li. 2023. Efficient and privacy-preserving tree-based inference via additive homomorphic encryption. *Inf. Sci.* 650, (December 2023), 119480. https://doi.org/10.1016/j.ins.2023.119480

[203] Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Lim. 2021. Exploiting Explanations for Model Inversion Attacks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. IEEE, Montreal, QC, Canada, 662–672. https://doi.org/10.1109/ICCV48922.2021.00072

[204] Xiubin Zhu, Dan Wang, Witold Pedrycz, and Zhiwu Li. 2022. Horizontal Federated Learning of Takagi–Sugeno Fuzzy Rule-Based Models. *IEEE Trans. Fuzzy Syst.* 30, 9 (September 2022), 3537–3547. https://doi.org/10.1109/TFUZZ.2021.3118733

[205] Covidence. Retrieved May 5, 2024 from https://www.covidence.org