A View on Vulnerabilites: The Security Challenges of XAI

Elisabeth Pachl
Fraunhofer Institute for Cognitive Systems, Munich, Germany

Fabian Langer
TÜV Informationstechnik GmbH, Artificial Intelligence, Essen, Germany

Thora Markert
TÜV Informationstechnik GmbH, Artificial Intelligence, Essen, Germany

Jeanette Miriam Lorenz
Fraunhofer Institute for Cognitive Systems, Munich, Germany

Abstract

Modern deep learning methods have long been considered as black-boxes due to their opaque decision-making processes. Explainable Artificial Intelligence (XAI), however, has turned the tables: it provides insight into how these models work, promoting transparency that is crucial for accountability. Yet, recent developments in adversarial machine learning have highlighted vulnerabilities in XAI methods, raising concerns about security, reliability and trustworthiness, particularly in sensitive areas like healthcare and autonomous systems. Awareness of the potential risks associated with XAI is needed as its adoption increases, driven in part by the need to enhance compliance to regulations. This survey provides a holistic perspective on the security and safety landscape surrounding XAI, categorizing research on adversarial attacks against XAI and the misuse of explainability to enhance attacks on AI systems, such as evasion and privacy breaches. Our contribution includes identifying current insecurities in XAI and outlining future research directions in adversarial XAI. This work serves as an accessible foundation and outlook to recognize potential research gaps and define future directions. It identifies data modalities, such as time-series or graph data, and XAI methods that have not been extensively investigated for vulnerabilities in current research.

2012 ACM Subject Classification Computing methodologies → Machine learning

Keywords and phrases Explainability, XAI, Transparency, Adversarial Machine Learning, Security, Vulnerabilities

Digital Object Identifier 10.4230/OASIcs.SAIA.2024.12

Category Academic Track

Funding This research was funded by the Bavarian Ministry for Economic Affairs, Regional Development and Energy, as part of a project to support the thematic development of the Fraunhofer Institute for Cognitive Systems.

1 Introduction

Ever since the wide adoption of machine learning (ML), the scientific community has striven for ways to make decision-making processes based on artificial intelligence (AI) transparent, creating the field of explainable AI (XAI) [92, 46]. Transparency is critical for maintaining accountability, especially in high-risk scenarios like autonomous vehicles encountering obstacles or medical AI systems determining patient treatments. Stakeholders – including professionals utilizing AI (e.g., physicians), end-users affected by AI decisions (e.g., patients), and AI developers – benefit from understanding AI-based decisions.

The increased adoption of XAI methods is driven in part by the need to enhance compliance with regulatory frameworks such as the General Data Protection Regulation (GDPR) [82] and the EU AI Act [83]. The GDPR preserves the *"right to explanation"* [54], encouraging organizations to provide understandable reasoning behind AI-driven decisions related to personal data. Similarly, the EU AI Act mandates transparency and human oversight for high-risk AI systems. Beyond compliance to regulations, XAI fosters trust among stakeholders, enhances the detection of biases or inaccuracies, and aligns with ethical principles such as fairness and accountability. Additionally, XAI aids continuous improvement by enabling developers to refine models based on insights from comprehensible decision-making processes.

Despite these advantages, adversarial ML (AdvML) [56, 89, 70] has become increasingly prevalent in XAI research, raising concerns regarding trustworthiness, robustness, and security [80]. This trend underscores the importance of scrutinizing XAI for potential vulnerabilities that adversarial attacks might exploit. While XAI aims to make AI systems more transparent and fair, the misuse of explainability can paradoxically be harnessed to amplify attacks on AI systems, posing threats such as evasion and privacy breaches.

Our study makes several significant contributions. We present a systematic categorization of the XAI attack surface, drawing insights from a comprehensive review of over 70 publications. This categorization helps in understanding and addressing the vulnerabilities inherent in the application of XAI. By providing concrete examples from scientific literature, we highlight specific risks associated with XAI usage. The categorization organizes attacks according to classes of XAI methods and their application domains, enabling readers to identify relevant attack vectors quickly. We discuss several aspects for the secure and robust development of AI systems incorporating XAI along the AI lifecycle. These considerations shall support the mitigation of the introduced vulnerabilities of XAI. Lastly, our work serves as a foundation for recognizing potential research gaps and defining future directions. It identifies domains and XAI methods scrutinized for vulnerabilities, guiding further investigation into countermeasures against documented attacks. It also highlights areas with scarce published attacks, suggesting potential novel attack vectors for exploration.

To the best of our knowledge, the provided information reflects the status up to March 2024. We incorporated papers from surveys on the robustness and reliability of XAI against attacks [15, 23, 77] and included notable papers from major ML conferences and journals, leveraging their citation networks to identify other relevant works.

The rest of the paper is organized as follows: Section 2 describes the background and positioning of our work within existing surveys on XAI, privacy and AdvML. Section 3 presents an overview of the attack landscape surrounding XAI, focusing on attacks on XAI and XAI-enhanced attacks. We discuss how our work can identify new attack vectors and research gaps, providing practical insights for different stakeholders of AI systems. To mitigate vulnerabilities of XAI methods, we examine certain aspects connected to the secure development of AI systems utilizing XAI in Section 4. Finally, Section 5 concludes the paper with an outlook.

2 Background

Here, we provide a brief introduction to the methodology in AdvML and XAI. Readers familiar with basic concepts of these can skip to Section 3.

2.1 Notation

Based on Baniecki and Biecek [15], we adopt a simplified notation for our work. We mainly focus on supervised classification tasks where a model $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$, parameterized by θ, maps a d-dimensional input $\mathbf{x}$ from the feature space $\mathcal{X} \in \mathbb{R}$ to probability scores for each possible class $c \in [C]$ as a C-dimensional vector in $\mathcal{Y} \in [0,1]^C$. The predicted class is determined by selecting the class index with the highest probability. For simplicity, we will refer to the prediction model as f. Let $\mathbf{x} \in \mathcal{X}$ represent the input vector for which we seek to explain the prediction $f(\mathbf{x})$. Consider an explanation function $g(\cdot,\cdot)$, where both the model f and $\mathbf{x}$ serve as inputs, yielding varying outputs depending on the underlying XAI method. To facilitate the categorization of the attack surface, we use specific symbols: $\rightarrow$ denotes a change in the given object, e.g., a small perturbation in the input: $\mathbf{x} \rightarrow \mathbf{x'}$; $\approx$ and $\neq$ denote similarity between two values, e.g., similar predictions $f(\mathbf{x}) \approx f(\mathbf{x'})$ or input features $\mathbf{x} \approx \mathbf{x'}$ and dissimilarity, e.g., different explanations $g(f,\mathbf{x}) \neq g(f,\mathbf{x'})$, respectively.

2.2 Explainable Artificial Intelligence

Similar to ML, XAI is a particularly wide field of research. Thus, in this section, we step back to detail the scope considered in our work. We emphasize that we do not attempt to summarize the field of XAI and refer the reader to surveys on the topic [3, 21, 26, 40, 115, 91].

XAI has found utility across various domains, including regulatory audits [57], cybersecurity [89], drug discovery [52] or model debugging [47].

Different XAI methods can be categorized based on different perspectives (Figure 1). Firstly, we distinguish between intrinsic explainable ML models and analyzing the model's outcome after training (post-hoc XAI methods). Intrinsic explainable ML models, like decision trees or attention-based neural networks, generate explanations concurrently with predictions [10]. In contrast, post-hoc explanations involve XAI methods applied at inference (e.g., Local Interpretable Model-agnostic Explanations (LIME) [87] or Gradient-weighted Class Activation Mapping (Grad-CAM) [93]). Secondly, XAI techniques can be classified as model-specific or model-agnostic. Model-specific methods are tailored to explain one specific model or a model group, while model-agnostic approaches can be applied to any ML model. The latter analyze feature importance without accessing internal model information such as weights. Examples include LIME [87], Shapley Additive Explanations (SHAP) [67], Saliency Map [51], Grad-CAM [93] or counterfactual explanations [105]. Local explainability focuses on why a specific decision was made for a single prediction instance. In contrast, global explainability offers insights into the overall decision-making process for the entire dataset. Local post-hoc feature attribution, can be obtained using perturbations-based XAI methods like Shapely values [67]. Specifically for tabular data, these allow to explain individual predictions in a model-agnostic manner. Contrary, gradient-based [68] and propagation-based [11] local post-hoc XAI methods, summarized as backpopagration-based methods in this work, e.g., Grad-CAM [93] or saliency maps [51], are specific to neural networks. Those methods leverage the principles of gradient descent to attribute importance to input features, necessitating access to the internals of the model f. Counterfactual examples are a popular approach that shows how much a specific input feature needs to change to alter the prediction outcome. Complementary to local explanations, global explanations summarize consistent patterns in model predictions across the data, such as feature importance and feature effect visualizations (e.g., partial dependence plots). For deep neural networks, concept-based explanations [43] relate human-understandable concepts to predicted classes, such as how a "stop sign" prediction is influenced by the presence of an octagon shape in an image.

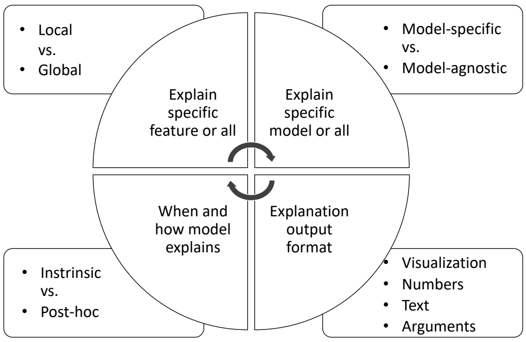

Figure 1 An overview diagram showing the categorization of XAI in different aspects. Adopted from Zhang et al. [116].

In this study, we focus on post-hoc explanation methods, which offer the advantage of being versatile and applicable to a wide range of models. Additionally, the separation between the learning task and explaining its outcomes allows for evaluating threat models for the XAI method independently of the learning task. However adversaries could exploit this disparity between model's inference and explanation, leading to discrepancies between reported predictions and explanations.

2.3 Adversarial Machine Learning

AdvML has emerged over the last 20 years as a critical research area. One major goal of advML is to – unnoticeably – alter the model's behaviour. The most explored class of attacks focuses on the vulnerabilities of ML models to malicious inputs, known as adversarial examples [16, 31]. These adversarial inputs are strategically crafted with the intent to fool a model into misclassification, thereby posing significant threats to the reliability and trustworthiness of AI systems deployed in real-world. Adversarial examples can manifest across various data modalities, including text, tabular data, and images. For instance, in text data [7], adversaries may manipulate input text to introduce subtle changes like misspellings, and swapping words or characters [90]. Similarly, in tabular data [16, 12], adversaries may tamper with specific features to alter model predictions, whereas in image data, adversarial patches or pixels, unrecognizable by humans, added to input images can cause models to misclassify them [19, 102]. Various techniques have been developed to generate adversarial examples effectively. These techniques include gradient-based methods such as the Fast Gradient Sign Method (FGSM) [37], iterative approaches like Projected Gradient Descent (PGD) [71], and optimization-based methods like the Carlini & Wagner (C&W) attack for images [20]. Possible defenses include augmenting training data with diverse examples, model regularization [39], such as dropout and weight decay, and distillation [81], which involves training a more robust "teacher" model on original data and using its predictions as soft labels to train a "student" model.

In addition to adversarial examples, adversaries can exploit other attack vectors to compromise ML models. Backdoor attacks involve injecting malicious triggers or patterns into training data, leading to targeted misclassifications during inference [25]. These attacks typically involve poisoning the training data to ensure that the adversarial model remains indistinguishable from the desired one. Moreover, various poisoning attacks have been proposed targeting different adversarial goals, including decreasing classification accuracy or causing targeted misclassifications to evade detection. We refer to [27], for a comprehensive systematization of poisoning attacks and defenses related to model predictions.

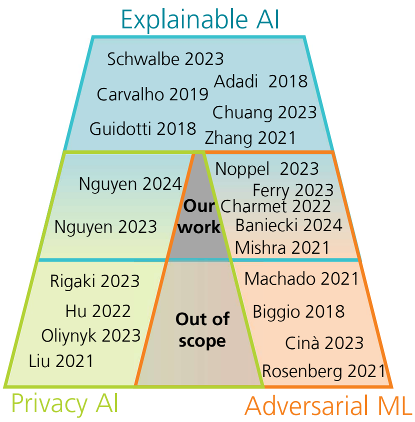

Figure 2 Overview of our holistic approach combining XAI, privacy breaches, and AdvML compared to existing surveys.

Privacy attacks and model stealing attacks pose additional threats, compromising the integrity and security of AI systems. Privacy attacks seek to extract sensitive information from models [80], while model stealing attacks involve reverse-engineering model architectures or parameters using predictions or access to black-box APIs [78]. For a detailed overview of privacy attacks and defense strategies, readers are directed to the work of Rigaki et al. [88].

2.4 Comparison to Existing Surveys

Many surveys have categorized different XAI methods for ML methods and provided guidance in selecting suitable techniques for desired explanations, e.g., [3, 21, 26, 40, 115, 91], while others summarized issues related to model and data privacy in AI systems, e.g., [77, 33, 23, 15, 73] or reviewed problems related to AdvML, e.g., [70, 16, 27, 89].

Some existing surveys cover the intersection between two of the aforementioned categories (Figure 2). On the intersection between XAI and AdvML, Ferry et al. [33] examined the interplay between interpretability, fairness, and robustness, whereas Baniecki et al. [15] surveyed adversarial attacks on model explanations and fairness metrics, offering a unified taxonomy for clarity across related research areas and discussing defenses against such attacks. Noppel et al. [77] summarized attacks designed to subvert explanations based on their objectives, e.g., preserving or altering explanations, formalized notions of adversarial robustness in the presence of explanation-aware attacks, and presented a taxonomy of existing defenses. Charmet et al. [23] focused on adversarial attacks targeting XAI methods within the cybersecurity domain. They explored various attack vectors and proposed defensive strategies to maintain the fairness and integrity of XAI models. Mishra et al. [73] focused on the robustness of XAI and attacks against it. They unify existing definitions of robustness of XAI, introduce a taxonomy to classify different robustness approaches as well as some pointers about extending current robustness analysis approaches so as to identify reliable XAI methods. To the best of our knowledge, Nguyen et al. [75, 74] are the first to summarize in-depth the knowledge in the intersection between XAI and privacy AI. However, these papers only present partial coverage of the entire safety and security landscape surrounding XAI. We note, that there are works on the intersection between privacy AI and AdvML, but this is not the focus of this work.

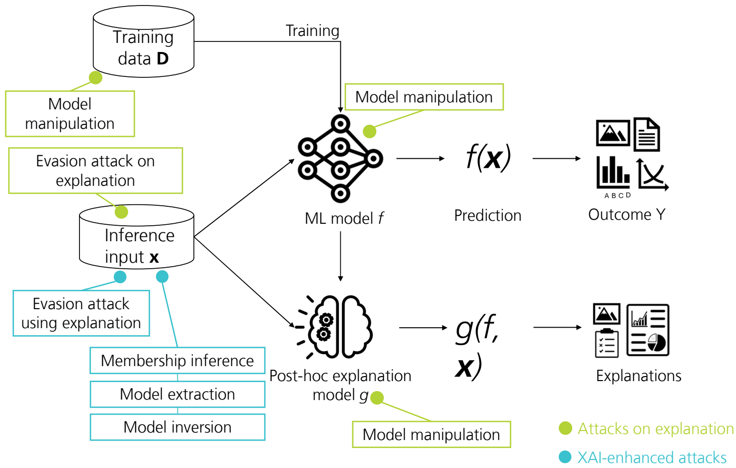

Figure 3 Attack surface against XAI systems. Trained model f predicts class label y for input $\mathbf{x}$. g represents a post-hoc XAI method deriving an explanation of the input sample. Attacks can be directly against explanations (green) or XAI knowledge can be used to enhance privacy attacks or attacks against predictions (blue).

Our contribution

Our survey presents an in-depth examination of evasion and privacy breaches related to XAI, diverging from previous work by its comprehensive nature and addressing the full spectrum of possible attack vectors. We delve into the underlying principles, methodologies, and taxonomies, while also mapping out potential trajectories for future research. Especially, our work goes beyond the individual matching of defenses to specific attacks, as seen in previous studies [15, 23]. Instead, we comprehensively discussed several aspects for the secure and robust development of AI systems incorporating explicit XAI considerations along the AI lifecycle. These considerations shall support the mitigation of the introduced vulnerabilities and evolving threats of XAI.

3 Attack Landscape Surrounding XAI

While prior efforts focused predominantly on the robustness and reliability of XAI [73, 15, 77], attacks on predictions [70, 27] or XAI in AdvML [66], our work distinguishes itself by also focusing on the misuse of explainability to amplify attacks on AI systems. Broadly, the attack surface can be categorized into *attacks on XAI* and *XAI-enhanced attacks* (Figure 3). The robustness of post-hoc XAI methods and their vulnerability to adversarial examples is addressed in the context of attacks on explanations. On the other hand, when XAI is used to enhance attacks on AI systems, such as altering model predictions or compromising model privacy, these attacks fall into the category of XAI-enhanced attacks.

Table 1 and 2 lists attacks across both categories, specifying the application domain e.g., computer vision using image data, and the type of attacked XAI e.g., local vs. global or backpropagation-based vs. perturbation-based.

3.1 Attacks on XAI

We proceed to specify different types of attacks that alter explanations. We use the terms *evasion attack* and *model manipulation* based on the attack point.

3.1.1 Evasion Attack on XAI

In this scenario, an adversarial example, based on a benign inference input, is crafted to manipulate the explanation without impacting the prediction of a deployed AI system:

$$\mathbf{x} \rightarrow \mathbf{x'} \Longrightarrow \begin{cases} g(f, \mathbf{x}) \neq g(f, \mathbf{x'}) \\ f(\mathbf{x}) \approx f(\mathbf{x'}) \end{cases}$$

Here, the adversarial example $\mathbf{x'}$ is constructed so that its explanation matches a target explanation, while maintaining the prediction model's f output [31, 35]. Note that the target explanation differs from the original explanation of $\mathbf{x}$. For instance, in medical imaging, an attacker could alter an AI-interpreted CT image, changing the highlighted region indicative of malignancy or benignancy of cancer while preserving the diagnosis. This could mislead a radiologist in selecting biopsy locations, potentially compromising patient care and outcomes.

3.1.2 Model Manipulation on XAI

In this scenario, the attack involves manipulating the prediction model f or the local post-hoc explanation model g. In model manipulation of f, such as weight manipulation, fine-tuning through an expanded loss function [30, 44] or poisoning of training data [113], the altered model f' generates different explanations for the same input data $\mathbf{x}$, while maintaining similar predictions (e.g., [44, 30, 76]):

$$f \rightarrow f' \Longrightarrow \begin{cases} g(f, \mathbf{x}) \neq g(f', \mathbf{x}) \\ f(\mathbf{x}) \approx f'(\mathbf{x}) \end{cases}$$

Further, neural networks can have backdoors triggered by specific input patterns to retrieve original explanations [104, 76]. A few works also consider how original or manipulated explanations can be used to cover an adversarial change in the model's prediction such as misclassifications [76]. This can be used to, e.g., disguise fraudulent activities in a financial fraud detection system. For instance, in a financial fraud detection system using LIME or SHAP, an attacker could manipulate the neural network's weights to disguise fraudulent transactions as legitimate. This manipulation alters the explanations to make fraudulent activity appear normal, justifying decisions that label fraudulent transactions as legitimate.

Similarly, by manipulating the local post-hoc explanation model g, the altered model g' produces different explanations for the same input data x, despite identical predictions by f [60]: $g \rightarrow g' \Longrightarrow g(f, \mathbf{x}) \neq g'(f, \mathbf{x})$

While the majority of attacks are on local XAI methods, a few specifically target global XAI [60, 13, 18, 14, 62]: $g \rightarrow g' \Longrightarrow \forall x \in X g(f, \mathbf{x}) \neq g'(f, \mathbf{x})$.

3.1.3 Observation

The majority of proposed attacks on explanations assume prior knowledge about the model's architecture, weights, and the XAI method used. For evasion attacks, attackers need access to the model's parameters to craft adversarial examples effectively, typically using gradient-based methods to modify inputs, changing explanations without altering predictions [44, 35]. Such attacks can be executed by individuals with significant technical expertise, such as AI developers or malicious insiders with model access. The same level of knowledge and access is necessary for model manipulation attacks, involving altering the model, XAI method, or leaving a backdoor in the model. Although generally less technically equipped, sophisticated end users with malicious intent might exploit available tools and methods to perform attacks if they can gain sufficient access to the model [29].

When considering the implementation of XAI methods, it is crucial to evaluate the advantages and disadvantages of local and global approaches. Local explanations provide insights into individual predictions, useful for case-by-case assessments, but are susceptible to adversarial manipulation, leading to significant global impacts on model behavior [44, 30, 76, 9]. Global explanations offer a comprehensive view of the model's decision-making process but can also be manipulated to affect local explanations [60, 13, 59]. Improving detection mechanisms for one type of explanation could enhance detectability across the board [55, 35].

Generally, research on attacks aiming to alter only the explanation while preserving the prediction, such as *XAI-washing* or *fair-washing*, is sparse. We found most works studying perturbation- and permutation-based XAI methods, whereas only few exist on concept-based explanations [18], counterfactual explanations [100], and interpretable models like decision trees [62]. Additionally, explanations for language models based on text [98, 22, 50], graphs [63], and time-series data like audio [45] remain underexplored. Understanding the robustness of post-hoc XAI models to adversarial attacks in real-world applications based on underexplored data modalities is crucial. In healthcare, text data from patient records, graph data from molecular structures, and time-series data from patient monitoring systems are commonly used. In finance, transaction data, customer feedback, and network analysis for fraud detection are critical.

3.2 XAI-enhanced Attacks on Predictions

XAI methods can be exploited by adversaries to enhance attacks on AI systems. In XAI-enhanced attacks on predictions, adversarial examples are crafted using additional knowledge from XAI methods to fool AI models into making inaccurate predictions while maintaining similar explanations:

$$\mathbf{x} \to \mathbf{x'} \implies \begin{cases} g(f, \mathbf{x}) \approx g(f, \mathbf{x'}) \vee g(f, \mathbf{x}) \neq g(f, \mathbf{x'}) \\ f(\mathbf{x}) \neq f(\mathbf{x'}) \end{cases}$$

Here, the primary goal is to make the model produce wrong predictions with consistent explanations, unlike evasion attacks where the focus is solely on altering explanations. Attackers may also aim to change both predictions and explanations to fully disguise the AI system [58], though this is more detectable due to changes in the model's behavior.

For XAI-enhanced adversarial example crafting, gradient-based or perturbation-based explanations are used for images or tabular data to identify important pixels or features. Perturbations are added only to these areas to deceive the classifier, maintaining a high attack success rate with fewer pixel changes and reducing the optimization space and redundancy of local perturbations [41, 53, 64, 114, 2, 65]. This makes these attacks more efficient and less resource-intensive. Furthermore, with additional knowledge from XAI methods, XAI-enhanced attacks are also possible in black-box settings without any knowledge of the target model and its coupled interpreter[2, 58, 112] in contrast to white-box settings, where the attacker has full knowledge of the model [1, 53, 64]. Abdukhamidov et al. [2] demonstrated a transfer-based and score-based technique using a microbial genetic algorithm, achieving high attack success with minimal queries and high similarity in interpretations between adversarial and benign samples.

Observation

XAI-enhanced attacks on predictions pose a critical concern for AI providers due to their increased feasibility and lower execution barriers compared to attacks on explanations. We found studies across different data modalities, including image [41, 53, 64, 65, 114, 112, 1, 2, 8, 42], tabular [58], textual [107, 22], and graph data [24, 63, 108]. However, time-series data, including audio in natural language processing or sensor data from vehicle systems or patient monitoring in ICUs, remain underexplored. Overall, the potential for using XAI knowledge to enhance evasion attacks is promising but largely untapped. Although there is a substantial body of literature on the subject of evasion attacks, including work on training data poisoning, backdoor attacks and adversarial example crafting, our focus remains on XAI-enhanced attacks. A review of the literature revealed no previous studies in this field, indicating that this represents an as yet unidentified threat.

3.3 XAI-enhanced Attacks on Privacy

With the growing use of XAI methods, new vectors for privacy breaches in AI systems emerge. This section covers three primary categories of XAI-enhanced attacks on privacy: *model inversion*, *model extraction*, and *membership inferences*. These attacks leverage the additional information provided by explanations to enhance their efficiency and effectiveness.

3.3.1 Model Inversion

Model inversion attacks aim to reconstruct input data from model outputs, potentially revealing sensitive information about individuals in the training set. For example, a gender recognition scenario, an attacker might use XAI-enhanced model inversion to reconstruct facial images from the outputs of a gender classification model (prediction and explanations), leading to unauthorized re-identification and privacy violations. These attacks typically assume a black-box scenario with query-access only, where the attacker receives model predictions and explanations for a given instance $\mathbf{x}$. Studies have shown that explanations from backpropagation-based methods (e.g., Gradient, Grad-CAM) can significantly improve reconstruction accuracy compared to using predictions alone [117, 32, 69].

Zhao et al. [117] demonstrated enhanced model inversion attacks using XAI-aware model inversion architectures, such as multi-modal, spatially-aware CNNs. They found vulnerability varies by explanation method as they provide different levels of additional information: LRP < Gradient < CAM < Gradient x Input. Duddu et al. [32] showed that sensitive attributes can be inferred from model explanations and predictions, even when not explicitly included in input or outcome. Attacks were more successful using backpropagation-based explanations like SmoothGrad or IntegratedGradients compared to predictions alone. Luo et al. [69] focused on feature inference attacks using Shapley value, demonstrating significant advantages over prediction-only attacks in reconstructing private model inputs.

3.3.2 Model Extraction

Model extraction attacks aim to steal the functionality of a ML model by creating a surrogate model that mimics the target model's decision behaviour. Typically, the target model is deployed through an API, providing the attacker with black-box access. These attacks involve three steps: (1) collecting or synthesizing an initial unlabeled dataset, (2) querying the target model with inputs, and (3) training a surrogate model using the attack dataset annotated by the target model. It is often assumed that the attacker has an auxiliary dataset following the same distribution as the target model's training data. XAI-enhanced model extraction attacks additionally leverage explanations of queried instances. For example, in the financial sector, an attacker could use XAI-enhanced model extraction to steal a proprietary credit scoring model, undermining the original model owner's competitive advantage by replicating the sophisticated decision-making process.

Milli et al. [72] demonstrated that gradient-based explanations reveal model information more efficiently than traditional label-only queries. Their experiments showed that achieving 95% accuracy required only 10 gradient-queries, receiving predictions and explanations, compared to 1000 label-only queries for a convolutional model (CNN) on MNIST. Yan et al. [111] introduced XAMEA, an explanation-guided model extraction attack, achieving a 25% reduction in required queries for CIFAR-10 compared to traditional methods. They found Grad-CAM explanations posed the greatest risk of privacy leakage. Aïvodji et al. [5] focused on model extraction attacks using counterfactual explanations, showing an attacker could achieve over 90% fidelity with only 250 queries, significantly outperforming baseline attacks without explanations. Their findings underscored that counterfactual explanations enable high-fidelity and high-accuracy model extractions even under limited query budgets. Wang et al. [106] proposed DualCF, a querying strategy that greatly reduces the number of required queries for model extraction. Their method sequentially queries the target model with counterfactual explanations, achieving better agreement scores and lower sensitivity to sampling procedures compared to baseline methods.

3.3.3 Membership Inference

Membership inference attacks (MIAs) pose a significant threat by allowing adversaries to determine if specific data points were part of a model's training set. This can be particularly problematic in critical areas like healthcare, where sensitive information could be inferred without consent. For instance, an adversary could query a hospital's ML model for rare disease diagnosis and determine whether specific individuals' medical records were used in training.

MIAs typically assume a black-box scenario with access to model predictions and explanations. They aim to predict the membership status of data points within the attack set. XAI-enhanced and prediction-only MIAs use two main strategies: threshold-based and reference model-based attacks [96]. Threshold-based attacks rely on output variance, assuming training set data points yield lower variance in predictions and explanations due to the model's familiarity with the data. Reference model-based attacks use shadow models to simulate the target model's behavior and derive membership inference thresholds. This method assumes access to similar data and knowledge of the model's architecture and hyperparameters.

Shokri et al. [96] pioneered investigating using model explanations for inferring private information about training data. They proposed a threshold-based attack using prediction and explanation variance, revealing significant privacy risks with backpropagation-based explanations in tabular datasets, but not in image datasets. They attributed this to fluctuating gradient variance. While it is entirely possible that perturbation-based methods are vulnerable to membership inference, the authors conjecture that this is not the case. Pawelczyk et al. [84] highlighted privacy risks from algorithmic recourse, introducing counterfactual distance-based attacks that infer membership without auxiliary data or model details. These attacks excelled with overfitting models and high data dimensionality. Goethals et al. [36] introduced explanation linkage attacks, where adversaries use quasi-identifiers from counterfactual explanations to re-identify individuals by linking with background information. They also proposed k-anonymous counterfactual explanations to mitigate these risks.

3.3.4 Observation

XAI-enhanced privacy attacks significantly increase the effectiveness and efficiency of model and data privacy breaches, making them more feasible in real-world scenarios. These attacks require minimal prior knowledge and model access and reduce the number of queries needed for successful breaches. The effectiveness of MIAs varies by data modality, with tabular and high-dimensional data being more susceptible [96]. Additionally, backpropagation-based methods are more vulnerable to MIAs than perturbation-based methods [96]. Counterfactual explanations pose significant risks for both MIAs and model extraction attacks [5, 106, 84, 36], although no work has been found addressing model inversion using counterfactual explanations. Research gaps exist in studying XAI methods' vulnerability in MIAs for tabular data and model inversion attacks using counterfactual explanations. Current model inversion studies focus primarily on tabular data [32, 69], with no work on textual data.

3.4 Practical Applications of the XAI Attack Vector Classification Table

The application of XAI methods can have a significant impact on a system's security and safety. Depending on the system at hand and the implemented XAI method, different attack vectors may apply. Tables 1 and 2 provide an overview of published attacks against XAI and XAI-enhanced attacks against AI systems, extending Baniecki et al.'s work [15].

The tables arranges studies by data modalities, groups of XAI methods, and attack types (privacy, prediction, and attacks on XAI). More granular attack subcategories further describe the type of attack presented in the referenced works. Table 1 covers the topic of computer vision, while the papers introduced in Table 2 deal with graphs, textual, numerical and time-series/audio data.

These tables serve multiple stakeholders: *Developers* can identify potential vulnerabilities early in the design and development stage. By understanding specific attack vectors associated with different XAI methods, they can proactively implement countermeasures and design more secure models. Section 4 shall support the development of secure and safe AI systems leveraging XAI. *Users* gain insights into limitations and risks associated with system explanations, recognizing potential compromises or errors [17]. An overview of attacks based on explainability methods equips *evaluators* with the necessary knowledge to conduct thorough and informed risk assessments and later on perform targeted vulnerability testing to verify the system's robustness against these kind of attacks. *Researchers* can identify knowledge gaps, explore new attack vectors, develop novel defense mechanisms, and enhance existing XAI methods.

4 Aspects of XAI Attack Mitigation

Our comprehensive analysis of potential attacks on and enhanced by XAI should not deter its use but rather highlight latent risks. Despite these risks, however, explainability offers such significant benefits that it should not be dispensed with. Under certain circumstances, it may even be necessary to use XAI methods in order to improve adherence with transparency obligations, such as those stated in the EU AI Act [83].

In the following, we present aspects connected to the responsible implementation and use of XAI methods throughout the first phases of the AI life cycle in accordance with ISO/IEC 22989 [49]. For the phases from inception to verification and validation, specific considerations are highlighted in order to mitigate potential risks and ensure the secure and safe use of XAI (Figure 4).

Table 1 Summary of adversarial attacks on explanations and XAI-enhanced attacks on model predictions and privacy for data modality of images.

			XAI-enhanced on Privacy			XAI-enhanced Attacks on Predictions	Attacks on XAI	
Data Modality	**XAI Method**		**Model Inversion**	**Model Extraction**	**Membership Inference**	**Evasion Attack**	**Model Manipulation**	**Evasion Attack**
Computer Vision	Local	Backpropagation	Zhao [117]	Yan [111], Milli [72], Yan [110], Yan [109]	Shokri [96]	Guo [41], Jing [53], Liu [64], Zhan [112], Zhang [114], Abdukhamidov [1], Abdukhamidov [2]	Heo [44], Noppel [76], Kindermans [55], Viering [104], Aïvodji [4], Zhang [113]	Anders [9], De Aguiar [29], Dombrowski [31], Ghorbani [35], Göpfert [38], Galli [34], Huang [48], Le [61], Pandya [79], Rasaee [85], Renkhoff [86], Song [101], Subramanya [103], Zhang [114], Kindermans [55], Si [97]
		Perturbation		Yan [111], Yan [110], Yan [109]	Shokri [96]	Amich [8], Liu [65]		Göpfert [38], Pandya [79]
		Counterfactual						
		Perturbation						
		Decision Tree				Hada [42]		
		Concept-based					Brown [18]	

Table 2 Summary of adversarial attacks on explanations and XAI-enhanced attacks on model predictions and privacy for data modality of textual, numerical, graph and time-series (TS) inlcuding audio data.

			XAI-enhanced on Privacy			XAI-enhanced Attacks on Predictions	Attacks on XAI	
Data Modality	**XAI Method**		**Model Inversion**	**Model Extraction**	**Membership Inference**	**Evasion Attack**	**Model Manipulation**	**Evasion Attack**
Textual & Numerical	Local	Backpropagation	Duddu [32]			Kuppa and Le-Khac [58], Xu and Du [107]	Dimanov [30], Zhang [113]	Ali [6], Anders [9], Ivankay [50], Sinha [98], Kuppa and Le-Khac [58]
		Perturbation	Duddu [32]			Chai [22]	Baniecki and Biecek [13], Dimanov [30], Severi [94]	Ali [6], Sinha [98], Slack [99]
		Counterfactual		Aïvodji [5], Wang [106]	Pawelczyk [84], Goethals [36]		Slack [100]	
	Global	Backpropagation	Luo [69]					
		Perturbation					Baniecki and Biecek [13], Baniecki [14]	Laberge [59], Lakkaraju [60]
		Interpretable Models					Le Merrer [62]	
Graph	Local	Backpropagation + perturbation				Chen [24], Li [63], Xu [108]		Li [63]
Audio/TS	Local	Backpropagation						Hoedt [45]
		Perturbation						Hoedt [45]

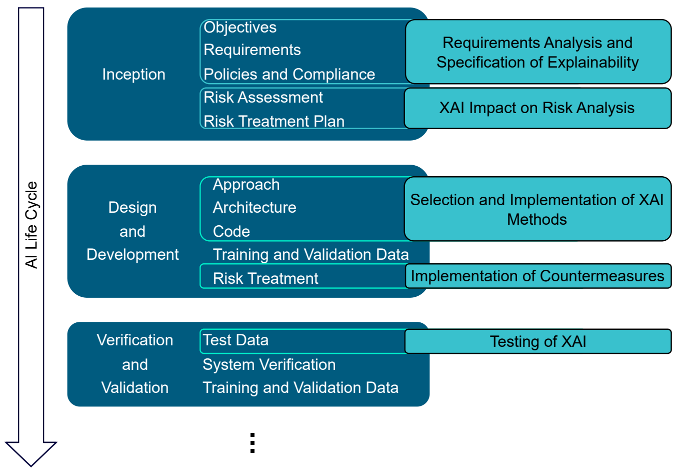

Figure 4 Aspects for the secure development of AI systems incorporating XAI along the AI life cycle.

Requirements Analysis and Specification of Explainability

In inception, assessing the necessity and benefits of explainability is crucial. The superior reason for integrating methods for explainability into a system is creating transparency for different stakeholders, providing insight into the system's general functionality or specific model operations. It is an important step to be clear in advance about the requirements coming from various sides that need to be fulfilled. For example, these requirements may originate from regulation (e.g., Article 13 of the EU AI Act [83]), adaption of standards and best-practices (e.g., Microsoft's Responsible AI Standard [28]) or business goals. Subsequently, the identified requirements must be specified regarding use case (including used data), the planned system architecture and environment. The goal is to formulate concise requirements for explainability, so that only the required level of transparency is provided and no unnecessary information is disclosed.

XAI Impact on Risk Analysis

The integration of XAI into a system can introduce new risks and potential attack vectors, significantly affecting risk analysis. While XAI enhances transparency and trust in AI systems by providing clear and interpretable insights, it also necessitates a thorough reassessment of security vulnerabilities and risk management strategies. One primary risk introduced by XAI is the potential exposure of the model's inner workings to adversaries. XAI methods reveal how models make decisions, inadvertently disclosing sensitive aspects like feature importance and decision pathways. This transparency can be exploited to launch targeted attacks (Section 3.2). Thus, detailed insights provided by XAI necessitate robust security measures to protect the model from exploitation. Additionally, XAI techniques can increase the risk of privacy attacks (Section 3.3). This is particularly concerning in applications involving personal or confidential data, such as healthcare or financial services. The enhanced interpretability offered by XAI can make it easier for attackers to infer training data and thus private information. Corresponding privacy-preserving techniques shall be considered. The integrity of the explanations themselves is another critical concern. If explanations can be manipulated (Section 3.1), the trustworthiness of the entire AI system can be compromised. Attackers might alter explanations to hide malicious activities or to falsely assure users of the model's reliability. The potential attack vectors depend on the selected XAI method and the domain the system is operating in. Table 1 and Table 2 provide a state-of-the-art overview of potential attacks based on various XAI methods in different domains to support the risk analysis process.

Despite these challenges, integrating XAI can enhance overall risk management by providing clearer insights into model behavior and decision-making processes. This transparency can help identifying potential biases and vulnerabilities within the model, enabling more effective mitigation strategies. By understanding how models arrive at their decisions, organizations can implement targeted defenses against specific risks and continuously monitor and improve the AI system's security posture.

Selection and Implementation of XAI and Countermeasures

As mentioned, implementing XAI methods introduces further risks and attack vectors based on the information obtained by these methods. Therefore, mitigating these threats requires balancing transparency with security. Providing too much detail in explanations can expose the model to various attacks, while insufficient transparency can undermine XAI's purpose, which is to build trust and understanding. Striking the right balance involves carefully selecting suitable methods to provide necessary insights without disclosing sensitive information that could be exploited. The XAI method should be chosen strictly based on the determined requirements for the type and extent of explainability needed. The explanations themselves can be a target for tampering. Ensuring the integrity and authenticity of explanations through cryptographic techniques like digital signatures can help verify that the explanations have not been altered and are legitimate [95]. Furthermore, explainability methods can inadvertently expose sensitive aspects of the "explained" AI model, such as proprietary algorithms or business logic. Role-based access controls can ensure that only authorized personnel view detailed explanations, protecting intellectual property and sensitive information. Employing robust adversarial training techniques can also help the model resist adversarial attacks based on or enhanced by XAI.

Testing of XAI

Initial testing should focus on verifying the introduced requirements for XAI. Test cases shall ensure that only necessary information is published, but that explanations are still effective. Therefore, testing XAI has to be conducted especially from the viewpoint of the target group of the explanations. Additionally, vulnerability testing shall be conducted with regard to XAI-related attacks, e.g., as listed in Table 1 and Table 2. Research for state-of-the-art attacks should always be carried out and relevant attacks are to be incorporated in the vulnerability testing activities.

5 Conclusion

As XAI methods move from research to practical applications, concerns about malicious use and adversarial attacks have increased. This work provides a comprehensive overview of security and robustness issues in XAI, categorizing research on adversarial attacks targeting ML explanations and the exploitation of explainability to enhance attacks on AI systems. Most studies focus on predictive models using imaging and tabular datasets with backpropagation and perturbation-based XAI techniques. Further research is needed on adversarial attacks in other data modalities, such as language, graphs, time series, multimodal systems, and explanations for reinforcement learning agents and transformer-based generative AI like large language models. Additionally, this review highlights the need to evaluate vulnerabilities in intrinsically explainable ML architectures, such as decision trees and attention-based neural networks, and how their explanations could enhance attacks.

Practically, integrating XAI into AI systems requires awareness of its dual-edged nature. While XAI offers benefits like compliance, user trust, and system debugging, it also introduces security risks that must be mitigated to ensure the safe development and deployment. Therefore, the integration of XAI into AI systems requires a thorough assessment of the potential risks and corresponding countermeasures. XAI methods should be selected carefully to ensure explanations are informative without revealing sensitive information that could facilitate attacks on the AI system.

References

1 E. Abdukhamidov, M. Abuhamad, F. Juraev, E. Chan-Tin, and T. AbuHmed. AdvEdge: Optimizing Adversarial Perturbations Against Interpretable Deep Learning. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 13116 LNCS, pages 93–105, 2021. doi:10.1007/978-3-030-91434-9_9.

2 E. Abdukhamidov, F. Juraev, M. Abuhamad, and T. Abuhmed. Black-box and Target-specific Attack Against Interpretable Deep Learning Systems. In *ASIA CCS 2022 - Proceedings of the 2022 ACM Asia Conference on Computer and Communications Security*, pages 1216–1218, 2022. doi:10.1145/3488932.3527283.

3 Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018. doi:10.1109/ACCESS.2018.2870052.

4 U. Aïvodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, and A. Tapp. Fairwashing: The risk of rationalization. In *36th International Conference on Machine Learning, ICML 2019*, volume 2019-June, pages 240–252, 2019.

5 Ulrich Aïvodji, Alexandre Bolot, and Sébastien Gambs. Model extraction from counterfactual explanations, September 2020. arXiv:2009.01884, doi:10.48550/arXiv.2009.01884.

6 H. Ali, M.S. Khan, A. Al-Fuqaha, and J. Qadir. Tamp-X: Attacking explainable natural language classifiers through tampered activations. *Computers and Security*, 120, 2022. doi:10.1016/j.cose.2022.102791.

7 Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*, 2018. arXiv:1804.07998.

8 Abderrahmen Amich and Birhanu Eshete. EG-Booster: Explanation-Guided Booster of ML Evasion Attacks. In *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy*, CODASPY '22, pages 16–28, New York, NY, USA, April 2022. Association for Computing Machinery. doi:10.1145/3508398.3511510.

9 C.J. Anders, P. Pasliev, A.-K. Dombrowski, K.-R. Muller, and P. Kessel. Fairwashing explanations with off-manifold detergent. In *37th International Conference on Machine Learning, ICML 2020*, volume PartF168147-1, pages 291–300, 2020.

10 Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*, 2019.

11 Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

12 Vincent Ballet, Xavier Renard, Jonathan Aigrain, Thibault Laugel, Pascal Frossard, and Marcin Detyniecki. Imperceptible adversarial attacks on tabular data. *arXiv preprint arXiv:1911.03274*, 2019. `arXiv:1911.03274`.

13 H. Baniecki and P. Biecek. Manipulating SHAP via Adversarial Data Perturbations (Student Abstract). In *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, volume 36, pages 12907–12908, 2022.

14 H. Baniecki, W. Kretowicz, and P. Biecek. Fooling Partial Dependence via Data Poisoning. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 13715 LNAI, pages 121–136, 2023. `doi:10.1007/978-3-031-26409-2_8`.

15 Hubert Baniecki and Przemyslaw Biecek. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, page 102303, 2024. `doi:10.1016/J.INFFUS.2024.102303`.

16 Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, pages 387–402. Springer, 2013. `doi:10.1007/978-3-642-40994-3_25`.

17 Clara Bove, Thibault Laugel, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. Why do explanations fail? a typology and discussion on failures in xai. *arXiv preprint arXiv:2405.13474*, 2024. `doi:10.48550/arXiv.2405.13474`.

18 D. Brown and H. Kvinge. Making Corgis Important for Honeycomb Classification: Adversarial Attacks on Concept-based Explainability Tools. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2023-June, pages 620–627, 2023. `doi:10.1109/CVPRW59228.2023.00069`.

19 Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. `arXiv:1712.09665`.

20 Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. `doi:10.1109/SP.2017.49`.

21 Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

22 Y. Chai, R. Liang, S. Samtani, H. Zhu, M. Wang, Y. Liu, and Y. Jiang. Additive Feature Attribution Explainable Methods to Craft Adversarial Attacks for Text Classification and Text Regression. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14, 2023. `doi:10.1109/TKDE.2023.3270581`.

23 Fabien Charmet, Harry Chandra Tanuwidjaja, Solayman Ayoubi, Pierre-François Gimenez, Yufei Han, Houda Jmila, Gregory Blanc, Takeshi Takahashi, and Zonghua Zhang. Explainable artificial intelligence for cybersecurity: a literature survey. *Annals of Telecommunications*, 77(11):789–812, 2022. `doi:10.1007/S12243-022-00926-7`.

24 L. Chen, N. Yan, B. Zhang, Z. Wang, Y. Wen, and Y. Hu. A General Backdoor Attack to Graph Neural Networks Based on Explanation Method. In *Proceedings - 2022 IEEE 21st International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom 2022*, pages 759–768, 2022. `doi:10.1109/TrustCom56396.2022.00107`.

25 Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. `arXiv:1712.05526`.

26 Yu-Neng Chuang, Guanchu Wang, Fan Yang, Zirui Liu, Xuanting Cai, Mengnan Du, and Xia Hu. Efficient xai techniques: A taxonomic survey. *arXiv preprint arXiv:2302.03225*, 2023.

27 Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Computing Surveys*, 55(13s):1–39, 2023. doi:10.1145/3585385.

28 Microsoft Corporation. Microsoft responsible ai standard, v2, 2022. accessed 29 July 2024. URL: https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf.

29 E.J. De Aguiar, M.V.L. Costa, C. Traina, and A.J.M. Traina. Assessing Vulnerabilities of Deep Learning Explainability in Medical Image Analysis under Adversarial Settings. In *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, volume 2023-June, pages 13–16, 2023. doi:10.1109/CBMS58004.2023.00184.

30 Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods. In Huáscar Espinoza, José Hernández-Orallo, Xin Cynthia Chen, Seán S. ÓhÉigeartaigh, Xiaowei Huang, Mauricio Castillo-Effen, Richard Mallah, and John A. McDermid, editors, *Proceedings of the Workshop on Artificial Intelligence Safety, Co-Located with 34th AAAI Conference on Artificial Intelligence, SafeAI@AAAI 2020, New York City, NY, USA, February 7, 2020*, volume 2560 of *CEUR Workshop Proceedings*, pages 63–73. CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2560/paper8.pdf.

31 Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

32 V. Duddu and A. Boutet. Inferring Sensitive Attributes from Model Explanations. In *International Conference on Information and Knowledge Management, Proceedings*, pages 416–425, 2022. doi:10.1145/3511808.3557362.

33 Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Sok: Taming the triangle–on the interplays between fairness, interpretability and privacy in machine learning. *arXiv preprint arXiv:2312.16191*, 2023. doi:10.48550/arXiv.2312.16191.

34 A. Galli, S. Marrone, V. Moscato, and C. Sansone. Reliability of eXplainable Artificial Intelligence in Adversarial Perturbation Scenarios. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12663 LNCS, pages 243–256, 2021. doi:10.1007/978-3-030-68796-0_18.

35 A. Ghorbani, A. Abid, and J. Zou. Interpretation of neural networks is fragile. In *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pages 3681–3688, 2019.

36 Sofie Goethals, Kenneth Sörensen, and David Martens. The Privacy Issue of Counterfactual Explanations: Explanation Linkage Attacks. *ACM Transactions on Intelligent Systems and Technology*, 14(5):83:1–83:24, August 2023. doi:10.1145/3608482.

37 Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

38 Jan Philip Göpfert, Heiko Wersing, and Barbara Hammer. Recovering localized adversarial attacks. In *Artificial Neural Networks and Machine Learning–ICANN 2019: Theoretical Neural Computation: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part I 28*, pages 302–311. Springer, 2019. doi:10.1007/978-3-030-30487-4_24.

39 Shixiang Gu and Luca Rigazio. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

40 Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018. doi:10.1145/3236009.

41 S. Guo, S. Geng, T. Xiang, H. Liu, and R. Hou. ELAA: An efficient local adversarial attack using model interpreters. *International Journal of Intelligent Systems*, 37(12):10598–10620, 2022. doi:10.1002/int.22680.

42 S.S. Hada, M.Á. Carreira-Perpiñán, and A. Zharmagambetov. Sparse oblique decision trees: A tool to understand and manipulate neural net features. *Data Mining and Knowledge Discovery*, 2023. doi:10.1007/s10618-022-00892-7.

43 Lena Heidemann, Maureen Monnet, and Karsten Roscher. Concept correlation and its effects on concept-based models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4780–4788, 2023.

44 J. Heo, S. Joo, and T. Moon. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

45 K. Hoedt, V. Praher, A. Flexer, and G. Widmer. Constructing adversarial examples to investigate the plausibility of explanations in deep audio and image classifiers. *Neural Computing and Applications*, 35(14):10011–10029, 2023. doi:10.1007/s00521-022-07918-7.

46 Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek. Explainable ai methods-a brief overview. In *International workshop on extending explainable AI beyond deep models and classifiers*, pages 13–38. Springer, 2022.

47 Weronika Hryniewska, Przemysław Bombiński, Patryk Szatkowski, Paulina Tomaszewska, Artur Przelaskowski, and Przemysław Biecek. Checklist for responsible deep learning modeling of medical images based on covid-19 detection studies. *Pattern Recognition*, 118:108035, 2021. doi:10.1016/J.PATCOG.2021.108035.

48 Q. Huang, L. Chiang, M. Chiu, and H. Sun. Focus-Shifting Attack: An Adversarial Attack That Retains Saliency Map Information and Manipulates Model Explanations. *IEEE Transactions on Reliability*, pages 1–12, 2023. doi:10.1109/TR.2023.3303923.

49 International Standardization Organization (ISO). Iso/iec 22989:2022 artificial intelligence concepts and terminology, 2022.

50 A. Ivankay, I. Girardi, C. Marchiori, and P. Frossard. FOOLING EXPLANATIONS IN TEXT CLASSIFIERS. In *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.

51 Rahul Iyer, Yuezhang Li, Huao Li, Michael Lewis, Ramitha Sundar, and Katia Sycara. Transparency and explanation in deep reinforcement learning neural networks. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 144–150, 2018. doi:10.1145/3278721.3278776.

52 José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020. doi:10.1038/S42256-020-00236-4.

53 H. Jing, C. Meng, X. He, and W. Wei. Black Box Explanation Guided Decision-Based Adversarial Attacks. In *2019 IEEE 5th International Conference on Computer and Communications, ICCC 2019*, pages 1592–1596, 2019. doi:10.1109/ICCC47050.2019.9064243.

54 Margot E Kaminski. The right to explanation, explained (june 15, 2018). university of colorado law legal studies research paper no. 18-24. *Berkeley Technology Law Journal*, 34(1), 2019.

55 Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (Un)reliability of Saliency Methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 267–280. Springer International Publishing, Cham, 2019. doi:10.1007/978-3-030-28954-6_14.

56 Zico Kolter and Aleksander Madry. Adversarial robustness: Theory and practice. *Tutorial at NeurIPS*, page 3, 2018.

57 Satyapriya Krishna, Jiaqi Ma, and Himabindu Lakkaraju. Towards bridging the gaps between the right to explanation and the right to be forgotten. In *International Conference on Machine Learning*, pages 17808–17826. PMLR, 2023. URL: `https://proceedings.mlr.press/v202/krishna23a.html`.

58 A. Kuppa and N.-A. Le-Khac. Black Box Attacks on Explainable Artificial Intelligence(XAI) methods in Cyber Security. In *Proceedings of the International Joint Conference on Neural Networks*, 2020. `doi:10.1109/IJCNN48605.2020.9206780`.

59 Gabriel Laberge, Ulrich Aïvodji, Satoshi Hara, Mario Marchand, and Foutse Khomh. Fool SHAP with Stealthily Biased Sampling. In *International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, May 2023.

60 H. Lakkaraju and O. Bastani. "how do i fool you?": Manipulating user trust via misleading black box explanations. In *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85, 2020. `doi:10.1145/3375627.3375833`.

61 Thi-Thu-Huong Le, Hyoeun Kang, and Howon Kim. Robust Adversarial Attack Against Explainable Deep Classification Models Based on Adversarial Images With Different Patch Sizes and Perturbation Ratios. *IEEE Access*, 9:133049–133061, 2021. `doi:10.1109/ACCESS.2021.3115764`.

62 Erwan Le Merrer and Gilles Trédan. Remote explainability faces the bouncer problem. *Nature Machine Intelligence*, 2(9):529–539, September 2020. `doi:10.1038/s42256-020-0216-z`.

63 Yiqiao Li, Sunny Verma, Shuiqiao Yang, Jianlong Zhou, and Fang Chen. Are graph neural network explainers robust to graph noises? In *Australasian Joint Conference on Artificial Intelligence*, pages 161–174. Springer, 2022. `doi:10.1007/978-3-031-22695-3_12`.

64 Haohan Liu, Xingquan Zuo, Hai Huang, and Xing Wan. Saliency map-based local white-box adversarial attack against deep neural networks. In *CAAI International Conference on Artificial Intelligence*, pages 3–14. Springer, 2022. `doi:10.1007/978-3-031-20500-2_1`.

65 Mingting Liu, Xiaozhang Liu, Anli Yan, Yuan Qi, and Wei Li. Explanation-Guided Minimum Adversarial Attack. In Yuan Xu, Hongyang Yan, Huang Teng, Jun Cai, and Jin Li, editors, *Machine Learning for Cyber Security*, Lecture Notes in Computer Science, pages 257–270, Cham, 2023. Springer Nature Switzerland. `doi:10.1007/978-3-031-20096-0_20`.

66 Ninghao Liu, Mengnan Du, Ruocheng Guo, Huan Liu, and Xia Hu. Adversarial attacks and defenses: An interpretation perspective. *ACM SIGKDD Explorations Newsletter*, 23(1):86–99, 2021. `doi:10.1145/3468507.3468519`.

67 Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

68 Daniel D Lundstrom, Tianjian Huang, and Meisam Razaviyayn. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In *International Conference on Machine Learning*, pages 14485–14508. PMLR, 2022. URL: `https://proceedings.mlr.press/v162/lundstrom22a.html`.

69 X. Luo, Y. Jiang, and X. Xiao. Feature Inference Attack on Shapley Values. In *Proceedings of the ACM Conference on Computer and Communications Security*, pages 2233–2247, 2022. `doi:10.1145/3548606.3560573`.

70 Gabriel Resende Machado, Eugênio Silva, and Ronaldo Ribeiro Goldschmidt. Adversarial machine learning in image classification: A survey toward the defender's perspective. *ACM Computing Surveys (CSUR)*, 55(1):1–38, 2021.

71 Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. `arXiv:1706.06083`.

72 Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. Model Reconstruction from Model Explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 1–9, New York, NY, USA, January 2019. Association for Computing Machinery. `doi:10.1145/3287560.3287562`.

73 Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. A survey on the robustness of feature importance and counterfactual explanations. *arXiv preprint arXiv:2111.00358*, 2021. `arXiv:2111.00358`.

74 Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Thanh Toan Nguyen, Phi Le Nguyen, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of privacy-preserving model explanations: Privacy risks, attacks, and countermeasures. *arXiv preprint arXiv:2404.00673*, 2024. `doi:10.48550/arXiv.2404.00673`.

75 Truc Nguyen, Phung Lai, Hai Phan, and My T Thai. Xrand: Differentially private defense against explanation-guided attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11873–11881, 2023. `doi:10.1609/AAAI.V37I10.26401`.

76 Maximilian Noppel, Lukas Peter, and Christian Wressnegger. Disguising Attacks with Explanation-Aware Backdoors. In *2023 IEEE Symposium on Security and Privacy (SP)*, page 664, 2023. `doi:10.1109/SP46215.2023.10179308`.

77 Maximilian Noppel and Christian Wressnegger. Sok: Explainable machine learning in adversarial environments. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 21–21. IEEE Computer Society, 2023.

78 Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. I know what you trained last summer: A survey on stealing machine learning models and defences. *ACM Computing Surveys*, 55(14s):1–41, 2023. `doi:10.1145/3595292`.

79 M.A. Pandya, P.C. Siddalingaswamy, and S. Singh. Explainability of Image Classifiers for Targeted Adversarial Attack. In *INDICON 2022 - 2022 IEEE 19th India Council International Conference*, 2022. `doi:10.1109/INDICON56171.2022.10039871`.

80 Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. Sok: Security and privacy in machine learning. In *2018 IEEE European symposium on security and privacy (EuroS&P)*, pages 399–414. IEEE, 2018. `doi:10.1109/EUROSP.2018.00035`.

81 Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016. `doi:10.1109/SP.2016.41`.

82 The European Parliament and The Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016.

83 The European Parliament and The Council of the European Union. Artificial intelligence act, 2024. accessed 29 July 2024. URL: `https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf`.

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

85 H. Rasaee and H. Rivaz. Explainable AI and susceptibility to adversarial attacks: A case study in classification of breast ultrasound images. In *IEEE International Ultrasonics Symposium, IUS*, 2021. `doi:10.1109/IUS52206.2021.9593490`.

86 J. Renkhoff, W. Tan, A. Velasquez, W.Y. Wang, Y. Liu, J. Wang, S. Niu, L.B. Fazlic, G. Dartmann, and H. Song. Exploring Adversarial Attacks on Neural Networks: An Explainable Approach. In *Conference Proceedings of the IEEE International Performance, Computing, and Communications Conference*, volume 2022-November, pages 41–42, 2022. `doi:10.1109/IPCCC55026.2022.9894322`.

87 Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

88 Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, 2023. `doi:10.1145/3624010`.

89 Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021. doi:10.1145/3453158.

90 Barbara Rychalska, Dominika Basaj, Alicja Gosiewska, and Przemysław Biecek. Models in the wild: On corruption robustness of neural nlp systems. In *Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part III 26*, pages 235–247. Springer, 2019. doi:10.1007/978-3-030-36718-3_20.

91 Maresa Schröder, Alireza Zamanian, and Narges Ahmidi. Post-hoc saliency methods fail to capture latent feature importance in time series data. In *International Workshop on Trustworthy Machine Learning for Healthcare*, pages 106–121. Springer, 2023. doi:10.1007/978-3-031-39539-0_10.

92 Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pages 1–59, 2023.

93 Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. doi:10.1109/ICCV.2017.74.

94 G. Severi, J. Meyer, S. Coull, and A. Oprea. Explanation-guided backdoor poisoning attacks against malware classifiers. In *Proceedings of the 30th USENIX Security Symposium*, pages 1487–1504, 2021.

95 Rucha Shinde, Shruti Patil, Ketan Kotecha, Vidyasagar Potdar, Ganeshsree Selvachandran, and Ajith Abraham. Securing ai-based healthcare systems using blockchain technology: A state-of-the-art systematic literature review and future research directions. *Transactions on Emerging Telecommunications Technologies*, 2024.

96 R. Shokri, M. Strobel, and Y. Zick. On the Privacy Risks of Model Explanations. In *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 231–241, 2021. doi:10.1145/3461702.3462533.

97 N. Si, H. Chang, and Y. Li. A Simple and Effective Method to Defend Against Saliency Map Attack. In *ACM International Conference Proceeding Series*, 2021. doi:10.1145/3474198.3478141.

98 Sanchit Sinha, Hanjie Chen, Arshdeep Sekhon, Yangfeng Ji, and Yanjun Qi. Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 420–434, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.blackboxnlp-1.33.

99 D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020. doi:10.1145/3375627.3375830.

100 D. Slack, S. Hilgard, H. Lakkaraju, and S. Singh. Counterfactual Explanations Can Be Manipulated. In *Advances in Neural Information Processing Systems*, volume 1, pages 62–75, 2021.

101 Qianqian Song, Xiangwei Kong, and Ziming Wang. Fooling Neural Network Interpretations: Adversarial Noise to Attack Images. In Lu Fang, Yiran Chen, Guangtao Zhai, Jane Wang, Ruiping Wang, and Weisheng Dong, editors, *Artificial Intelligence*, Lecture Notes in Computer Science, pages 39–51, Cham, 2021. Springer International Publishing. doi:10.1007/978-3-030-93049-3_4.

102 Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. doi:10.1109/TEVC.2019.2890858.

103 A. Subramanya, V. Pillai, and H. Pirsiavash. Fooling network interpretation in image classification. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-October, pages 2020–2029, 2019. doi:10.1109/ICCV.2019.00211.

104 T. Viering, Ziqi Wang, M. Loog, and E. Eisemann. How to Manipulate CNNs to Make Them Lie: The GradCAM Case. *ArXiv*, July 2019.

105 Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

106 Y. Wang, H. Qian, and C. Miao. DualCF: Efficient Model Extraction Attack from Counterfactual Explanations. In *ACM International Conference Proceeding Series*, pages 1318–1329, 2022. doi:10.1145/3531146.3533188.

107 J. Xu and Q. Du. Adversarial attacks on text classification models using layer-wise relevance propagation. *International Journal of Intelligent Systems*, 35(9):1397–1415, 2020. doi:10.1002/int.22260.

108 J. Xu, M. Xue, and S. Picek. Explainability-based Backdoor Attacks against Graph Neural Networks. In *WiseML 2021 - Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning*, pages 31–36, 2021. doi:10.1145/3468218.3469046.

109 A. Yan, R. Hou, X. Liu, H. Yan, T. Huang, and X. Wang. Towards explainable model extraction attacks. *International Journal of Intelligent Systems*, 37(11):9936–9956, 2022. doi:10.1002/int.23022.

110 A. Yan, R. Hou, H. Yan, and X. Liu. Explanation-based data-free model extraction attacks. *World Wide Web*, 26(5):3081–3092, 2023. doi:10.1007/s11280-023-01150-6.

111 A. Yan, T. Huang, L. Ke, X. Liu, Q. Chen, and C. Dong. Explanation leaks: Explanation-guided model extraction attacks. *Information Sciences*, 632:269–284, 2023. doi:10.1016/j.ins.2023.03.020.

112 Y. Zhan, B. Zheng, Q. Wang, N. Mou, B. Guo, Q. Li, C. Shen, and C. Wang. Towards Black-Box Adversarial Attacks on Interpretable Deep Learning Systems. In *Proceedings - IEEE International Conference on Multimedia and Expo*, volume 2022-July, 2022. doi:10.1109/ICME52920.2022.9859856.

113 H. Zhang, J. Gao, and L. Su. Data Poisoning Attacks against Outcome Interpretations of Predictive Models. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2165–2173, 2021. doi:10.1145/3447548.3467405.

114 X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, and T. Wang. Interpretable deep learning under fire. In *Proceedings of the 29th USENIX Security Symposium*, pages 1659–1676, 2020.

115 Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021. doi:10.1109/TETCI.2021.3100641.

116 Zhibo Zhang, Hussam Al Hamadi, Ernesto Damiani, Chan Yeob Yeun, and Fatma Taher. Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEE Access*, 10:93104–93139, 2022. doi:10.1109/ACCESS.2022.3204051.

117 X. Zhao, W. Zhang, X. Xiao, and B. Lim. Exploiting Explanations for Model Inversion Attacks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 662–672, 2021. doi:10.1109/ICCV48922.2021.00072.