

Survey on Explainability-Weaponising Adversarial Attack Vectors against Deep Neural Networks and Artificial Intelligence

Marek Pawlicki^{1,2}, Ryszard Choraś², Rafał Kozik^{1,2}, and Michał Choraś^{1,2}

¹ITTI, Sp. z o.o., Poznań, Poland

²Department of Telecommunications, Computer Science and Electrical Engineering, Bydgoszcz University of Science and Technology, Bydgoszcz, Poland

ABSTRACT

Adversarial machine learning has revealed the fragility of deep neural networks, while explainable artificial intelligence has been introduced to improve the transparency and trust of AI. It has recently been demonstrated, however, that xAI can be weaponised, enabling adversaries to amplify the effectiveness and efficiency of adversarial attacks. This paper presents the first systematic survey dedicated to xAI-weaponising adversarial attacks. The literature is synthesised across four adversarial goals: evasion, poisoning/backdoors, privacy/inference, and model extraction. A unified taxonomy is proposed that organises attack vectors according to adversarial goals, operational roles of xAI, and attacker capabilities. The bibliographic methodology follows PRISMA guidelines, with structured queries applied to IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and Google Scholar, complemented by snowballing. The date range was set to 2020-2025. The findings indicate that evasion attacks dominate current literature, while poisoning and extraction attacks remain comparatively underexplored. Open challenges and research directions are identified. This survey reframes xAI from a purely diagnostic tool to a security-critical interface and provides a foundation for principled defence.

Keywords: Explainability, Adversarial Attacks, Artificial Intelligence, Deep Neural Networks, Adversarial Machine Learning

1 INTRODUCTION

Adversarial machine learning (AdvML) has revealed that deep neural networks (DNNs) and other artificial intelligence (AI) models remain vulnerable to carefully crafted manipulations. In parallel, the emergence of explainable AI (xAI) has been widely promoted as a mechanism for improving transparency, accountability, and human trust in complex systems.

However, it has recently been demonstrated that the xAI signals, designed to enhance interpretability, can be exploited by adversaries to strengthen their positioning. As a result, xAI is transitioning from being viewed solely as a defensive or diagnostic tool to a dual-use interface that may be weaponised. To the best knowledge of the authors, despite scattered studies highlighting such risks, no comprehensive survey has previously consolidated this emerging body of work into a systematic taxonomy.

Thus, this study provides three major contributions.

1. the literature on adversarial attacks that exploit xAI as an offensive signal is synthesised
2. a unified taxonomy is proposed that systematises xAI-weaponising attacks
3. open challenges and future research directions are identified

Since this is an emerging niche, a regular PRISMA exhaustive search yielded very mixed results, with structured queries applied to IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and Google Scholar. Search terms combined variations of “*adversarial attack*”, “*explainable AI*”, “*xAI*”, “*explanation-guided attacks*”, “*counterfactual*”. Only a handful of papers matched the eligibility criteria. The initial set of works was expanded through backward and forward snowballing of references. Inclusion criteria required that papers explicitly describe attacks in which xAI signals were used to plan, guide, or validate adversarial actions against ML models. Exclusion criteria removed works that studied only attacks on explanations themselves or papers that discussed xAI from any perspective other than weaponisation.

The remainder of this paper is structured as follows. Section II discusses related surveys and situates this work within the broader AdvML and xAI landscape. Section III presents a detailed survey of xAI-

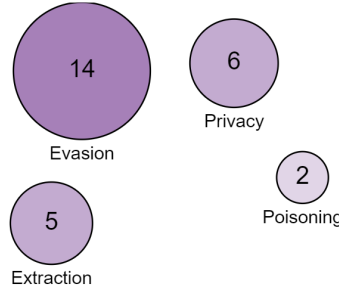


Figure 1. Distribution of surveyed papers across adversarial goals. The bubble area is proportional to the number of papers in each category: Evasion (14), Privacy/Inference (6), Model Extraction (5), and Poisoning/Backdoors (2).

enhanced adversarial attacks. Section IV introduces the proposed taxonomy for systematising these attacks. Section V highlights open challenges and future research directions. Section VI concludes the study.

2 RELATED WORKS

The notion of the relationship between xAI and AdvML has been approached from different angles. In (Baniecki and Biecek, 2024), the authors perform a survey of how explanations themselves can be attacked and defended, spanning both model explanations and fairness metrics. The paper provides a unified notation and taxonomy that formalise adversarial changes to data, models, and explanation functions, providing a common language for AdvML and xAI researchers. The authors perform a comprehensive synthesis of over 50 papers and clarify relationships across attack surfaces such as AdvML, data poisoning, model manipulation, backdoors, adversarial fairwashing, and trust manipulation. In (Vadillo et al., 2025), the authors provide a formalisation of AdvML in xAI by extending the notion of adversarial examples to scenarios where humans evaluate not just the model output but also its accompanying explanation, organising and unifying diverse attack paradigms previously studied in isolation. The focus is on attacks which preserve the label but alter the xAI. In (Tam Nguyen et al., 2024), the authors deliver a survey of privacy risks arising specifically from model explanations. The paper proposes a taxonomy that jointly organises privacy attack families like membership inference, re-identification, model inversion, attribute inference, and model extraction, together with the explanation mechanisms they target.

While prior surveys study attacks on xAI and privacy risks of xAI, to the best knowledge of the authors, this work is the first to systematically review attacks that use xAI as an offensive resource ('xAI-enhanced attacks'). This study analyses works which focus on how SHAP/LIME/Saliency Maps/Counterfactuals and other explainers can be utilised to facilitate attacks across evasion, poisoning/backdoor placement, model extraction, and privacy/inversion. This considers xAI as a signal, not as a target. This complements the scope of other surveys and positions our work as, to our best knowledge, the first comprehensive systematisation of xAI-weaponising attack vectors.

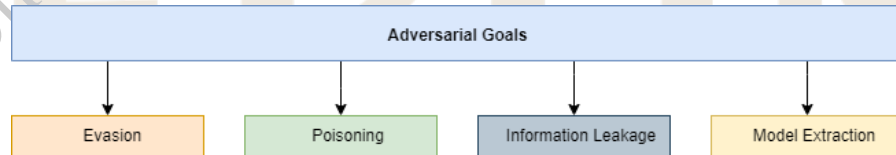


Figure 2. Overview of adversarial goals of xAI in ML, including evasion, poisoning, information leakage, and model extraction.

3 SURVEY OF XAI-ENHANCED ADVERSARIAL ATTACKS

This section presents categorised surveyed works on xAI-enhanced adversarial attacks.

3.1 xAI-enhanced Evasion Attacks

This subsection reviews works in which xAI is used to guide perturbation generation at test-time. Across the literature, xAI most commonly supports evasion through feature targeting, perturbation reduction, and query efficiency improvements.

In (Kumagai et al., 2023), the authors use xAI techniques to mark the regions which carry a strong influence over the results of the output of a classifier, to then leverage those regions in the adversarial attack generation process. Compared to conventional AdvML techniques, the xAI-guided method showed higher success rates in deceiving classifiers. Additionally, the xAI-guided techniques are more interpretable than the standard adversarial attack generation methods, which emphasises a synergy between attack approaches and xAI.

In (Vu et al., 2024), the authors propose xSub, a black-box AdvML method that exploits xAI tools. It identifies the most important input features, then substitutes them with corresponding features from a "golden sample" of another class. This perturbation can deceive the classifier using a constant number of queries, often with minimal modifications. The method chooses a representative sample for each target class, which features are highly important according to xAI methods. Then, replace the top features of the original input with those from the golden sample. Two amplification hyperparameters control the strength of the substitution. This approach can work with only a small number of model+explainer queries. The approach can be enriched with dataset poisoning, where instead of a 'golden sample', a backdoor trigger is used.

In (Bayer et al., 2024), the overall idea is to identify incorrectly learned patterns within a model by applying attribution-based explanation methods, such as LIME or SHAP, on misclassified instances. These explanations highlight words that disproportionately influence erroneous predictions, which are then used to construct AdvML. To ensure semantic validity, the authors propose a filtering mechanism which excludes words strongly correlated with the correct class, or those that alter the semantics of the input. XAI-Attack does not require gradient access or soft-label outputs, making it suitable for realistic black-box scenarios.

In (Zhu et al., 2024), the authors present LimeAttack, a hard-label AdvML for text classification that leverages local xAI methods. Hard-label settings are both challenging and realistic, since an adversary can only query the victim model and observe discrete prediction labels, without access to gradients or confidence scores. Existing approaches typically rely on random word substitutions followed by heuristic optimization, which leads to high query costs and low-quality AdvML. LimeAttack addresses this limitation by adopting LIME to approximate word importance ranking, thus guiding perturbations toward the most influential tokens. Extensive experiments on seven NLP datasets and multiple architectures (CNN, LSTM, BERT) show that LimeAttack outperforms prior hard-label baselines. The authors also evaluate LimeAttack on large language models, including GPT-3 and ChatGPT, demonstrating that even state-of-the-art LLMs remain vulnerable to such attacks.

The authors of (Yan et al., 2024) propose MEAttack, an explanation-guided AdvML framework tailored for black-box settings with limited query budgets. The paper introduces a novel model-agnostic explanation approach that relies on training an ensemble of lightweight substitute models to approximate the decision process of the target model. By extracting and ranking important features from the substitute model, the attacker identifies which parts of the input space drive the outputs of the model. Perturbations are then applied selectively to these features, significantly narrowing the perturbation scope. The local model is iteratively aligned with the target model. Experimental evaluation on MNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100 under both non-targeted and targeted settings demonstrates that MEAttack consistently surpasses baseline and state-of-the-art black-box attacks.

In (Okada et al., 2025), the authors extend their earlier work (Okada et al., 2024) on explanation-guided white-box attacks against deep learning-based network intrusion detection systems (NIDS) by presenting a more practical black-box variant. The proposed method leverages KernelSHAP to identify the most influential features contributing to detection decisions without requiring access to the internal structure of the targeted model. By focusing perturbations on only a few critical and relatively independent features, the attack resolves the non-invertibility problem of network traffic feature extraction, ensuring that manipulations in the feature space map to feasible transformations in the problem space. To demonstrate generalisability, the method was evaluated against multiple NIDS models trained on both the CIC-IDS2017 and TON_IoT datasets.

In (Kozik et al., 2024), an investigation is showcased on how explanations produced by xAI can be exploited by adversaries in the context of fake news detection. The authors demonstrate that an attacker with access to feature-level relevance scores, like word importance, can tweak fake news content to manipulate the output of the AI-based fake news detector. The experiments start with a standard AI classifier paired with an xAI mechanism. The attacker can query the detector with candidate text and receive explanations indicating which tokens have the most influence on the decision of the classifier. The attacker then modifies words, guided by the xAI feedback. The process is demonstrated over several rounds, showing a steady drop in fake-news score despite text meaning being preserved.

In 'EG-Booster: Explanation-Guided Booster of ML Evasion Attacks' (Amich and Eshete, 2022), the authors utilise feature-based explanations to guide the generation of AdvML, prioritising perturbations

which are most critical for decision-making, and avoiding those which have minimal effect. The approach is model-agnostic, supports both white-box and black-box threat models, and allows to apply a variety of distance metrics, making it both efficient and effective. Experiments validate the efficacy of EG-Booster on benchmark datasets (MNIST and CIFAR-10) and multiple architectures, including undefended neural networks and an adversarially trained ResNet. EG-Booster consistently improves evasion rates while requiring fewer perturbations than state-of-the-art baselines.

In (Liu et al., 2022b), the authors also provide a method which exploits feature importance information provided by xAI to sharply reduce the perturbation range necessary for successful evasion. Instead of spreading noise across the entire input, the attack focuses only on the small set of features deemed most critical by xAI, allowing for efficient, effective, minimal and imperceptible perturbations. Evaluations on CIFAR-10 confirm the efficiency of this strategy: the perturbation scope is up to 90% smaller than the Carlini & Wagner attack, yet the method maintains comparable success rates. The approach is effective in both white-box and black-box settings. In (Zhan et al., 2023), the proposed PSP-Mal framework enhances reinforcement learning-based malware evasion attacks by integrating Shapley value-driven priors into the training process. Instead of relying on sparse or noisy environment feedback, the method leverages Shapley values to quantify the expected influence of feature manipulations. This creates a Shapley prior that guides the agent toward more impactful operations. The authors of (Shi et al., 2024) introduce SHAPAttack, a multigranularity strategy for textual AdvML by combining perturbations at both the word and phrase level. This enlarges the perturbation space and improves the diversity of AdvML. SHAPAttack integrates a Shapley value-based constituent importance ranking that estimates the contribution of each word or phrase without relying on costly queries. This query-free guidance enables the attacker to focus perturbations on the most influential constituents, reducing query overhead while maintaining high attack strength.

In (Shu and Yan, 2023) the authors introduce a problem-space AdvML framework for Android malware detection that leverages local explanations to guide the search for evasive modifications. EAGLE uses explanation signals to determine which feature manipulations, both increases and decreases, are most likely to reduce classifier confidence. The framework is designed to be generic and adaptable, accommodating multiple types of count-based feature representations used in Android malware classifiers. EAGLE demonstrates high effectiveness in producing functional adversarial variants that successfully evade detection, as demonstrated on two Android malware datasets.

In (Pawlicki et al., 2024), the authors expose the security risks inherent in counterfactual xAI by showing how explanation mechanisms can be weaponised to mount AdvML. Specifically, the authors demonstrate that the Diverse Counterfactual Explanations algorithm, originally designed to improve transparency by providing alternative decision pathways, can be repurposed to systematically generate AdvML. These samples are tailored to flip classifier outputs, thereby compromising the integrity of ML-based detectors. The information leakage is showcased on Network Intrusion Detection data.

The paper (Chanda et al., 2025) introduces an xAI-driven AdvML framework for graph neural networks (GNNs) that exploits edge perturbations to degrade node classification performance. Unlike prior methods that treat all edges uniformly, the described method leverages GNN-specific xAI tools (GNExplainer and PGExplainer) to identify a subset of nodes and edges that hold the most importance in the decision-making process of the GNN, constructing an 'important subgraph.' Adversarial perturbations are then targeted within this subgraph. Comprehensive experiments on three benchmark citation datasets (Cora, CiteSeer, and PubMed) and across multiple GNN architectures (GCN, GAT, and GraphSAGE) validate the effectiveness of the proposed attack. The authors of (Liu et al., 2022a) propose a 'Saliency Map-based Local white-box Adversarial Attack method' (SMLAA) for images that also leverages xAI to constrain where perturbations are applied. The authors report that SMLAA maintains the high attack success rates typical of global gradient methods but with substantially less visible change. Similarly, in (Dai et al., 2023), a saliency-constrained framework for black-box AdvML that limits perturbations to a compact, visually salient region of the image is proposed. This restriction is designed to be plug-and-play with many existing black-box methods. The described method is a black-box algorithm that further refines perturbations within the salient region to optimise perceptual stealth. Experiments show that the approach obtains a higher 'true success rate' and improved query efficiency. By the same token, (Zhang et al., 2024) employs saliency maps to localise and solve for effective perturbations, obtaining high attack success within small perturbation budgets.

3.2 xAI-enhanced Poisoning Attacks

The reviewed poisoning attacks share a common theme: xAI is used to identify low-importance or decision-relevant regions where triggers can be inserted more stealthily, improving their persistence and impact.

The authors of (Chen et al., 2023) introduce Horn-Clause Attacks to Recommender Systems (H-CARS), a method for poisoning recommendation models through counterfactual explanations. The main idea of the

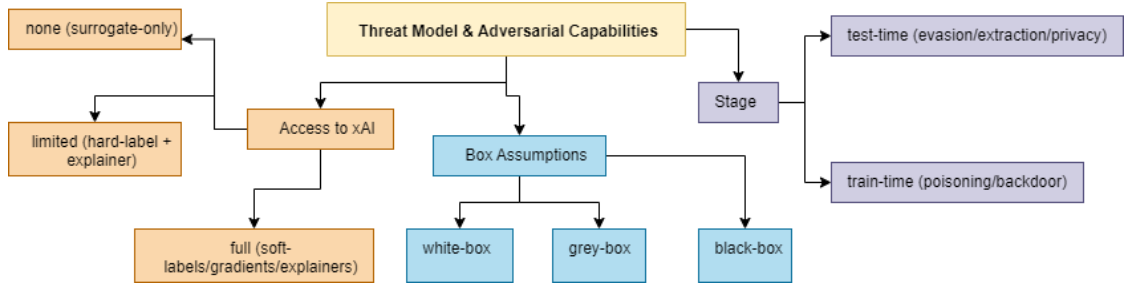


Figure 3. Framework outlining adversarial capabilities in xAI, structured by access level, box assumptions, and stage of interaction (train-time vs. test-time).

approach is as follows: The attacker extracts Counterfactuals from a target ML-based recommendation model, which will serve as observation data to build a logical-reasoning-based surrogate model that approximates the behaviour of the recommender. The attacker then generates synthetic user profiles and interactions tailored to mislead this surrogate to steer the real recommender system toward malicious recommendations. The method is tested on two real-world datasets. In (Xu et al., 2021), an explainability-guided framework for orchestrating backdoor attacks on GNNs is introduced. For graph classification, the authors leverage GNNExplainer to score node importances and then replace subgraphs at the least influential positions, aiming to maximise attack effectiveness while remaining hard to detect. A new feature-trigger backdoor is proposed. The study is positioned as the first to systematically examine how explainability can drive the selection of trigger injection positions in GNN backdoors across both graph- and node-level tasks.

3.3 xAI-enhanced Attacks on Privacy

A recurring finding in privacy attacks is that explanations leak distributional information that is strongly correlated with training membership or sensitive attributes, providing additional inference signals beyond model outputs alone.

The paper "On the privacy risks of model explanations" (Shokri et al., 2021) explores how providing explainability in ML can compromise privacy. The authors examine whether xAI, particularly feature-based methods such as integrated gradients and perturbation-based techniques, leak sensitive information about training data. Using membership inference attacks as their analytical framework, they show that backpropagation-based explanations can expose substantial information about whether a given datapoint was part of the training set. This leakage arises because explanations encode statistical signals about decision boundaries, making it easier for adversaries to distinguish training from non-training examples. The study demonstrates that the risk is especially pronounced in high-dimensional datasets where variance in explanation vectors aligns with membership information. The authors further compare different explanation methods to highlight a trade-off between transparency and privacy. Perturbation-based methods like SmoothGrad or LIME show greater resilience against inference attacks, but this comes at the expense of explanation fidelity, as they rely on out-of-distribution perturbations that reduce interpretability quality.

In (Liu et al., 2024a), the authors point out that while xAI helps humans understand models, it can unintentionally leak private information about the data used to train those models. The paper investigates how xAI methods, such as Grad-CAM, LIME, SHAP, SmoothGrad can increase the risk of membership inference attacks. The authors develop a membership inference attack that exploits information from the model output and from the explanations, which change when input data is perturbed. The authors note that data that was used for training tends to be more sensitive when perturbations are introduced in the important features identified by xAI, the model's confidence drops more compared to non-member data. This difference can be measured and used to differentiate members from non-members. The authors test their method on several datasets and model types, using seven explanation methods, and find non-trivial privacy leaks across the board.

Liu et al. (Liu et al., 2024b) investigate how explanation-guided attacks can be utilised. Their study shows that model explanations, when provided repeatedly, can be aggregated and amplified to significantly improve the success of inversion and membership inference attacks, effectively lowering the barrier for adversaries. The study highlights that xAI reveals semantic information that persists across queries. This is especially important in user-facing systems where explanations are generated dynamically.

The authors of (Zhao et al., 2021a) investigate a hidden privacy risk in AI: model explanations obtained via xAI methods, meant to help users understand AI decision-making process, can be exploited by attackers to reconstruct private input data even more effectively than using the output of the AI model alone. The risk

is highest when explanations provide more spatial or detailed information. The authors propose new attack models using U-Net and Flatten architectures to ingest both model output and explanations to reconstruct images. The experiments show that even if the target model does not provide explanations, but a similar surrogate model does, the attacker can still use explanations from the surrogate to boost their attack on the target. The authors prove that in some cases the xAI-augmented attacks are around 40 times more effective.

The authors of (Zhao et al., 2021b) extend the discussion by demonstrating that explanations can serve not only as signals for membership inference, but also as a basis for reconstructing sensitive attributes of training data. By leveraging feature attribution vectors, the authors show that adversaries can effectively mount model inversion attacks that recover class-representative images and, in some cases, even approximate individual training examples.

In (Luo et al., 2022), a study of the privacy risks of using Shapley values for model interpretability in machine learning is presented, especially when these explanations are provided as a service. The study shows that Shapley value explanations can leak private input features through feature inference attacks. In the presented method, two approaches are undertaken, one where the attacker has access to an auxiliary dataset, which has a similar distribution to the target data, and black-box query access to the model along with the model’s xAI, and one where the attacker only has access to the model decisions and xAI. Attacks were tested on Google Cloud, Microsoft Azure, and IBM AIX360 with multiple datasets and models. With just 100 queries, Adversary 1 could reconstruct private features with about 10% average error, and Adversary 2 could still reconstruct at least 30% of features with about 14% average error. Thus, sharing Shapley-based explanations, even without the raw input, can leak sensitive data.

3.4 xAI-enhanced Model Extraction

In model extraction settings, xAI commonly acts as a high-information auxiliary output, reducing query budget by revealing gradient-like directions or counterfactual decision boundaries.

The authors of (Aïvodji et al., 2020) point out that counterfactual explanations can leak important information about the underlying model, and by querying the model for both output labels and counterfactual explanations, an attacker can extract a copy of the original model. The attacker sends an input and gets both the class label and the counterfactual explanation from the model. By collecting many such input-output pairs, including the counterfactuals, the attacker gains an immense amount of information about how the model works. Then, using this dataset, the attacker can train their own model. The approach was tested on real datasets like Adult Income, COMPAS, and Default Credit. With only a few hundred to a thousand queries, the attacker could build a surrogate that performs almost as well as the original model. With access to multiple diverse counterfactuals per input, the attack becomes even more potent.

The paper (Yan et al., 2023b) investigates privacy and security risks for ML models that offer explanations, showing that xAI can help attackers extract the model, making it much easier to copy a proprietary model even if the attackers only have black-box access. In extraction attacks, the attackers query a black-box model with various inputs, collect the outputs and then train a substitute model to mimic the original. This is a privacy and intellectual property concern, especially for companies deploying valuable models behind APIs. The push for transparency in AI reveals information which could be utilised in extraction attempts. The authors propose a framework called XAI-aware Model Extraction Attacks (XaMEA), which collects data just like in traditional model extraction, and for each sample, the attacker gets both the output label and the explanation, and then uses a transfer learning approach to combine information from the explanations and outputs, training a model that more closely mimics the original. Some explanation methods, like Grad-CAM, are particularly hazardous by providing more useful information to attackers.

In (Oksuz et al., 2024), another way of boosting extraction attacks with xAI is presented. AUTOLYCUS uses the xAI tools to make educated guesses about the decision boundaries of the model. Instead of sending vast amounts of random inputs, AUTOLYCUS tweaks the most important features, as selected by xAI methods, rendering the attacks much more efficient. The authors demonstrate that with AUTOLYCUS, one can extract a model almost exactly, using significantly fewer queries than older methods like Equation-Solving or Path-Finding. AUTOLYCUS sends a query to the model, obtains the label and the xAI output, and perturbs the most important features one at a time to gauge the response of the model. By iterating on this process, the method maps out the decision boundaries of the model, and allows to train an almost-exact copy of the probed model with the produced set.

The paper “MEGEX: Data-Free Model Extraction Attack Against Gradient-Based Explainable AI” (Miura et al., 2024) introduces a threat model in which adversaries leverage gradient-based xAI to improve the efficiency and fidelity of data-free model extraction attacks. MEGEX directly leverages explanation outputs to obtain exact gradient information, making the approach a functional equivalent to white-box data-free knowledge distillation. MEGEX reduces the query budget compared to existing data-free attacks, with comparable performance.

Table 1. xAI-guided adversarial attacks surveyed (grouped by goal, then by domain).

Paper	Modality/Domain	xAI leveraged
Goal: Evasion		
(Kumagai et al., 2023)	Vision	Saliency/attribution
(Vu et al., 2024)	Vision	Feature importances
(Bayer et al., 2024)	NLP	LIME/SHAP
(Zhu et al., 2024)	NLP	LIME
(Yan et al., 2024)	Vision	Model-agnostic importances
(Okada et al., 2025)	NIDS	KernelSHAP
(Amich and Eshete, 2022)	Vision	Feature importances
(Liu et al., 2022b)	Vision	Saliency maps
(Liu et al., 2022a)	Vision	Saliency maps
(Dai et al., 2023)	Vision	Saliency maps
(Zhang et al., 2024)	Vision	Saliency maps
(Shu and Yan, 2023)	Android malware	Local explanations
(Pawlicki et al., 2024)	NIDS	Counterfactuals
(Shi et al., 2024)	NLP	Shapley values
Goal: Poisoning		
(Chen et al., 2023)	Recommender systems	Counterfactuals
(Xu et al., 2021)	GNNs	GNNExplainer
Goal: Privacy		
(Shokri et al., 2021)	Vision/NLP	Integrated Gradients/perturbation-based
(Liu et al., 2024a)	Vision/NLP	Grad-CAM/LIME/SHAP/SmoothGrad
(Liu et al., 2024b)	General	Aggregated explanations
(Zhao et al., 2021a)	Vision	Explanations (feature maps)
(Zhao et al., 2021b)	Vision/Tabular	Feature importances
(Luo et al., 2022)	Tabular	Shapley values
Goal: Model Extraction		
(Aïvodji et al., 2020)	Tabular	Counterfactuals
(Yan et al., 2023b)	Vision	Grad-CAM/LIME
(Oksuz et al., 2024)	General	Attribution-guided boundary probing
(Miura et al., 2024)	Vision	Gradient-based xAI
(Yan et al., 2023a)	Vision	Grad-CAM/LIME

In (Yan et al., 2023a), the authors develop Explanation-based data-free model extraction attacks (DMEAE). Unlike MEGEX, which exploits gradient-based explanation vectors to approximate white-box conditions, DMEAE generalises the attack surface by incorporating both model predictions and explanation tensors into the loss function for training surrogate models. DMEAE combines cross-entropy loss on predicted labels with mean square error on explanations, forcing the surrogate model to align not only with the victim’s outputs but also with its xAI patterns. The authors also explore variations in attack effectiveness across explanation modalities. Their findings suggest that model-specific techniques like Grad-CAM tend to leak more exploitable information than model-agnostic approaches such as LIME, owing to their finer localisation and closer coupling with internal gradients. Attempts to defend against DMEAE by perturbing explanation outputs had a negligible impact.

Figure 1 illustrates the distribution of papers across the four adversarial goals. Evasion dominates the landscape with the majority of studies (14 papers), followed by privacy-related inference attacks (6) and model extraction (5). Poisoning and backdoor attacks leveraging xAI remain relatively rare (2). This imbalance highlights how the community has primarily focused on test-time evasion, while train-time poisoning and backdooring with xAI support are still underexplored.

4 TAXONOMY

In this section, a taxonomy is proposed to systematise xAI-weaponising attacks to plan, steer, or validate adversarial actions against models or data. The taxonomy considers what the adversary ultimately aims to achieve and how xAI is obtained and used.

4.1 Adversarial Goals

1. **Evasion** (test-time misclassification): xAI guides where and how to perturb inputs to induce errors with minimal cost (Kumagai et al., 2023; Vu et al., 2024; Bayer et al., 2024; Zhu et al., 2024; Yan et al., 2024; Okada et al., 2025; Amich and Eshete, 2022; Liu et al., 2022b,a; Dai et al., 2023; Zhang et al., 2024; Shu and Yan, 2023; Pawlicki et al., 2024; Shi et al., 2024).
2. **Poisoning & Backdoors** (train-time compromise): xAI guides what to inject and where to place triggers to maximise effectiveness and stealth (Chen et al., 2023; Xu et al., 2021; Vu et al., 2024).
3. **Privacy/Inference** (information leakage) xAI vectors or trajectories are exploited to infer membership, reconstruct inputs, or infer hidden attributes/features (Shokri et al., 2021; Liu et al., 2024a,b; Zhao et al., 2021a,b; Luo et al., 2022).
4. **Model Extraction** (training surrogate models): xAI augments label outputs with gradients /attributions /counterfactual moves to fit high-fidelity substitutes with fewer queries (Aïvodji et al., 2020; Yan et al., 2023b; Oksuz et al., 2024; Miura et al., 2024; Yan et al., 2023a).

An overview of adversarial goals of xAI in ML, including evasion, poisoning, information leakage, and model extraction can be found in Figure 2.

4.2 Operational Roles of xAI (How Explanations Are Used)

xAI can support adversaries in several operational roles. **Targeting and Localisation** is used to identify influential features, regions, tokens or edges, allowing perturbations or trigger placement to be focused efficiently. **Search Guidance** assists optimisation, reinforcement-learning loops or black-box query selection. **Constraint Handling** enables attackers to maintain semantic validity or problem-space feasibility while perturbing only regions permitted by the domain. **Signal Aggregation** combines explanation vectors across multiple queries to strengthen leakage signals in privacy attacks or stabilise boundary estimates in extraction scenarios. **Synthetic Supervision** appears when explanations are treated as additional labels or learning signals for surrogates during model extraction. Finally, **Counterfactual Leveraging** uses decision-boundary-crossing samples to accelerate poisoning design or model replication. These roles often co-occur in practice, which proves that xAI commonly provides actionable guidance rather than merely interpretability.

The proposed taxonomy highlights that xAI serves not only as a transparency tool but also as an adversarial resource, enabling attackers to localise perturbations, optimise search strategies, enforce stealthy constraints, aggregate signals, or derive surrogate supervision. By systematising adversarial goals and operational roles, this taxonomy provides a structured foundation for analysing existing threats and guiding the development of defences against xAI-weaponising attacks.

4.3 Adversarial Capabilities

1. **Access to xAI**: none (surrogate-only), limited (hard-label + explainers), full (soft-labels /gradients/explainers).
2. **Box Assumptions**: white-box (gradients/params), grey-box (architecture/partial APIs), black-box (labels or hard-label).
3. **Stage**: test-time (evasion/extraction/privacy) vs. train-time (poisoning/backdoor).

The framework outlining adversarial capabilities in xAI, structured by access level, box assumptions, and stage of interaction is showcased in Figure 3.

5 OPEN CHALLENGES AND FUTURE DIRECTIONS

With the current state of the weaponisation of xAI for AdvML, the following open challenges can be identified. **Counterfactuals: dual-use, multiplicity, and controllability.** Counterfactual explanations enable efficient boundary discovery for extraction and poisoning (Aïvodji et al., 2020; Yan et al., 2023b; Chen et al., 2023; Pawlicki et al., 2024). **Privacy-preserving xAI with measurable utility-risk trade-offs.** Membership, inversion and feature inference via explanations is present across modalities (Shokri et al., 2021; Liu et al., 2024a; Zhao et al., 2021a,b; Liu et al., 2024b; Luo et al., 2022), and there is a lack of mechanisms to preserve privacy bounds. **Robust-by-design xAI that resists weaponisation.** Many xAI-guided attacks presume that importance attribution is faithful and stable (Amich and Eshete, 2022; Liu et al., 2022b,a; Dai et al., 2023; Zhang et al., 2024; Bayer et al., 2024), so future work may reduce attack-useful fidelity while maintaining interpretability. **Detection of xAI-guided query behaviour.** Black-box evasion

and extraction attacks often involve distinctive probing patterns (Vu et al., 2024; Zhu et al., 2024; Yan et al., 2024; Amich and Eshete, 2022), which may enable detection. **Model extraction mitigation.** Surrogate training benefits from explanation-aligned supervision (Yan et al., 2023b; Oksuz et al., 2024; Miura et al., 2024; Yan et al., 2023a), suggesting a need for tamper-resistant xAI outputs. **Forensics, logging, and accountability for xAI.** Given dual-use risks across privacy and extraction (Shokri et al., 2021; Yan et al., 2023b,a), robust logging of explanation requests could support auditing and post-incident analysis.

Collectively, these challenges argue for treating xAI as a regulated resource, rather than a good freely available to users. Future work should formalise risk-tiered xAI services, constrain xAI susceptibility to leakage, develop detection and response mechanisms for explanation-centric probing, evaluate under hostile conditions, evaluate xAI utility vs adversary payoff, and introduce auditable logging.

6 CONCLUSIONS

In this survey, the use of xAI as a resource in formulating attacks against AI was evaluated. The surveyed literature was organised through a taxonomy that specifies adversarial goals, the operational roles of xAI, and the Threat Model. The paper closes with a research agenda for handling xAI-weaponising attacks. Taken together, this survey reframes xAI from a purely diagnostic tool to a security-critical surface and provides the groundwork for principled defences and future empirical studies.

ACKNOWLEDGMENTS

REFERENCES

- Aivodji, U., Bolot, A., and Gambs, S. (2020). Model extraction from counterfactual explanations. *arXiv preprint arXiv:2009.01884*.
- Amich, A. and Eshete, B. (2022). Eg-booster: explanation-guided booster of ml evasion attacks. In *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy*, pages 16–28.
- Baniecki, H. and Biecek, P. (2024). Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, 107:102303.
- Bayer, M., Neiczer, M., Samsinger, M., Buchhold, B., and Reuter, C. (2024). Xai-attack: Utilizing explainable ai to find incorrectly learned patterns for black-box adversarial example creation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17725–17738.
- Chanda, D., Gheshlaghi, S. H., and Soltani, N. Y. (2025). Explainability-based adversarial attack on graphs through edge perturbation. *Knowledge-Based Systems*, 310:112895.
- Chen, Z., Silvestri, F., Wang, J., Zhang, Y., and Tolomei, G. (2023). The dark side of explanations: Poisoning recommender systems with counterfactual examples. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, page 2426–2430, New York, NY, USA. Association for Computing Machinery.
- Dai, Z., Liu, S., Li, Q., and Tang, K. (2023). Saliency attack: Towards imperceptible black-box adversarial attack. *ACM Transactions on Intelligent Systems and Technology*, 14(3):1–20.
- Kozik, R., Ficco, M., Pawlicka, A., Pawlicki, M., Palmieri, F., and Choraś, M. (2024). When explainability turns into a threat-using xai to fool a fake news detection method. *Computers & Security*, 137:103599.
- Kumagai, R., Takemoto, S., Nozaki, Y., and Yoshikawa, M. (2023). Explainable ai based adversarial examples and its evaluation. In *Proceedings of the 2023 6th International Conference on Electronics, Communications and Control Engineering*, pages 220–225.
- Liu, H., Wu, Y., Yu, Z., and Zhang, N. (2024a). Please tell me more: Privacy impact of explainability through the lens of membership inference attack. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 4791–4809.
- Liu, H., Wu, Y., Yu, Z., and Zhang, N. (2024b). Please tell me more: Privacy impact of explainability through the lens of membership inference attack. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 4791–4809. IEEE.

- 402 Liu, H., Zuo, X., Huang, H., and Wan, X. (2022a). Saliency map-based local white-box adversarial attack
403 against deep neural networks. In *CAAI International Conference on Artificial Intelligence*, pages
404 3–14. Springer.
- 405 Liu, M., Liu, X., Yan, A., Qi, Y., and Li, W. (2022b). Explanation-guided minimum adversarial attack. In
406 *International Conference on Machine Learning for Cyber Security*, pages 257–270. Springer.
- 407 Luo, X., Jiang, Y., and Xiao, X. (2022). Feature inference attack on shapley values. In *Proceedings of the*
408 *2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2233–2247.
- 409 Miura, T., Shibahara, T., and Yanai, N. (2024). Megex: Data-free model extraction attack against gradient-
410 based explainable ai. In *Proceedings of the 2nd ACM Workshop on Secure and Trustworthy Deep*
411 *Learning Systems*, pages 56–66.
- 412 Okada, S., Jmila, H., Akashi, K., Mitsunaga, T., Sekiya, Y., Takase, H., Blanc, G., and Nakamura, H.
413 (2024). Xai-driven adversarial attacks on network intrusion detectors. In *Proceedings of the 2024*
414 *European Interdisciplinary Cybersecurity Conference*, pages 65–73.
- 415 Okada, S., Jmila, H., Akashi, K., Mitsunaga, T., Sekiya, Y., Takase, H., Blanc, G., and Nakamura, H.
416 (2025). Xai-driven black-box adversarial attacks on network intrusion detectors. *International Journal*
417 *of Information Security*, 24(3):1–15.
- 418 Oksuz, A. C., Halimi, A., and Ayday, E. (2024). Autolycus: Exploiting explainable artificial intelligence
419 (xai) for model extraction attacks against interpretable models. *Proceedings on Privacy Enhancing*
420 *Technologies*.
- 421 Pawlicki, M., Pawlicka, A., Kozik, R., and Choraś, M. (2024). Explainability versus security: The
422 unintended consequences of xai in cybersecurity. In *Proceedings of the 2nd ACM Workshop on Secure*
423 *and Trustworthy Deep Learning Systems*, pages 1–7.
- 424 Shi, J., Li, L., and Zeng, D. (2024). Shapattack: Shapley-guided multigranularity adversarial attack against
425 text transformers. *IEEE Intelligent Systems*, 39(3):45–53.
- 426 Shokri, R., Strobel, M., and Zick, Y. (2021). On the privacy risks of model explanations. In *Proceedings of*
427 *the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 231–241.
- 428 Shu, Z. and Yan, G. (2023). Eagle: evasion attacks guided by local explanations against android malware
429 classification. *IEEE Transactions on Dependable and Secure Computing*, 21(4):3165–3182.
- 430 Tam Nguyen, T., Huynh, T. T., Ren, Z., Nguyen, T. T., Le Nguyen, P., Yin, H., and Nguyen, Q. V. H. (2024).
431 A survey of privacy-preserving model explanations: Privacy risks, attacks, and countermeasures.
432 *arXiv e-prints*, pages arXiv–2404.
- 433 Vaddillo, J., Santana, R., and Lozano, J. A. (2025). Adversarial attacks in explainable machine learning:
434 A survey of threats against models and humans. *Wiley Interdisciplinary Reviews: Data Mining and*
435 *Knowledge Discovery*, 15(1):e1567.
- 436 Vu, K., Lai, P., and Nguyen, T. (2024). Xsub: Explanation-driven adversarial attack against blackbox
437 classifiers via feature substitution. In *2024 IEEE International Conference on Big Data (BigData)*,
438 pages 1599–1604.
- 439 Xu, J., Xue, M., and Picek, S. (2021). Explainability-based backdoor attacks against graph neural networks.
440 In *Proceedings of the 3rd ACM workshop on wireless security and machine learning*, pages 31–36.
- 441 Yan, A., Hou, R., Yan, H., and Liu, X. (2023a). Explanation-based data-free model extraction attacks.
442 *World Wide Web*, 26(5):3081–3092.
- 443 Yan, A., Huang, T., Ke, L., Liu, X., Chen, Q., and Dong, C. (2023b). Explanation leaks: Explanation-guided
444 model extraction attacks. *Information Sciences*, 632:269–284.
- 445 Yan, A., Liu, X., Li, W., Ye, H., and Li, L. (2024). Explanation-guided adversarial example attacks. *Big*
446 *Data Research*, 36:100451.
- 447 Zhan, D., Bai, W., Liu, X., Hu, Y., Zhang, L., Guo, S., and Pan, Z. (2023). Psp-mal: Evading malware
448 detection via prioritized experience-based reinforcement learning with shapley prior. In *Proceedings*
449 *of the 39th Annual Computer Security Applications Conference*, pages 580–593.

- 450 Zhang, D., Dong, Y., and Yang, Y. (2024). Sample-analysis based adversarial attack with saliency map.
451 *Applied Soft Computing*, 161:111733.
- 452 Zhao, X., Zhang, W., Xiao, X., and Lim, B. (2021a). Exploiting explanations for model inversion attacks. In
453 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 682–692.
- 454 Zhao, X., Zhang, W., Xiao, X., and Lim, B. (2021b). Exploiting explanations for model inversion attacks.
455 In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 682–692.
- 456 Zhu, H., Zhao, Q., Shang, W., Wu, Y., and Liu, K. (2024). Limeattack: Local explainable method for
457 textual hard-label adversarial attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*,
458 volume 38, pages 19759–19767.

