SoK: Explainable Machine Learning in Adversarial Environments

Maximilian Noppel
KASTEL Security Research Labs
Karlsruhe Institute of Technology
Germany

Christian Wressnegger
KASTEL Security Research Labs
Karlsruhe Institute of Technology
Germany

***Abstract*—Modern deep learning methods have long been considered black boxes due to the lack of insights into their decision-making process. However, recent advances in explainable machine learning have turned the tables. Post-hoc explanation methods enable precise relevance attribution of input features for otherwise opaque models such as deep neural networks. This progression has raised expectations that these techniques can uncover attacks against learning-based systems such as adversarial examples or neural backdoors. Unfortunately, current methods are not robust against manipulations themselves. In this paper, we set out to systematize attacks against post-hoc explanation methods to lay the groundwork for developing more robust explainable machine learning. If explanation methods cannot be misled by an adversary, they can serve as an effective tool against attacks, marking a turning point in adversarial machine learning. We present a hierarchy of explanation-aware robustness notions and relate existing defenses to it. In doing so, we uncover synergies, research gaps, and future directions toward more reliable explanations robust against manipulations.**

***Index Terms*—Explainable Machine Learning, XAI, Attacks, Defenses, Robustness Notions**

1. Introduction

Ever since the wide adoption of machine learning, the community has striven for ways to explain the inner workings of learned models [92, 100]. The complexity and non-linearity of modern deep learning schemes has, however, raised the bar to do so distinctively [139]. In contrast to simple linear models, neural networks are not inherently explainable [115, 137]. Recent research has thus brought forward various *post-hoc explanation* methods that can precisely attribute relevance to input features, explaining the decision-making process of the model [2, 67, 199]. These techniques are applied to existing machine learning models, operate on individual input samples, and are either model agnostic [e.g., 90, 104, 129, 130] or are tailored to the specific model at hand [e.g., 12, 141, 144, 147, 160, 191]. The former are referred to as *"black box explanations"* as they only consider high-level input-output relations, while the latter, so-called *"white box explanations"*, assume perfect knowledge of all model parameters and involved computations [23].

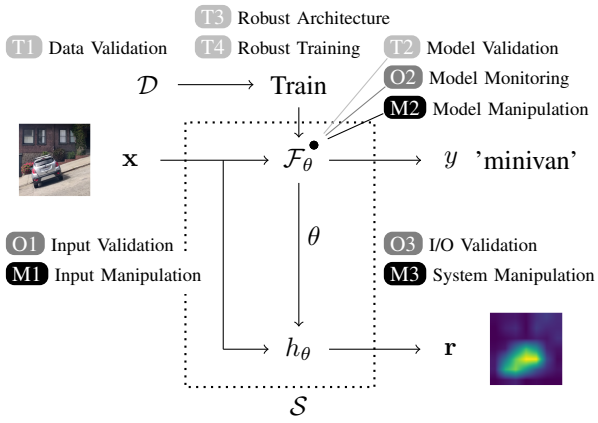

Figure 1: Attacks (M) against an XAI-system $\mathcal{S}$ and defense strategies during model training (T) and system operation (O). $\mathcal{F}_\theta$ predicts class label y for input $\mathbf{x}$ based on model θ that is trained on a training dataset $\mathcal{D}$. h_θ represents a post-hoc explanation method deriving an explanation $\mathbf{r}$ of the input sample.

With the advent of adversarial machine learning [120] the need to reliably explain learned models continues unabated [62, 119]. Spurious correlations can cause shortcuts in the decision-making process [91] which may be used to bypass detection [103, 173]. It thus is oftentimes recommended applying explanation techniques "to gain a better view of the capabilities of a learning-based systems" [10]. However, explanations themselves can be forged through input manipulations [48, 152, 198] and various model manipulations [118, 151] similar to adversarial examples [161] and neural backdoors [65, 173], respectively. Thus, it is questionable whether current post-hoc explanation methods can indeed accomplish this goal in practice.

In this paper, we lay the groundwork for understanding the limits and potentials of explainable machine learning in adversarial environments. We systematize attacks against explanation methods, so-called *explanation-aware attacks*. Fig. 1 shows the different leverage points of an adversary, that is the different threat models for input manipulation (M1), model manipulation (M2), and whole system manipulation (M3). The latter represents a particularly strong adver-

sary, capable of manipulating any aspect of the XAI-system at any point of the pipeline. Additionally, our systematization covers different attack objectives in terms of the attack's outreach and scope. While the first considers whether the explanations or predictions of the original model are preserved or both are manipulated. The latter differentiates targeted, semi-targeted, and untargeted attacks. This categorization covers traditional input-level [e.g., 48, 86, 198] and model-level manipulations [e.g., 71, 118, 195], but also more recent system-level threats such as fairwashing [e.g., 3, 8, 151].

For further analysis, we then define robustness notions of post-hoc explanation methods, modeling pairs of constraints and restrictions that may be enforced to yield robust explanations. One central finding of this systematized *hierarchy of robustness notions* is that different authors describe different understandings of robustness with the same key phrase. This underlines the necessity of a common understanding of explanation-aware robustness to push forward research in this domain. In particular, robustness guarantees for post-hoc explanation methods can be key to effectively defend against adversaries attacking learning-based systems.

Based on this fundamental understanding of attacks and robustness requirements, we proceed to taxonomize existing defenses to counter explanation-aware attacks. In Fig. 1, we mark locations in the machine learning pipeline where defenses can be applied during training (T1 – T4) and at inference time during operation (O1 – O3).

Compared to defenses against prediction-only attacks [e.g., 14, 44, 106, 126, 138, 173], defenses against explanation-aware attacks are scarce, as shown in a concurrent survey [16]. For our systematization, we thus explicitly relate to defenses that have been explored for conventional input and model manipulation attacks against a classifier's prediction. In doing so, we motivate the community to step back and unify the knowledge of both fields to find more effective defenses. We pose open research questions, whose answers will eventually close the gap between these highly related subfields. Most pressingly *"What robustness guarantees can we provide for post-hoc explanation methods?"*

We are confident that our systematization can act as a signpost toward effective defenses against explanation-aware attacks. Most importantly, however, we build a foundation and provide guidelines for creating and proving robust explanation methods.

In summary, we make the following contributions:

- **Systematization of attacks against explanations.** We systematize attacks against post-hoc explanations along the adversary's capabilities, constraints, and objectives. In doing so, we highlight crucial differences and similarities of threat models and attack objectives specific to explanation-aware attacks.
- **Explanation-aware robustness notions.** We formalize robustness against explanation-aware attacks and form a hierarchy of robustness notions, that enables us to compare, link, and unify efforts in the community. Moreover, we find that robustness notions might be conflicting for adversaries with diverging objectives.
- **Taxonomy of defenses.** We taxonomize existing defenses based on the usual machine learning pipeline and relate to defenses for prediction-only attacks. In each of the identified categories, we raise research questions that will help advance defense against explanation-aware attacks.
- **Future research directions.** We point out specific directions for future research on explanation-aware robustness in order to help advance the field.

2. Background on Explainable ML

Similar to machine learning and artificial intelligence, explainable machine learning and explainable artificial intelligence (XAI) are particularly wide fields of research. Hence, in this section, we take a step back to detail the scope considered in the remainder of the paper, laying the groundwork for discussing attacks and defense within this scope.

We consider *post-hoc explanation methods* that are applied to an existing (potentially difficult to explain) machine learning model, explaining a single input at a time. These methods have the benefit of being generic and applicable to a large group of models. Moreover, performing a learning task and explaining its results are fully disjoint components, allowing for methodological improvement of both aspects separate from each other. For this reason, models that are intrinsically explainable, e.g., linear support vector machines [143] or decision trees [133], are not part of our systematization. The downside of a separate (post-hoc) consideration of explanations is that this different viewpoint on the model potentially diverges from the model's actual inference [60, 137]. An adversary may use this difference to make the prediction (the model's inference) report one thing while the explanation reports something else.

In the following, we detail the notation used in the remainder of the paper and formally define post-hoc explanation methods. Next, we briefly outline how post-hoc explanations are generated and how they are presented to the user. We emphasize that we do not attempt to summarize the field of XAI and refer the reader to surveys on the topic [2, 31, 36, 39, 45, 46, 67, 109, 155, 199]. The notation below forms a general framework for systematizing attacks against post-hoc explanations that fits the state of the art of these explanation techniques.

Notation. A machine-learning model is represented by its parameters $\theta = (w, \mathbb{A}) \in \Theta$, defined as an element of all possible models Θ. The model's weights and its architecture are referred to as w and $\mathbb{A}$, respectively. Moreover, we denote the set of models with a fixed architecture $\Theta_{\mathbb{A}} \subset \Theta$. Each model maintains a decision function $f_\theta : \mathcal{X} \to \mathcal{Y}$ that accepts a d-dimensional input $\mathbf{x}$ from the feature space $\mathcal{X} \subseteq \mathbb{R}^d$ and derives a probability score for every possible class $c \in [C]$ as a C-dimensional vector in $\mathcal{Y} \subseteq [0,1]^C$. The winning class is determined by $\mathcal{F}_\theta(\mathbf{x}) := \arg\max_i f_\theta(\mathbf{x})_i$. The considered dataset of n inputs $X \subseteq \mathcal{X}$ and corresponding ground-truth labels $Y \in [C]^n$ is denoted as $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$.

The dataset is then segmented to form training, validation, and testing dataset accordingly.

A post-hoc explanation method determines relevance of an input's features to $\mathcal{F}_\theta$ for arriving at that prediction for a specific input. White box explanation methods take recourse to knowledge about the model's parameters $\theta \in \Theta$ implementing the function

$$h_\theta^w : \mathcal{X} \times \Theta \to \mathcal{E} \ ,$$

where $\mathcal{E}$ is the explanation space, that is, the input's representation space for which relevance values are determined. While this space can have different instantiations [156, 171], in most settings, however, $\mathcal{E}$ is a subset of the input's feature space $\mathcal{X}$. Black box explanation methods, in turn, implement the function

$$h_\theta^b : \mathcal{X} \times \{f_\theta\}_{\theta \in \Theta} \to \mathcal{E} \ ,$$

where $\{f_\theta\}_{\theta \in \Theta}$ represents black box access to a model. Such methods have access to the input-output behavior of the respective decision function, $f_\theta(\mathbf{x})$, only.

For formal consideration in the remainder of the paper, we additionally assume both $\mathcal{X}$ and $\mathcal{E}$ to be metric spaces, $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{E}, d_\mathcal{E})$, with their respective distance metrics $d_\mathcal{X}$ and $d_\mathcal{E}$ (cf. Section 5).

For two particular results, we assume $\mathcal{X}$ and $\mathcal{E}$ to also be convex. This assumption holds whenever the inputs are vectors or matrices of floating point numbers. However, for categorical and discrete features this is not the case and these two results do not necessarily apply. We emphasize these limitations at the corresponding positions in the paper.

Generating Explanations. The strategies for generating explanations differ greatly, ranging from perturbation-based [56, 90, 104, 129, 192], gradient-based [13, 83, 147, 160], and propagation-based methods [12, 110–112, 146, 196], over concept-based [61, 82, 190, 197] and example-based [26, 41, 84] explanations, to activation-based techniques [17, 32, 94, 144, 201]. Note that this is not an all-encompassing list of methods but rather a rough overview. Additionally, multiple techniques can be utilized at once, e.g., SmoothGrad [154] can be considered perturbation-based and gradient-based, and GradCAM [144] can be regarded as a combination of activation-based and gradient-based. However, these techniques have in common that they fit the proposed notation of mapping relevance/importance to features in an explanation space $\mathcal{E}$.

Please refer to dedicated surveys on explanation techniques for a more comprehensive discussion [2, 31, 36, 39, 45, 46, 67, 109, 155, 199]. Whether individual approaches have shortcomings such as redundancy, incompleteness, or inconsistency [155, 182] is a discussion orthogonal to our systematization of attacks against explanation methods.

Representing Explanations. Most commonly, post-hoc explanations are visualized as heatmaps or saliency maps, indicating the relevance of a specific feature or set of features of the feature space with more or less bright colors. Examples of different representations and thus explanation spaces include pixel-based images [12, 147], text or program code [11, 68], tabular data [129, 130], or graphs [141, 191]. Assuming the explanation space is equivalent to the feature space or a subspace of it, e.g., for feature space $\mathbb{R}^{c \times h \times w}$ the $\mathcal{E}$ might be $\mathbb{R}^{h \times w}$, meaning, it assigns relevance values to pixels instead of individual color values. This can be enforced by reducing the channel dimension via averaging or taking the maximum [147].

Additional forms of representation include contrastive and counterfactual explanations [169, 172], explainable surrogate models and rule sets [66, 89, 123, 127, 129, 130], relations comparing two instances [52, 130], or multimodal explanations that, for instance, use text to describe images [79, 121]. Again, this is not an exhaustive list, but a rough overview. We regard the specific representation of an explanation as independent of our systematization and refer the reader to related work of this adjacent research area [108, 171].

3. Threat Models

We begin to systematize the adversary's capabilities and their associated constraints. A schematic depiction of the considered attack vectors is provided in Fig. 1. More formally, an adversary $\mathcal{A}$ attacking local post-hoc explanation methods faces an optimization problem, abstractly defined as:

$$\min_{\mathcal{A}(\mathfrak{K})} obj \text{ s.t. } \omega_1(\ldots) \wedge \ldots \wedge \omega_N(\ldots)$$

The adversary operates on certain *knowledge* $\mathfrak{K}$ to manipulate inputs $\mathbf{x}_i$ (M1), the model θ (M2), and/or the entire XAI-system $\mathcal{S}$ (M3). She may know (a subset of) all original inputs $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ or any information about the model and the original system $\mathcal{S}$ (e.g., the model's architecture $\mathbb{A}$, its weights w, or the used explanation method h and its parameters). However, depending on the threat model, the available knowledge and the output of the adversary varies. Additionally, constraints ω_k restrict the adversary and define her capabilities. For instance, she may perturb inputs within a certain ϵ only, must not change the model's architecture $\mathbb{A}$, or needs to apply the same manipulation on every input.

In the following, we detail threat models for manipulating inputs (Section 3.1), the model (Section 3.2), and the overall XAI-system (Section 3.3). Each has different constraints and allows for varying degree of knowledge $\mathfrak{K}$, thus demanding distinct attack capabilities and forcing the adversary to produce different outputs, $\star \leftarrow \mathcal{A}(\mathfrak{K})$. Note that the attack objective obj itself is detailed in Section 4, independently of the threat models. Additionally, Table 2 lists works using these threat models.

3.1. Input Manipulation M1

In the context of explanation-aware attacks the ability to manipulate an input $\mathbf{x}$ is most frequently investigated [1, 5, 21, 48, 76, 150, 198]. Commonly, the adversary knows about a specific sample $\mathbf{x}$ and perturbs it within the attack's constraints, yielding a malicious input $\tilde{\mathbf{x}} \leftarrow \mathcal{A}(\mathbf{x}, \ldots)$. In the following, we discuss different constraints:

Imperceptibility. The adversary aims for imperceptible perturbations. Hence, the difference between $\mathbf{x}$ and $\tilde{\mathbf{x}}$ should be small, meaning both inputs need to be as similar as possible:

$$\omega_k(\mathbf{x}, \tilde{\mathbf{x}}) := d_{\mathcal{X}}(\mathbf{x}, \tilde{\mathbf{x}}) \leq \epsilon \ ,$$

for a limit $\epsilon \in \mathbb{R}^+$. A straightforward choice for $d_{\mathcal{X}}$ frequently considered in literature [e.g., 29, 48, 62, 161] are the L_p norms, $\|\mathbf{x} - \tilde{\mathbf{x}}\|_{p \in \{0,1,2,\infty\}}$. For visual and audible inputs there exist more sophisticated ways to measure the perceptability to a human analyst, e.g., the "Structural Similarity Index (SSIM)" for images [180].

Universality. Moreover, an adversary may operate on single inputs or consider manipulations that universally apply to multiple inputs at once. Universal input manipulations are restricted to operations that perturb the input independent of the input. Their evasive properties (defined by the overall objective) need to apply for all inputs $\mathbf{x}_i$. Examples include, adding a constant additive offset [113] or patch replacement [159]. The latter also enforces a constraint on the "input region", meaning the features have to be manipulated as described in the following.

Feature Subspaces. Occasionally the adversary only controls a small, specific portion of each input, enforcing a masked input manipulation [22, 97, 187]. More generally, the constraint limits the perturbation's feature space and is modeled by a mask vector $\mathbf{m} \in \{0,1\}^d$, which is either a fixed parameter of the threat model or chosen by the adversary. We formalize this as:

$$\omega_k(\mathbf{x}, \tilde{\mathbf{x}}) := ((\mathbf{x} \neq \tilde{\mathbf{x}}) \vee \mathbf{m}) = \mathbf{m} \ ,$$

where $\vee$ and $\neq$ are interpreted as element-wise operations. Various extension exist, for instance: the mask might be required to be a square [97, 187] or small in size $\|\mathbf{m}\|_0 \leq \epsilon_0$ [22, 170]. If $\mathbf{m}$ is chosen by the adversary, the latter links to imperceptibility under L_0 over a set of samples $\mathbf{x}_i$ and their manipulated counterparts $\tilde{\mathbf{x}}_i$: $\left\| \bigvee_i (\mathbf{x}_i \neq \tilde{\mathbf{x}}_i) \right\|_0 \leq \epsilon_0$, where $\vee$ and $\neq$ are interpreted element-wise again.

3.2. Model Manipulation M2

An adversary manipulating the model, $\tilde{\theta} \leftarrow \mathcal{A}(\mathfrak{K})$, may change the model's weights (and biases) w and may even alter the underlying architecture $\mathbb{A}$. We define the weight manipulated model as $\tilde{\theta} := (w + \delta_w, \mathbb{A})$. For full model manipulation, in turn, the adversary creates an independent model from scratch $\tilde{\theta} \in \Theta$ without any restrictions on the model's architecture $\mathbb{A}$.

Despite the capability to fully manipulate the model, the adversary is constrained regarding the side effects her changes may have. A manipulated model needs to either (a) achieve a similar validation/testing accuracy or (b) make the same mistakes as the original model in order to bypass functionality tests before deployment. Based on the notion of ϵ-accuracy and ϵ-agreement defined below, we formulate model parameter manipulation as:

$$\min_{\delta_w \leftarrow \mathcal{A}(\mathfrak{K})} obj \text{ s.t. } \omega_k := \begin{cases} \mathcal{F}_{(w+\delta_w, \mathbb{A})} \text{ is } \epsilon\text{-accurate} \\ \mathcal{F}_\theta, \ \mathcal{F}_{(w+\delta_w, \mathbb{A})} \text{ are } \epsilon\text{-agreeing,} \end{cases}$$

and model manipulation as:

$$\min_{\tilde{\theta} \leftarrow \mathcal{A}(\mathfrak{K})} obj \text{ s.t. } \omega_k := \begin{cases} \mathcal{F}_{\tilde{\theta}} \text{ is } \epsilon\text{-accurate} \\ \mathcal{F}_\theta \text{ and } \mathcal{F}_{\tilde{\theta}} \text{ are } \epsilon\text{-agreeing} \ . \end{cases}$$

Note that these constraints apply to the overall model operation rather than single inputs. Moreover, we denote $\mathcal{F}_{\tilde{\theta}}$ in the following sections for the sake of simplicity, even if only the learned parameters w are perturbed.

ϵ-Accuracy. The manipulated model should achieve a similar accuracy (on testing data) but may differ in classification errors compared to the original model. Formally, we measure the accuracy as

$$acc(\mathcal{F}_{\tilde{\theta}}) = \mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}_{test}}[\mathcal{F}_{\tilde{\theta}}(\mathbf{x}) = y] \ .$$

In comparison to a benignly trained reference model, we define a manipulated model as ϵ-accurate as follows:

Definition 1 (ϵ-accurate Model). — *We denote the manipulated model $\mathcal{F}_{\tilde{\theta}}$ as* ϵ-accurate *model regarding a benignly trained model $\mathcal{F}_\theta$ if its test accuracy is not significantly lower than the accuracy of the reference model $\mathcal{F}_\theta$*

$$acc(\mathcal{F}_\theta) - acc(\mathcal{F}_{\tilde{\theta}}) \leq \epsilon$$

for a limit $\epsilon \in \mathbb{R}^+$.

ϵ-Agreement. If a manipulated model's predictions match the decisions of the original model, we denote both as having a high fidelity. In line with related work [3, 4, 43], we define the fidelity (and the equivalent "agreement rate") of two models as follows:

Definition 2 (Fidelity/Agreement Rate). — *We denote the* fidelity *and* agreement rate *between two classifiers $\mathcal{F}_\theta$ and $\mathcal{F}_{\tilde{\theta}}$ based on a distribution or set of samples P as*

$$\mathit{fid}_P(\mathcal{F}_\theta, \mathcal{F}_{\tilde{\theta}}) := \mathbb{E}_{\mathbf{x} \sim P}[\mathcal{F}_\theta(\mathbf{x}) = \mathcal{F}_{\tilde{\theta}}(\mathbf{x})] \ .$$

We refer to the above as *local fidelity around* $\mathbf{x}$ if P is the neighborhood $\mathcal{N}_\mathbf{x}$ of a sample $\mathbf{x}$, and as *global fidelity* if P is the complete feature space $\mathcal{X}$. The agreement rate as defined above is the exact inverse of the occasionally used "disagreement rate" [89].

Definition 3 (ϵ-agreeing Models). — *We denote two classifiers $\mathcal{F}_\theta$ and $\mathcal{F}_{\tilde{\theta}}$ as* ϵ-globally agreeing *if*

$$1 - \mathit{fid}_{\mathcal{X}}(\mathcal{F}_\theta, \mathcal{F}_{\tilde{\theta}}) \leq \epsilon \ .$$

Intuitively, the empirical measurement of the probability that both models arrive at the same (potentially wrong) prediction should be close to 1.

Other Constraints. Next to the above descriptions, the adversary may be subject to further application-specific constraints. Similar to prediction-only attacks that have been

shown for a vast number of corner cases. For instance, manipulations may be constrained to a few bit-flips [128]; weight perturbations need to remain below a specific threshold, $\|\delta_w\|_\infty \leq \epsilon$ [58]; or the attack takes effect only if the model is compressed [165] or quantized [73]. Similar constraints are easily conceivable for explanation-aware attacks as well.

Dataset Manipulation. A model can also be *indirectly* manipulated by poisoning the training data [18]. While popular for prediction-only attacks [59, 77, 78, 145], the relevance of data poisoning for explanation-aware attacks remains an open research question. It is not immediately apparent how such an attack would be instantiated by perturbing inputs or labels only. One option is to encircle target samples with poisoned samples to alter their explanations [195]. This process, however, is rather costly for attacking a few samples.

Moreover, various works augment the training process and its training data to guide the model in its decision-making [33, 51, 132, 134, 184, 202]. An added loss term considers the distance to a ground truth explanation per sample as regularization. This additional regularization supposedly suppresses spurious correlations by considering domain knowledge. For instance, Chefer et al. [33] let the model focus on the object rather than the background. The similarity to model manipulations that facilitate explanation-aware attacks raises the question whether the model indeed learns to avoid spurious correlations or if the explanations are accidentally forged.

> **Observation.** Guiding the learning process through ground truth explanations shares properties with explanation-aware model manipulations. These similarities (a) pose the risk of guidance actually being an unwilling manipulation and simultaneously (b) may provide insights on how to channel positive effects of explanation-aware manipulations.

3.3. System Manipulation M3

System manipulations represent the strongest adversarial capability that we consider in this work. Also, while input and model manipulations are common in conventional attacks against machine learning, the capability to manipulate the whole system is most significant in explainable machine learning. As an example, a potential adversarial goal is to hide the unfair or biased reasoning of the system [3, 8, 151]. Whether the bias is introduced on purpose is secondary. For hiding biases, the adversary may, for instance, learn a set of unbiased surrogate models for the biased black box model and report explanations for the surrogate that yields the most similar prediction to the biased model [3]. With the explanation being generated on the unbiased surrogate model, the adversary feigns a fair decision.

We denote the complete functionality of generating a prediction and an explanation as XAI-system $\mathcal{S}$

$$\mathcal{S} : \mathcal{X} \times \Theta \to Y \times \mathcal{E} \ ,$$

as depicted in Fig. 1.

We write $\mathcal{S}_f$ or $\mathcal{S}_\mathcal{F}$ if we are concerned about the prediction output and $\mathcal{S}_h$ for the explanation output respectively. Hence, we can describe the general optimization problem as

$$\min_{\tilde{\mathcal{S}} \leftarrow \mathcal{A}(\mathfrak{K})} obj \text{ s.t. } \omega_k := \begin{cases} \tilde{\mathcal{S}}_\mathcal{F} \text{ is } \epsilon\text{-accurate} \\ \mathcal{S}_\mathcal{F} \text{ and } \tilde{\mathcal{S}}_\mathcal{F} \text{ are } \epsilon\text{-agreeing} \ . \end{cases}$$

4. Attack Objectives

An adversary may target two fundamental aspects of a learning-based system: (1) the predictions [120] or (2) the explanations. Each target can be pursued in isolation or in combinations, giving rise to different attack classes. Prediction-only attacks do not consider explanations at all, while explanation-aware attacks do.

The adversary's objective may be to either alter, preserve, or ignore a specific target. In the following, we briefly discuss each option before we elaborate on the different combinations

(a) Prediction Preservation. A common objective is to preserve the prediction [48, 76, 150, 177]. How well predictions are preserved can be measured by the cross entropy loss, denoted as $\mathcal{L}_{CE}(\tilde{\mathcal{S}}_\mathcal{F}(\tilde{\mathbf{x}}), \cdot)$.

(b) Prediction Alteration. Altering the prediction alone is extensively discussed in literature, covering input manipulations [62, 113, 161] and model manipulations [65, 102, 173]. In the context of this paper, however, we do not consider these details. Instead, we generalize and simply write $\mathrm{att}_\mathcal{F}(\tilde{\mathcal{S}}_\mathcal{F}(\tilde{\mathbf{x}}), \cdot)$ as an *attack* objective on the prediction.

(c) Explanations Preservation. Here, the explanation should be as similar to the corresponding benign explanation as possible. For instance, to disguise an ongoing input manipulation [198] or model manipulation [118]. We denote this objective by $d_\mathcal{E}(\mathcal{S}_h(\mathbf{x}), \tilde{\mathcal{S}}_h(\tilde{\mathbf{x}}))$.

(d) Explanation Alteration. Finally, the adversary may choose to alter the explanation, for instance, to distract the analyst through input manipulations [48] or make a model fair while it is not [3]. In the following, we denote this *attack* objective by $\mathrm{att}_h(\tilde{\mathcal{S}}_h(\tilde{\mathbf{x}}), \cdot)$.

Based on the presented criteria we categorize explanation-aware attacks in Section 4.1. Afterwards, we elaborate on specific instantiations of the "Explanation Alteration" objective in Section 4.2.

4.1. Explanation-Aware Attacks

The combination of two objectives (preservation and alteration) for two targets (prediction and explanation) yields the three subclasses of explanation-aware attacks that we consider in this paper. Note that the fourth option, where the adversary preserves the prediction as well as the explanation is not an attack [142]. Table 1 provides an overview, while Table 2 lists corresponding related work.

TABLE 1: Prediction-only adversaries (first line, grayed out) that do not address explanations are not considered in this paper. Instead, we focus on explanation-aware attacks, that we categorize by objectives in explanation-preserving (EP), prediction-preserving (PP), and dual (D) attacks.

Name	Prediction	Explanation
Prediction-Only (AE/BD)	alter	ignore
Explanation-Preserving (EP)	alter	preserve
Prediction-Preserving (PP)	preserve	alter
Dual (D)	alter	alter

Explanation-Preserving Attack. The adversary attempts to keep the explanation unchanged while attacking the prediction, leading to the following optimization problem:

$$\min_{\tilde{\mathbf{x}},\tilde{\mathcal{S}}\leftarrow\mathcal{A}(\mathfrak{K})} \mathrm{att}_{\mathcal{F}}\big(\tilde{\mathcal{S}}_{\mathcal{F}}(\tilde{\mathbf{x}}),\cdot\big) + d_{\mathcal{E}}\big(\mathcal{S}_h(\mathbf{x}),\tilde{\mathcal{S}}_h(\tilde{\mathbf{x}})\big) \ .$$

Prediction-Preserving Attack. The adversary attempts to keep the prediction as accurate or agreeing with the benign prediction as possible. Simultaneously, she attacks the explanation method, leading to the following optimization problem:

$$\min_{\tilde{\mathbf{x}},\tilde{\mathcal{S}}\leftarrow\mathcal{A}(\mathfrak{K})} \mathcal{L}_{CE}\big(\tilde{\mathcal{S}}_{\mathcal{F}}(\tilde{\mathbf{x}}),\cdot\big) + \mathrm{att}_h\big(\tilde{\mathcal{S}}_h(\tilde{\mathbf{x}}),\cdot\big)$$

> **Observation.** *Perfect fidelity* can be easily achieved through manipulating the XAI-system by using the original model for the prediction output.

Dual Attack. The adversary targets both, the prediction and the explanation. In our systematization, this attack corresponds to the following optimization problem:

$$\min_{\tilde{\mathbf{x}},\tilde{\mathcal{S}}\leftarrow\mathcal{A}(\mathfrak{K})} \mathrm{att}_{\mathcal{F}}\big(\tilde{\mathcal{S}}_{\mathcal{F}}(\tilde{\mathbf{x}}),\cdot\big) + \mathrm{att}_h\big(\tilde{\mathcal{S}}_h(\tilde{\mathbf{x}}),\cdot\big) \ .$$

4.2. Scope of Explanation Alteration

We proceed to specify different types of attacks that alter explanations. In line with prior research on prediction-only attacks [120], we use the terms *untargeted* and *targeted*, and adapt them to explanation-aware attacks [5, 163]. Additionally, we consider *semi-targeted* attacks that are specific to explanation-aware attacks.

4.2.1. Untargeted Attacks. In this setting, the objective is to yield an explanation, that should be maximally different to the benign explanation. Hence, the set of optimal solutions is dependent on the benign explanation and potentially contains more than one optimal explanation. Whether this is possible, depends on the used metric and setting. We formalize the *untargeted attack objective* for altering explanations as:

$$\mathrm{att}_h^{untar} := \frac{1}{d_{\mathcal{E}}\big(\tilde{\mathcal{S}}_h(\tilde{\mathbf{x}}),\mathcal{S}_h(\mathbf{x})\big)} \ .$$

Note that potentially every input yields a different forged explanation, hence, the attack resembles a `n:n` relation from inputs to explanations.

Example: Top-k Fooling. Minimizing the top-k overlap between the benign explanation and the manipulated one can be considered as an untargeted attack [71, 98]. The originally k-most relevant features should receive as little relevance as possible in the manipulated explanation. However, the exact ranking of features is not crucial and, thus, multiple explanations can meet the objective.

4.2.2. Targeted Attacks. Here, the objective is to minimize the distance between the manipulated explanation and a fixed target explanation $\mathbf{r}_t$ [48, 118]. In contrast to untargeted attacks, a single optimal solution exists, namely the target explanation. We formalize the *targeted attack objective* for altering the explanation as:

$$\mathrm{att}_h^{tar} := d_{\mathcal{E}}\big(\tilde{\mathcal{S}}_h(\tilde{\mathbf{x}}),\mathbf{r}_t\big) \ .$$

> **Observation.** A targeted attack can be evaluated by using the benign explanation of an "in-distribution" reference sample as a target [48, 131]. This way, the yield explanation is guaranteed to be plausible. However, certain target explanations, like attributing zero relevance to each feature, might be impossible.

4.2.3. Semi-Targeted Attacks. Finally, an attack can be neither untargeted nor targeted. This is the case when the exact target explanation depends on the input, e.g., inverting the benign explanation [118], swapping the explanations of two classes [71], or suppressing the relevance at a specific location but keeping the remaining explanation intact [159]. For attacking a single sample there is no difference to a targeted attack. However, for multiple instances semi-targeted attacks may implement arbitrary functions in $\mathcal{E}$. We generalize semi-targeted attacks by a function $\mu : \mathcal{E} \times \mathcal{X} \rightarrow \mathcal{E}$ that is applied to the benign explanation yielding a sample-specific target explanation:

$$\mathrm{att}_h^{semi} := d_{\mathcal{E}}\Big(\tilde{\mathcal{S}}_h(\tilde{\mathbf{x}}),\mu\big(\mathcal{S}_h(\mathbf{x}),\mathbf{x}\big)\Big) \ .$$

To emphasize the importance of this setting for explanation-aware attacks, we discuss two examples found in literature.

Example: Inverting Explanations. A model can be manipulated to yield inverted explanations, whenever a backdooring trigger is present on the input [118]. That is, regions with high relevance in the benign explanation receive low relevance in the attacked explanation and vice versa.

Example: Location Fooling. For input manipulations, the adversary may only be allowed to change a small region of the input only [22]. Without further provisions this region contributes highly to the prediction and an explanation method would highlight the manipulated region, revealing the attack. To go unnoticed, the manipulated area can be penalized for high absolute relevance values when generating the adversarial input [159].

TABLE 2: Works on explanation-aware attacks categorized by threat model (cf. Section 3), type of attack (cf. Section 4.1), and attack scope (cf. Section 4.2) including untargeted (⁙), semi-targeted (⁘) and targeted (⁖) attacks. Additionally, we specify whether the corresponding paper attack white box (□) or black box (■) post-hoc explanation method.

	Paper	Threat Model			Attack Type			Scope			XAI Methods	
		Input	Model	System	EP	PP	D	⁙	⁘	⁖	□	■
Attacks	Zhang et al. [198]	●	–	–	●	–	●	–	–	⁖	□	■
	Dombrowski et al. [49]	●	–	–	–	●	–	⁙	–	⁖	□	–
	Kuppa and Le-Khac [86]	●	–	–	–	●	●	–	–	⁖	□	–
	Dombrowski et al. [48]	●	–	–	–	●	–	–	–	⁖	□	–
	Ghorbani et al. [60]	●	–	–	–	●	–	⁙	–	–	□	–
	Fan et al. [53]	●	–	–	–	–	●	⁙	–	–	□	–
	Carbone et al. [28]	●	–	–	–	–	●	⁙	–	–	□	–
	Alvarez-Melis and Jaakkola [7]	●	–	–	–	●	–	⁙	–	–	□	■
	Wang et al. [176]	●	–	–	–	–	●	–	–	⁖	□	–
	Tamam et al. [162]	●	–	–	–	●	–	–	–	⁖	□	–
	Ivankay et al. [76]	●	–	–	–	●	–	⁙	–	–	□	–
	Sinha et al. [150]	●	–	–	–	●	–	⁙	–	–	□	■
	Abdukhamidov et al. [1]	●	–	–	–	–	●	–	–	⁖	□	■
	Sarkar et al. [140]	●	–	–	–	●	–	⁙	–	–	□	–
	Subramanya et al. [159]	●	–	–	–	–	●	–	⁘	–	□	–
	Tang et al. [163]	●	–	–	–	●	–	⁙	–	⁖	□	–
	Heo et al. [71]	–	●	–	–	●	–	⁙	⁘	⁖	□	–
	Wang et al. [177]	–	●	–	–	●	–	–	⁘	⁖	□	–
	Ali et al. [6]	–	●	–	–	●	–	⁙	–	–	□	■
	Zhang et al. [195]	–	●	–	–	●	–	–	–	⁖	□	–
	Slack et al. [153] (a)	–	●	–	–	●	–	⁙	–	–	–	■
	Slack et al. [153] (b)	○	●	–	–	●	–	⁙	–	–	□	–
	Slack et al. [152]	○	●	–	–	●	–	⁙	–	–	□	–
	Noppel et al. [118]	○	●	–	●	●	●	⁙	⁘	⁖	□	–
	Fang and Choromanska [54]	○	●	–	–	●	–	⁙	–	⁖	□	–
	Viering et al. [170]	○	●	–	–	●	–	–	–	⁖	□	–
Fairwashing	Anders et al. [8]	–	●	–	–	●	–	–	–	⁖	□	–
	Aïvodji et al. [3]	–	–	●	–	●	–	–	⁘	–	–	■
	Aïvodji et al. [4]	–	–	●	–	●	–	–	⁘	–	–	■
	Lakkaraju and Bastani [88]	–	–	●	–	●	–	–	⁘	–	–	■
	Slack et al. [151]	–	–	●	–	●	–	–	⁘	–	–	■

5. Explanation-Aware Robustness Notions

As discussed in the previous section, robustness against prediction-only attacks and explanation-aware attacks is strongly related. In this section, we zoom in on the robustness of post-hoc explanations and systematize the relation of various notions of robustness against explanation-aware attacks. This formalization serves as an important building block toward certifiable robust XAI-systems.

We focus on robustness at inference time with regard to an input $\mathbf{x}$ and its worst-case counterpart $\tilde{\mathbf{x}}$. However, a preceded model manipulation can benefit the adversary. In Section 5.1, we define templates for two families of robustness notions and introduce different constraints and restrictions for explanation-aware robustness. Combinations of restrictions and constraints allow us to define notions of varying strictness, which we then use to build a hierarchy of robustness notions in Section 5.2. On this basis and the gained understanding of robustness, we provide an outlook on how robustness can be guaranteed in Section 5.3.

5.1. Definitions

A robustness notion is defined by two conjugated sets, namely the *restrictions* (Section 5.1.1) and the *constraints* (Section 5.1.2). Both sets operate on input tuples $(\mathbf{x}, \tilde{\mathbf{x}})$, making recourse to the associated predictions and explanations. To name the notions, we use $\mathtt{R}|\mathtt{C}$ as a template where $\mathtt{R}$ stands for the restrictions (joined by +) and $\mathtt{C}$ for the specific constraint. Given this name, we use two robustness definitions with different scopes:

Definition 4 ($\mathtt{R}|\mathtt{C}$-Robustness around $\mathbf{x}$). — *We denote a system $\mathcal{S}$ as* $\mathtt{R}|\mathtt{C}$-robust around $\mathbf{x}$ *if*

$$\forall \tilde{\mathbf{x}} \in \mathcal{X} \; \textit{restrictions}(\mathbf{x}, \tilde{\mathbf{x}}) \rightarrow \; \textit{constraints}(\mathbf{x}, \tilde{\mathbf{x}}) \; .$$

This definition models an input's worst-case manipulation and can only be achieved or not achieved. Achieving the above robustness notion around *any* input $\mathbf{x}$ yields the following stronger and more general definition:

Definition 5 (R|C-Robustness). — *We denote a system $\mathcal{S}$ as R|C-robust if*

$$\forall \mathbf{x}, \tilde{\mathbf{x}} \in \mathcal{X} \ \textit{restrictions}(\mathbf{x}, \tilde{\mathbf{x}}) \rightarrow \ \textit{constraints}(\mathbf{x}, \tilde{\mathbf{x}}) \ .$$

To emphasize the distinction to *robustness around* $\mathbf{x}$ we occasionally denote this definition as *general robustness*. While robustness can also be measured as a scalar value, we refrain from doing so for the sake of simplicity. Hence, this general robustness can again be achieved or not achieved. In Appendix A, we review different variations for more fine-grained measures [28]. Moreover, we assume ($\mathcal{X}$,$d_\mathcal{X}$) and ($\mathcal{E}$,$d_\mathcal{E}$) to be metric spaces with their associated metric. This is applicable in most applications. However, for two specific findings, we additionally require fully-connectedness or convexity. We emphasize that both can be assumed whenever the space $\mathcal{X}$ consists of floating point numbers only.

In the following, we describe important constraints and restrictions used in literature to define explanation-aware robustness. In Appendix B, we provide an overview table of the constraints and restrictions we define in the following. Some of these definitions require additional parameters, e.g., a limit ϵ, a perturbation δ, or a Lipschitz constant K, which are always fix and part of the notion itself. The same holds true for the concretely applied metrics which are also considered parameters of the notion. *Consequently, two notions with different parameters are considered different notions.*

5.1.1. Constraints. Below, we define three constraints on explanations, common in literature to define explanation-aware robustness.

LIP$^{d_\mathcal{E},d_\mathcal{X},K}$: Lipschitz Continuity. The Lipschitz continuity requires the difference between two explanations to be less or equal to a fixed multiple of their distance. Hence, we define the Lipschitz continuity constraint as follows:

$$\exists \gamma \in \mathbb{R}^+ \ d_\mathcal{E}\big(h_\theta(\mathbf{x}), \gamma h_\theta(\tilde{\mathbf{x}})\big) \leq K d_\mathcal{X}(\mathbf{x}, \tilde{\mathbf{x}}) \ .$$

The positive scaling factor $\gamma \in \mathbb{R}^+$ ensures that we only compare relative differences in the explanations [163]. Note that this can also be modeled as a requirement on the explanation methods producing normalized explanations. Being Lipschitz continuous therefore is equivalent to having a bounded gradient. The smallest such bound is denoted as the Lipschitz constant K. We abbreviate notions that require Lipschitz continuity by LIP$^{d_\mathcal{E},d_\mathcal{X},K}$.

EXPLSIM$^{d_\mathcal{E},\epsilon}$: Explanation Similarity. We require the maximal difference between two explanations to be bounded by a fixed ϵ, formally given as

$$\exists \gamma \in \mathbb{R}^+ \ d_\mathcal{E}(h_\theta(\mathbf{x}), \gamma h_\theta(\tilde{\mathbf{x}})) \leq \epsilon \ .$$

We denote such notions as EXPLSIM$^{d_\mathcal{E},\epsilon}$.

EXPLEQ: Explanation Equivalence. Equivalence is a special case of the EXPLSIM$^{d_\mathcal{E},\epsilon}$ constraint with $\epsilon = 0$. We require

$$\exists \gamma \in \mathbb{R}^+ \ h_\theta(\mathbf{x}) = \gamma h_\theta(\tilde{\mathbf{x}}) \ .$$

The transitivity of the equivalence relation makes this constraint extremely strict in general robustness according to Definition 5. Hence, it is mostly applied together with a proper restriction to subspaces. It is mainly used for certified robustness as we discuss in Section 5.3. We denote notions that require explanation equivalence by EXPLEQ.

5.1.2. Restrictions. Restrictions define the subspace of tuples for which the constraints need to be satisfied. In the following, we motivate three different instantiations.

CLSEQ: Classification Equivalence. Suppose a model predicts two nearby samples as the class "`dog`" and "`cat`", respectively, despite their proximity. This may happen if the model is vulnerable to prediction-only attacks or the samples are simply close to the decision boundary. If the predicted class changes, we have to allow larger changes in the explanation as well because it answers a different question: *"Why is this cat?"* rather than *"Why is this a dog?"*. Consequently, we restrict to input tuples with identical predictions instead:

$$\mathcal{F}_\theta(\mathbf{x}) = \mathcal{F}_\theta(\tilde{\mathbf{x}}) \ .$$

We denote notions that apply this restriction by CLSEQ. Note that these subspaces are strictly smaller for every non-constant $\mathcal{F}_\theta$, but also neither necessarily convex nor fully-connected, even if $\mathcal{X}$ is convex. This fact is crucial for the hierarchy of robustness notions we present in Section 5.2.

LOC$^{d_\mathcal{X},\delta}$: Local Vicinity. Another restriction is to only consider tuples of nearby inputs:

$$d_\mathcal{X}(\mathbf{x}, \tilde{\mathbf{x}}) \leq \delta \ ,$$

where δ denotes the fixed limit of the distance. We denote those notions as LOC$^{d_\mathcal{X},\delta}$.

No Restrictions. The last option is to have no restriction at all. Obviously, this is only reasonable if the constraint somehow considers the distance between the inputs, e.g., LIP$^{d_\mathcal{E},d_\mathcal{X},K}$. We consider this setting the default and use no specific symbol to denote it.

5.2. Hierarchy of Robustness Notions

Next, we set up the hierarchy of robustness notions and provide intuitions on the relations between them as a starting point to unify the field. We visualize the resulting implications in Fig. 2. For the sake of the notion's simplicity, we drop superscripts whenever possible and not relevant for understanding. For instance, we write LIP instead of LIP$^{d_\mathcal{E},d_\mathcal{X},K}$. Further, we assume reasonable assignments of the respective parameters, like the used distance metrics.

Arrows denote a logical implication from a stricter to a weaker notion, and the stricter the notion, the higher the robustness against attacks. The two dashed arrows are only valid for general robustness and require additional assumptions, like the convexity of $\mathcal{X}$. Intuitively they then arise from the transitive properties of EXPLEQ and LIP. The gray boxes indicate notions that are too strict to be relevant: ❶ EXPLEQ-robustness requires every two explanations to be equivalent, which is not useful in practice, of course.

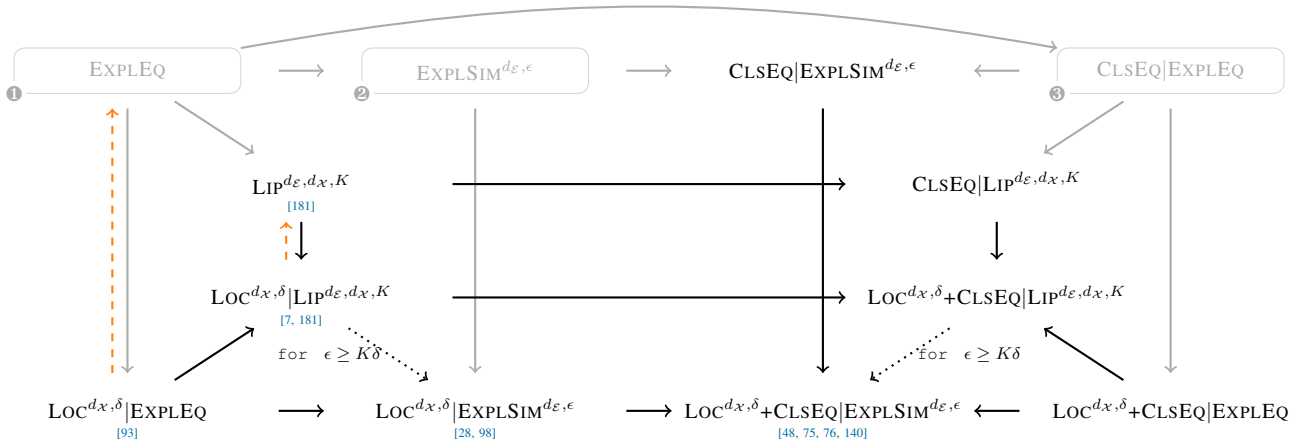

Figure 2: Hierarchy of explanation-aware robustness notions. Arrows denote implications from strict to a weak notions: the stricter the notion, the higher the robustness. Dashed arrows are only valid for general robustness, and require fully-connectedness or even convexity. Dotted arrows are only valid for specific combinations of parameters. Notions ❶, ❷, and ❸ are considered too strict to be relevant.

❷ ExplSim-robustness requires that explanations differ only slightly. Hence, explanations must be either pairwise similar (general robustness) or every explanation must be similar to $h_\theta(\mathbf{x})$ (robustness around $\mathbf{x}$). Both limits the usability of explanations dramatically.

> **Observation.** According to the triangle inequality every system that is ExplSim$^{d_\mathcal{E},\epsilon}$-robust around $\mathbf{x}$ is also generally ExplSim$^{d,2\epsilon}$-robust.

❸ ClsEq|ExplEq-robustness can only be satisfied if there is only one explanation per class. ClsEq|ExplEq-robustness around $\mathbf{x}$ is weaker and requires each sample of the class $\mathcal{F}_\theta(\mathbf{x})$ to produce the same explanation as $\mathbf{x}$. We consider both as not relevant in practice.

While some other notions are also very strict, they might still be helpful for further analysis. We discuss their relation and relevance in the following.

Details on the Relations. ExplEq-robustness and Loc|ExplEq-robustness are equivalent if $\mathcal{X}$ is fully-connected. The reason is that the equivalence relation transitive. Interestingly, Loc +ClsEq|ExplEq-robustness and ClsEq|ExplEq-robustness are not equivalent because the subspace restricted by ClsEq might not be fully-connected even if $\mathcal{X}$ is fully-connected. The other way around the implication trivially holds.

Our argumentation for the equivalence between Lip-robustness and Loc|Lip-robustness is slightly different and even requires $\mathcal{X}$ and $\mathcal{E}$ to be convex.

Lemma 1. — *Given two convex metric space $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{E}, d_\mathcal{E})$. Every* Loc$^{d_\mathcal{X},\delta}$|Lip$^{d_\mathcal{E},d_\mathcal{X},K}$*-robust system $(\mathcal{F}_\theta, h_\theta)$, is also* Lip$^{d_\mathcal{E},d_\mathcal{X},K}$*-robust:*

$$\text{Loc}^{d_\mathcal{X},\delta}|\text{Lip}^{d_\mathcal{E},d_\mathcal{X},K} \Rightarrow \text{Lip}^{d_\mathcal{E},d_\mathcal{X},K} \ .$$

Proof. We assume two points $\mathbf{x}_0, \mathbf{x}_2 \in \mathcal{X}$ such that $d_\mathcal{X}(\mathbf{x}_0, \mathbf{x}_2) > \delta$. W.l.o.g. we assume $d_\mathcal{X}(\mathbf{x}_0, \mathbf{x}_2) < 2\delta$. As $\mathcal{X}$ is convex, there exists an equidistant point $\mathbf{x}_1 \in \mathcal{X}$, that lies directly between $\mathbf{x}_0$ and $\mathbf{x}_2$ and satisfies the triangle equation with equality:

$$d_\mathcal{X}(\mathbf{x}_0, \mathbf{x}_2) = d_\mathcal{X}(\mathbf{x}_0, \mathbf{x}_1) + d_\mathcal{X}(\mathbf{x}_1, \mathbf{x}_2) \tag{1}$$

and

$$d_\mathcal{X}(\mathbf{x}_0, \mathbf{x}_1) < \delta \text{ and } d_\mathcal{X}(\mathbf{x}_1, \mathbf{x}_2) < \delta \ .$$

Hence, $\mathbf{x}_1$ lies in the δ-neighborhood of $\mathbf{x}_0$ and $\mathbf{x}_2$. Due to the fact that the system is Loc$^{d_\mathcal{X},\delta}$|Lip$^{d_\mathcal{E},d_\mathcal{X},K}$-robust, we can pose the following two equations:

$$\begin{aligned} \exists\gamma \ d_\mathcal{E}(h_\theta(\mathbf{x}_0), \gamma h_\theta(\mathbf{x}_1)) &\leq K \cdot d_\mathcal{X}(\mathbf{x}_0, \mathbf{x}_1) \\ \exists\gamma \ d_\mathcal{E}(h_\theta(\mathbf{x}_1), \gamma h_\theta(\mathbf{x}_2)) &\leq K \cdot d_\mathcal{X}(\mathbf{x}_1, \mathbf{x}_2) \end{aligned}$$

Adding both together, applying the triangle equation, and Eq. (1) results in

$$\exists\gamma \ d_\mathcal{E}(h_\theta(\mathbf{x}_0), \gamma h_\theta(\mathbf{x}_2)) \leq K \cdot d_\mathcal{X}(\mathbf{x}_0, \mathbf{x}_2)$$

□

Intuitively the proof relies on $\mathcal{X}$ being convex and the norm satisfying the triangular inequality with equality for the existing equidistant input in the center. Note that this argumentation does not hold for robustness around a fixed $\mathbf{x}$. Further ClsEq|Lip-robustness implies Loc+ClsEq|Lip-robustness, but not the other way around as the ClsEq subspace might not be convex.

Lemma 2. — *Every* Loc$^{d_\mathcal{X},\delta}$|Lip$^{d_\mathcal{E},d_\mathcal{X},K}$*-robust system $(\mathcal{F}_\theta, h_\theta)$ is also* Loc$^{d_\mathcal{X},\delta}$|ExplSim$^{d_\mathcal{E},\epsilon}$*-robust for $\epsilon \leq K\delta$:*

$$\text{Loc}^{d_\mathcal{X},\delta}|\text{Lip}^{d_\mathcal{E},d_\mathcal{X},K} \Rightarrow \text{Loc}^{d_\mathcal{X},\delta}|\text{ExplSim}^{d_\mathcal{E},\epsilon} \ .$$

Proof. In order to show $A \to B \Rightarrow A \to C$ it is sufficient to show that $A \to (B \to C)$. Hence, we emphasize that the following

$$\begin{aligned} \exists\gamma \ d_\mathcal{E}(h_\theta(\mathbf{x}_0), \gamma h_\theta(\mathbf{x}_1)) &\leq K d_\mathcal{X}(\mathbf{x}_0, \mathbf{x}_1) \\ &\Rightarrow \\ \exists\gamma \ d_\mathcal{E}(h_\theta(\mathbf{x}_0), \gamma h_\theta(\mathbf{x}_1)) &\leq \epsilon \end{aligned}$$

holds for any choice of $\mathbf{x}_0, \mathbf{x}_1 \in \mathcal{X}$ if

$$\epsilon \geq K d_{\mathcal{X}}(\mathbf{x}_0, \mathbf{x}_1) \ .$$

Under the restriction $d_{\mathcal{X}}(\mathbf{x}_0, \mathbf{x}_1) \leq \delta$ this holds exactly if

$$\epsilon \geq K\delta \ .$$

□

Intuitively, if $h_\theta(\cdot)$ is Lipschitz continuous within a certain radius δ, then the maximal explanation dissimilarity within the same ball is $\epsilon = K\delta$. The same idea applies for:

$$\text{LOC+CLSEQ|LIP} \Rightarrow \text{LOC+CLSEQ|EXPLSIM}$$

In the other direction, however, both last findings do not apply. Intuitively speaking, the reason is that $\text{LIP}^{d_\mathcal{E}, d_\mathcal{X}, K}$ requires an infinitesimal small difference in the explanations of infinitesimal nearby inputs. Hence, the required difference can always be forced to be below *any* ϵ by picking two inputs that are just close enough together. Further, for $\text{EXPLSIM}^{d_\mathcal{E}, \epsilon}$ the transitivity argument, as applied above, does not hold and hence

$$\begin{aligned} \text{EXPLSIM} &\Rightarrow \text{LOC|EXPLSIM} \\ \text{CLSEQ+EXPLSIM} &\Rightarrow \text{LOC+CLSEQ|EXPLSIM} \end{aligned}$$

holds, while it does not necessarily hold the other way around. The argumentation for the relations within *robustness around* $\mathbf{x}$ are similar, except for the fact that transitivity can not be applied. In Fig. 2, we indicate relations that only hold for *general robustness*, and fully-connected or convex spaces as dashed arrows. Not explicitly shown relations are considered as trivial.

Existing Notions in Literature. Our hierarchy arises from the complete combination of the building blocks proposed in Section 5.1. Since these notions have varying relevance to the community, in the following, we provide links to notions proposed in related work:

(a) Wang et al. [181] propose LIP and LOC|LIP-robustness around $\mathbf{x}$ as *Attribution(al) Robustness*,

(b) Ivankay et al. [75, 76] and Sarkar et al. [140] use the same term but refer to LOC +CLSEQ|EXPLSIM-robustness around $\mathbf{x}$.

(c) Levine et al. [98] work with the LOC|EXPLSIM-robustness around $\mathbf{x}$, but without naming it.

(d) Dombrowski et al. [48] prove an upper bound on ϵ for $\text{LOC}^{d_\mathcal{X}, \delta}$ +CLSEQ|EXPLSIM-robustness around $\mathbf{x}$ using the maximal principle curvature. They choose the parameter δ such that the union of the neighborhood around the fixed $\mathbf{x}$ and the hyperspace of equal classification is fully-connected.

Observation. Inferring *robustness* from *robustness around* $\mathbf{x}$ is in general non-trivial. Determining a reasonable δ that holds around *any* $\mathbf{x}$ can be impossible, as illustrated in Fig. 3. Intuitively speaking, the fully-connectedness forces δ to be upper bounded at point a, but also lower bounded at point b. The only solution is $\delta = 0$, which however is an unreasonable notion.

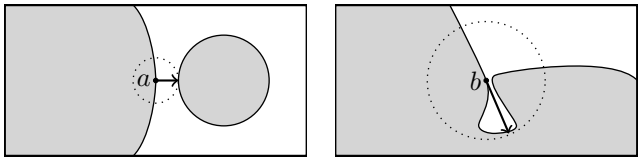

Figure 3: Geometric intuition in a two-class problem that mapping *robustness around* $\mathbf{x}$ to *robustness* is not trivial in general. The system is $\text{LOC}^{d_\mathcal{X}, \delta}$+CLSEQ|EXPLSIM-robust around a and b for different values of δ. Finding a common value δ to satisfy general $\text{LOC}^{d_\mathcal{X}, \delta}$+CLSEQ|EXPLSIM-robustness is impossible.

Moreover, various authors motivate that explanations should only be compared by rank and hence the top-k intersection or the Kendall correlation [81] should be considered [28, 35, 60, 75, 98, 140, 163, 174, 175]. We emphasize that k in this case is way bigger than a usual norm limit ϵ and also that the distance metrics' required properties might not be satisfied. Still, those ideas can be mapped to versions of $\text{EXPLSIM}^{d_\mathcal{E}, \epsilon}$ or EXPLEQ-robustness.

Link to Attack Objectives. Our robustness notions are linked to specific attack objectives (Section 4). Let's consider a concrete example for clarification: A defender that expects a prediction-preserving adversary may strive for a notion with the CLSEQ restriction (right side of Fig. 2) because the considered attack keeps the classification intact anyway. Consequently, explanation-aware robustness needs to be preserved within the same class only.

However, these CLSEQ notions are too weak to simultaneously counter a dual adversary (Table 1), who is willing to change the classification as well. Consequently, the defender should rather pick the notions *without* the CLSEQ restriction (left side of Fig. 2). **But**, any adversary that strives for maintaining the benign explanation, while attacking the prediction (*explanation-preserving attacks*) still outmaneuvers the above defender. Even worse, by implementing any of our robustness notions, the defender actually *supports* the adversarial objectives of an explanation-preserving adversary.

To counter explanation-preserving adversaries additional requirements are necessary, e.g., explanations should indicate when a class flip occurs. This behavior can be trivially achieved by requiring the explanation's changes to exceed a certain threshold whenever $\mathbf{x}$ and $\tilde{\mathbf{x}}$ yield different predictions. Unfortunately, this would then benefit an untargeted dual adversary, leaving the defender in yet another dilemma.

Observation. A defender can adjust to expected adversaries, fulfilling a specific robustness notion. However, the required properties might be conflicting for multiple adversaries with diverging objectives.

This observation leads to the following research question:

Open RQ 1. To which extent do different threat models and attack objectives conflict in their required robustness notions?

5.3. Robustness Guarantees

So far, we have identified various robustness notions. However, we aim for guarantees that a certain notion is satisfied, i.e., we want *Certified Robustness*. Instead of simply proving a notion, most often the exact guaranteed notion parameters (K, ϵ, or δ) are calculated in practice as every instantiation of the parameters yields a different notion. In the following, we present two examples of how this goal can be accomplished in practice.

Example: Linear Regions. Given an input $\mathbf{x}$, the system operator can calculate the maximal norm ball around $\mathbf{x}$ that fits within the corresponding linear region of a ReLU network [93]. Therefore, the active linear segments of the activation function of each neuron are collected in so-called *Activation Patterns* [125]. Each activation pattern corresponds to one linear region of the decision surface. The trick is that each region is convex, as it is encircled by lines. Consequently, the maximal norm ball that fits in can be approximated efficiently for L_1 and L_2 [93]. However, this guarantee does not apply for every explanation method. In fact, the only guarantee is that the gradient is constant within the norm ball. As a matter of fact, for the Simple Gradients explanation method this corresponds to a constant explanation and hence a guarantee for LOC|EXPLEQ-robustness around $\mathbf{x}$.

> **Open RQ 2.** How to provide guarantees for the robustness around $\mathbf{x}$ for explanation methods, other than gradient-based methods?

Example: Randomized Smoothing. Levine et al. [98] demonstrate a probabilistic robustness guarantee via randomized smoothing [40]. Therefore, the model is sampled with noisy inputs and depending on the results an overlap of the top-k features can be probabilistically guaranteed in a norm ball with fixed radius. This matches a $\text{LOC}^{d_{\mathcal{X}},\delta}|\text{EXPLSIM}^{d_{\mathcal{E}},\epsilon}$ notion, where δ is fixed, and the explanation distance is set to the guaranteed top-k overlap.

> **Open RQ 3.** How to generalize guarantees from randomized smoothing to other dissimilarity metrics and explanation methods, e.g., other than gradient-based methods?

6. Robust Models for XAI-Systems

The research field on robust models is rather novel, albeit being heavily inspired by works on prediction-only attacks. We start by discussing the sanitization and validation of training data (Section 6.1) and the model (Section 6.2). Then we present certain robust model architectures (Section 6.3) and elaborate on possibilities to increase the robustness during training (Section 6.4).

6.1. Training Data Validation and Sanitization T1

Errors, spurious correlations, and intentionally poisoned data can lead to misbehaving models [9, 38, 92, 158]. Hence, for training a model from scratch, the training data needs to be validated first. Unfortunately, we are not aware of any work that discusses data sanitization and validation specifically to prevent explanation-aware attacks. To prevent prediction-only attacks, in turn, a huge body of work proposes to sanitize and validate data [34, 157, 167, 200]. Some of them, however, propose techniques based on explanation methods [9, 37, 47, 92, 179].

Example: Class Artifact Compensation. Explanations are generated for each training sample, clustered, and aggregated per cluster to highlight regions that cause spurious correlations [9]. This method heavily relies on the correctness of explanations, which raises the question:

> **Open RQ 4.** To what extent can explanations be fooled in a data manipulation setting to bypass sanitization? In particular, bypassing explanation-based data sanitization techniques, seems to be a practicable goal.

6.2. Model Sanitization and Validation T2

Downloaded models and models that have been trained by MLaaS providers need to be sanitized and validated before deployment [54, 71, 71, 118, 152, 153]. However, this process requires additional resources like a small dataset that is guaranteed to be clean [101].

Sanitization of Manipulated Models. Model sanitization should be applied prophylactically whenever the integrity of the training process is questionable. The benign performance is barely affected [173]. To the best of our knowledge, there is no work on the sanitization specifically for explanation-aware attacks. However, many authors discuss sanitization to prevent prediction-only attacks [99, 101, 185, 193]. Concretely, Liu et al. [101] propose an alternation between finetuning and pruning steps on clean data, denoted as finepruning. That way, the model is forced to forget a backdooring trigger it learned previously. We expect these techniques to also remove injected explanation-aware-backdoors. However, its effectiveness against prediction-preserving model manipulations is not clearly answerable.

Detecting Manipulated Models. Prediction-preserving adversaries resemble a new type model manipulation threat specific to XAI-systems, which is not present in prediction-only settings [71, 152, 153]. For instance, Heo et al. [71] demonstrate that the explanations of two classes can be swapped or, alternatively, a model can be manipulated to always show a specific target explanation, for any input. However, explanation-aware backdooring attacks have been proposed in all three distinct flavors of explanation-aware attacks by various authors [15, 54, 117, 118]. However, to the best of our knowledge there is no work that deeply investigates the detection of such explanation-aware attacks. Albeit,

there are plenty of works on the detection of prediction-only attacks [34, 57, 74, 168, 173, 185, 188, 189]. They identify anomalies in weight distributions [57, 185] and neuron activation [34], or measure distances to potential target classes [173]. We strongly assume that similar techniques would work for explanation-aware model manipulations as well. We pose the following open research question:

> **Open RQ 5.** To what extent are recent model sanitizations and validations effective against explanation-preserving, dual, and in particular, prediction-preserving attacks?

6.3. Robust Architectures T3

Every robust XAI-system starts with choosing a robust model architecture. Recent research shows that some design choices favor robustness and explainability more than others [27, 28, 48, 49, 136]. In the following, we briefly present two directions.

Smooth Activation Functions. The ReLU activation function induces kinks in the decision surface whenever the zero-line is crossed. These kinks lead to a sudden increase in gradients that many explanation methods heavily pick up on. Smooth approximations of ReLU, e.g., the Softplus-β, can reduce kinks and allow for a more smooth transition in explanations for similar inputs [48, 49]. The β parameters specifies how close ReLU is approximated: Infinite β equals ReLU, while low values render loose approximations [48].

Bayesian Networks. Recent work suggests that Bayesian neural networks provide significantly more robustness against adversarial attacks [27, 28]. According to Carbone et al. [28], this is also true for explanation-aware attacks and Bykov et al. [24] demonstrate how these bayesian networks can be explained with LRP. Supplementing the importance scores with uncertainty scores helps to make the explanation more robust, or at least inform the user about a high uncertainty.

6.4. Robust Training T4

Non-smooth decision surfaces can induce significant gradient changes, thus a small principle curvature is a strong indicator for a robust model [49, 114]. Ideally, we aim for a small curvature during the training process [49]. In the following, we present various training approaches leading to models more robust against explanation-aware attacks.

Double Backpropagation. Double backpropagation refers to penalizing large second derivatives and is known for improving generalization [50, 72]. It is formalized by adding a λ-weighted term $\lambda\|\frac{\partial^2 f_\theta}{\partial \mathbf{x}^2}\|$ to the loss function. A major disadvantage is the high computational effort, though.

Regularization of the Hessian Matrix. The Hessian matrix $\mathbf{H}$ contains the second order partial mixed derivatives and defines the local curvature of the decision surface. The integral of the Frobenius norm of the Hessian $\|\mathbf{H}\|_F$ on the path between $\mathbf{x}$ and $\tilde{\mathbf{x}}$ bounds the difference in their explanation [49]:

$$d_{\mathcal{E}}(h_\theta(\mathbf{x}), h_\theta(\tilde{\mathbf{x}})) \leq \int_{-\infty}^{+\infty} \|\mathbf{H}_{\mathcal{F}}(\tau_{\mathbf{x}}(t))\|_F \ \mathrm{d}t \ ,$$

where τ defines a path between $\mathbf{x}$ and $\tilde{\mathbf{x}}$ such that $\tau(-\infty) = \mathbf{x}$ and $\tau(+\infty) = \tilde{\mathbf{x}}$. This bound is shown for simplified Simple Gradients explanations [147], without aggregation or absolute values [49]. Moreover, the fundamental theorem of calculus for line integrals needs to hold for h_θ [160]. However, the Hessian matrix is expensive to calculate, making an approximation necessary [49].

Regularize Weights and Weight Decay. This technique is proven to improve the generalization of neural networks [69, 85, 183, 194]. It adds a λ-weighted penalization for the norm of the weights, formalized as $\lambda\|w\|$, where w represents the weight components of a model's parameters. Regularizing the weights of a neural network using the Frobenius norm $\|\cdot\|_F$ amounts to bounding $\mathbf{H}$. This has the effect of smoothing the decision surface and makes attacks against gradient-based explanations harder [49].

Regularizing the Largest Eigenvalues. *Smooth Surface Regularization* (SSR) [181] can be used to minimize the distance between explanations of nearby points. This is done through penalizing the largest eigenvalue of the $\mathbf{H}$ with respect to all data samples from a training distribution. The authors formalize this as a λ-weighted term in the loss function: $\lambda \max_i |\xi_i|$. Moreover, Singla et al. [149] propose a closed-form solution to efficiently find the eigenvalues in ReLU networks and demonstrate that the largest eigenvalue is almost parallel to the gradient of the loss.

Adversarial Training for Explanations (ATEX). Vanilla adversarial training for explanations is computational expensive due to the necessary multiple derivations. ATEX [163] uses an efficient way to approximate the required changes by relying on the fact that the prediction and the explanation changes are orthogonal to each other.

Explanation-Based Optimization (ExpO). Plumb et al. [122] generalize the loss function to include a λ-weighted term $\lambda R(\mathbf{x}, \mathcal{N}_\mathbf{x})$ that depends on the neighborhood of $\mathbf{x}$. They propose two instantiations of R. ExpO-Fidelity defines $R(\mathbf{x}, \mathcal{N}_\mathbf{x})$ as

$$R(\mathbf{x}, \mathcal{N}_\mathbf{x}) := \mathbb{E}_{\mathbf{x}' \sim \mathcal{N}_\mathbf{x}}[(\mathcal{F}_\theta(\mathbf{x}') - \mathbf{w}^T\mathbf{x}' + \mathbf{b})^2] \ ,$$

where $\mathbf{w}$ and $\mathbf{b}$ are weights and biases learned on $\mathcal{N}_\mathbf{x}$. Hence, it measures how well each neighborhood can be approximated linearly. ExpO-Stability defines $R(\mathbf{x}, \mathcal{N}_\mathbf{x})$ as

$$R(\mathbf{x}, \mathcal{N}_\mathbf{x}) := \mathbb{E}_{\mathbf{x}' \sim \mathcal{N}_\mathbf{x}}[\|h_\theta(\mathbf{x}), h_\theta(\mathbf{x}')\|_2^2] \ .$$

It measures how stable the explanations are in $\mathcal{N}_\mathbf{x}$. Summarizing, we pose the following research question:

> **Open RQ 6.** What are effective and efficient regularization techniques for general explanation methods? How can we reduce the required computational efforts of current approaches?

7. Robust Operation of XAI-Systems

Training time defenses ideally are complemented by defenses and monitoring during operation [124]. Hence, we consider the sanitization and detection of malicious inputs as initial defense (Section 7.1). Additionally, potentially malicious inputs can be collected and used to re-validate the model continuously during operation (Section 7.2). Finally, explanations can be used to verify the input-output behavior of black box systems (Section 7.3).

7.1. Input Sanitization and Validation O1

An upstream algorithm can intercept and sanitize any submitted input and, depending on the application scenario, detect and reject inputs with malicious components [80, 166]. We detail both approaches in the following.

Sanitization of Malicious Inputs. To the best of our knowledge, no sanitization has been evaluated specifically for explanation-aware attacks yet. However, research on its effectiveness against prediction-only attacks is omnipresent [19, 64, 116, 186]. Proposed techniques apply denoising or partial impainting through generative models, e.g., variational autoencoder decoder (VAE) [64], generative adversarial models (GAN) [70], or diffusion models [19, 116, 186]. In the hope of generalizing on future adversarial perturbations, the model is trained to reconstruct images from diverse perturbations, including adversarial ones. As explanation-aware attacks require a higher or comparable level of perturbation, we assume a great transferability towards explanation-aware attacks [20]. Note that we exclude methods from this assumption that heavily rely on robust explanations. As an example, Februus [47] uses GradCAM explanations to determine the position of backdoor triggers, and thus can be bypassed by explanation-aware attacks [118].

Detecting and Rejecting Adversarial Inputs. Detection of adversarial inputs can be based on three (potentially overlapping) approaches: (1) Indicators of known attack classes in the input [42, 63, 95, 105, 107, 135], (2) effects of extracted components on clean data [37], and (3) side effects during processing, e.g., uncommon activations in intermediate layers [164, 178]. We assume the same detectability for explanation-aware attacks, except for detection techniques that utilize explanations for that task [37, 47, 55, 80, 178]. Adversaries that are aware of the vulnerabilities of explanations might easily evade these detection techniques. On the other side, a strong robust explanation can contribute significantly to detecting malicious inputs, e.g., explanations highlight backdooring trigger [37, 118]. Summarizing this subsection, we pose the following open research question:

> **Open RQ 7.** How do the introduced techniques perform in sanitizing and detecting explanation-aware input manipulations, specifically for prediction-preserving attacks, where only the explanation is attacked? To what extent can explanation-aware attacks be utilized to bypass recent explanation-based detection techniques?

7.2. Continuous Model Monitoring O2

In addition to the input sanitization and validation, the model must be monitored during deployment and regularly re-validated with the potentially malicious inputs [168]. The difference to O1 is that here we do not decide on single individual inputs, but sets of potentially malicious inputs collected over time. The model may be corrupted early on during training already by full or partial replacement [124] or even hardware side channels [128].

7.3. Validating the Input-Output Behavior O3

The input-output behavior of a prediction-only system only consists of the input and the one dimensional label. In contrast, XAI-system have an additional high-dimensional output in form of the explanation. This output gives outsiders in general more information to audit the system, e.g., verifying its fairness [30, 87, 96]. But also other properties like the used model architecture, if the explanations are fitting to the decision-making, or the number of neural networks might be inferable by querying the XAI-system properly. We pose the following research question:

> **Open RQ 8.** To what extent and how can public authorities audit and verify the inner working of an explainable system to ensure a fair decision-making?

8. Robust Explanation Methods

Beside using robust models and monitoring the system during operation, the robustness of the explanation method itself must be enhanced [8, 131, 181]. In the following, we present different directions.

Smooth Adaptations. Smoothed explanations are less vulnerable to input manipulations [5, 48] and can be yield in various ways. SmoothGrad aggregates the gradients over multiple noisy inputs in the vicinity of the input [154]. Beside Gaussian noise, like in SmoothGrad, others propose to use uniform noise, denoted as UniGrad [181]. NoiseGrad [25], in turn, stochastically perturbs the model weights instead of the inputs. In practice, there exists a runtime trade-off inflicted by the number averaged samples. These approaches are applicable to every explanation method and model type.

However, using a smooth activation functions as discussed in Section 6.3 are advantageous over smooth adaptions [5, 48]. Smooth activation only require one forward and backward pass through the model. On the other hand, smooth adaptions require one forward and backward pass per sample or weight perturbation.

Adversarially Trained Surrogate Models. Lakkaraju et al. [90] propose a black box method that additionally performs adversarial training on the surrogate model. As a result, the surrogate model will resemble the black box model's functionality for out-of-distribution samples and, thus, becomes more robust.

Ensembles of Attribution Methods. Instead of using a single explanation method, ensembles of (multiple) aggregated explanation methods are more robust against attacks [131]. This is because attacks are usually targeted toward specific explanation methods and are generally not transferable to other explanation methods. A certain degree of transferability can be observed between methods from the same family or if the attack is specifically targeting the transferability [118, 131].

Tangent Space Projections (TSP). Tangent space projected explanations post-process explanations by projecting them along tangential directions of the data manifold [8]. This method is theoretically motivated by differential geometry and manifold learning and motivated by the fact that the relevant training data only lie on a low-dimensional small submanifold in the manifold of all possible data points. As these TSPs only rely on the dataset, they are also effective against model manipulations.

Certifiable Robustness. Research on the robustness of explanation methods focuses on the gradient $\nabla_{\mathbf{x}} \mathcal{F}_\theta(\mathbf{x})$ for which many interesting results have been found [48, 49, 163]. However, these results do not directly apply to explanation methods used in practice due to subtle differences. As an example, Simple Gradients [147] uses the gradient's absolute value and aggregates it to yield one importance score per pixel. Also, while all gradient-based explanations use the model's gradients, it is not clear whether they are equally vulnerable. Introducing axioms for explanation methods, such as sensitivity, completeness, and implementation invariance, enables us to generalize robustness better. One primary result is the importance of the completeness axiom which corresponds to the fundamental theorem of line integrals

$$\int_{-\infty}^{+\infty} \nabla h_\theta(\tau(t)) \, \mathrm{d}t = h_\theta(\mathbf{x}) - h_\theta(\tilde{\mathbf{x}}) \ ,$$

where τ is again a path from $\mathbf{x}$ to $\tilde{\mathbf{x}}$. Many recent robustness proofs rely on this axiom to hold [20, 35, 75, 148]. However, more axioms exist, such as relevance conservation, positivity, and continuity [12, 110] and many explanation methods have been proposed without pointing out which axioms they satisfy. This drawback hinders research and renders the provided guarantees inaccessible.

> **Observation.** Robustness proofs need to be decoupled from specific explanation methods. Fundamental axioms such as sensitivity, completeness, and implementation invariance [160], or relevance conservation, positivity, and continuity [12, 110] can serve as building blocks for describing explanation techniques.

While the implications of an explanation method's completeness are well studied already, other axioms have not been investigated in this intensity yet. Explanation methods need to be mapped to the axioms they satisfy, specifying the applying conditions. These axioms then allow to pinpoint the robustness guarantees for the explanation method at hand in a specific setting.

9. Conclusion

We systematize attacks against post-hoc explanation methods along the adversary's capabilities, constraints, and objectives. We reveal relations of different classes of explanation-aware attacks and even find similarities to applications that use explanations to improve learning. Interestingly, these works use techniques related to data poisoning for model manipulation, raising the question whether and how they correlate.

Moreover, we formalize robustness notions for post-hoc explanations that (a) highlight the requirements for defenses and (b) enable the community to unify their efforts. Our taxonomy of existing defenses shows that plenty of research on mitigating prediction-only attacks exists, but methods for fending off explanation-aware attacks are scarce. It is not immediately apparent whether defenses for prediction-only attacks work for explanation-aware attacks out-of-the-box. Perhaps even more importantly, the community actively works on developing robust post-hoc explanation methods and investigates the certified robustness of explanations. Both directions are auspicious as many underlying concerns in applying explanations can be addressed if explanations methods are provable robust and reliable.

We are confident that our systematization and the made observations, the hierarchy of robustness notions as a foundation for understanding robustness requirements, and the highlighted future research directions will help advance the field toward robust post-hoc explanation methods.

Acknowledgement

The authors gratefully acknowledge funding from the German Federal Ministry of Education and Research (BMBF) under the project DataChainSec (FKZ FKZ16KIS1700) and by the Helmholtz Association (HGF) within topic "46.23 Engineering Secure Systems".

References

[1] E. Abdukhamidov, M. Abuhamad, S. S. Woo, E. Chan-Tin, and T. Abuhmed, "Interpretations cannot be trusted: Stealthy and effective adversarial perturbations against interpretable deep learning," *CoRR*, vol. abs/2211.15926, 2022.

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[3] U. Aïvodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, and A. Tapp, "Fairwashing: The risk of rationalization," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 97, 2019, pp. 161–170.

[4] U. Aïvodji, H. Arai, S. Gambs, and S. Hara, "Characterizing the risk of fairwashing," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[5] A. Ajalloeian, S.-M. Moosavi-Dezfooli, M. Vlachos, and P. Frossard, "On smoothed explanations: Quality and robustness," in *Proc. of ACM International Conference on Information and Knowledge Management (CIKM)*, 2022, pp. 15–25.

[6] H. Ali, M. S. Khan, A. I. Al-Fuqaha, and J. Qadir, "Tamp-X: Attacking explainable natural language classifiers through tampered activations," *Comput. Secur.*, vol. 120, p. 102791, 2022.

[7] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," *Proc. of the ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2018.

[8] C. J. Anders, P. Pasliev, A.-K. Dombrowski, K.-R. Müller, and P. Kessel, "Fairwashing explanations with off-manifold detergent," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 119, 2020, pp. 314–323.

[9] C. J. Anders, L. Weber, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin, "Finding and removing clever hans: Using explanation methods to debug and improve deep models," *Information Fusion*, vol. 77, pp. 261–295, 2022.

[10] D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro, and K. Rieck, "Dos and don'ts of machine learning in computer security," in *Proc. of the USENIX Security Symposium*, Aug. 2022.

[11] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek, ""what is relevant in a text document?": An interpretable machine learning approach," *PLOS ONE*, 2016.

[12] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, p. 46, 2015.

[13] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *Journal of Machine Learning Research (JMLR)*, vol. 11, pp. 1803–1831, 2010.

[14] M. Bafna, J. Murtagh, and N. Vyas, "Thwarting adversarial examples: An l_0-robust sparse fourier transform," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[15] E. Bagdasaryan and V. Shmatikov, "Blind backdoors in deep learning models," in *Proc. of the USENIX Security Symposium*, 2021, pp. 1505–1521.

[16] H. Baniecki and P. Biecek, "Adversarial attacks and defenses in explainable artificial intelligence: A survey," in *Proc. of the IJCAI Workshop of explainable AI (XAI)*, 2023.

[17] O. Barkan, E. Hauon, A. Caciularu, O. Katz, I. Malkiel, O. Armstrong, and N. Koenigstein, "Grad-SAM: Explaining transformers via gradient self-attention maps," in *Proc. of ACM International Conference on Information and Knowledge Management (CIKM)*, 2021, pp. 2882–2887.

[18] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proc. of the International Conference on Machine Learning (ICML)*, 2012.

[19] T. Blau, R. Ganz, B. Kawar, A. M. Bronstein, and M. Elad, "Threat model-agnostic adversarial defense using diffusion models," *CoRR*, vol. abs/2207.08089, 2022.

[20] A. Boopathy, S. Liu, G. Zhang, C. Liu, P.-Y. Chen, S. Chang, and L. Daniel, "Proper network interpretability helps adversarial robustness in classification," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 119, 2020, pp. 1014–1023.

[21] D. Brown and H. Kvinge, "Brittle interpretations: The vulnerability of TCAV and other concept-based explainability tools to adversarial attack," *CoRR*, vol. abs/2110.07120, 2021.

[22] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *CoRR*, vol. abs/1712.09665, 2017.

[23] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.

[24] K. Bykov, M. M.-C. Höhne, K.-R. Müller, S. Nakajima, and M. Kloft, "How much can i trust you? - quantifying uncertainties in explaining neural networks," *CoRR*, vol. abs/2006.09000, 2020.

[25] K. Bykov, A. Hedström, S. Nakajima, and M. M.-C. Höhne, "NoiseGrad: Enhancing explanations by introducing stochasticity to model weights," in *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2022, pp. 6132–6140.

[26] C. J. Cai, J. Jongejan, and J. Holbrook, "The effects of example-based explanations in a machine learning interface," in *Proc. of the International Conference on Intelligent User Interfaces (IUI)*, 2019, pp. 258–262.

[27] G. Carbone, M. Wicker, L. Laurenti, A. Patané, L. Bortolussi, and G. Sanguinetti, "Robustness of bayesian neural networks to gradient-based attacks," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[28] G. Carbone, L. Bortolussi, and G. Sanguinetti, "Resilience of bayesian layer-wise explanations under adversarial attacks," in *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8.

[29] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *Proc. of the IEEE Symposium on Security and Privacy (S&P)*, pp. 39–57, 2017.

[30] Z. Carmichael and W. J. Scheirer, "Unfooling perturbation-based post hoc explainers," in *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2023.

[31] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronicsweek*, vol. 8, no. 8, 2019.

[32] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 839–847.

[33] H. Chefer, I. Schwartz, and L. Wolf, "Optimizing relevance maps of vision transformers improves robustness," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[34] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. M. Molloy, and B. Srivastava, "Detecting backdoor attacks on deep neural networks by activation clustering," in *Proc. of the Workshop on Artificial Intelligence Safety Co-Located with the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[35] J. Chen, X. Wu, V. Rastogi, Y. Liang, and S. Jha, "Robust attribution regularization," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 14 300–14 310.

[36] V. Chen, J. Li, J. S. Kim, G. Plumb, and A. Talwalkar, "Towards connecting use cases and methods in interpretable machine learning," *CoRR*, vol. abs/2103.06254, 2021.

[37] E. Chou, F. Tramèr, and G. Pellegrino, "SentiNet: Detecting localized universal attacks against deep learning systems," in *Proc. of the IEEE Symposium on Security and Privacy Workshops*, 2020, pp. 48–54.

[38] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, "Data cleaning: Overview and emerging challenges," in *Proc. of the International Conference on Management of Data (SIGMOD)*, 2016, pp. 2201–2206.

[39] Y.-N. Chuang, G. Wang, F. Yang, Z. Liu, X. Cai, M. Du, and X. Hu, "Efficient XAI techniques: A taxonomic survey," 2023.

[40] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 97, 2019, pp. 1310–1320.

[41] R. D. Cook and S. Weisberg, "Characterizations of an empirical influence function for detecting influential cases in regression," vol. 22, no. 4, 1980.

[42] F. Craighero, F. Angaroni, F. Stella, C. Damiani, M. Antoniotti, and A. Graudenzi, "Unity is strength: Improving the detection of adversarial examples with ensemble approaches," *CoRR*, vol. abs/2111.12631, 2022.

[43] M. W. Craven and J. W. Shavlik, "Extracting tree-structured representations of trained networks," in *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, 1995, pp. 24–30.

[44] F. Croce, S. Gowal, T. Brunner, E. Shelhamer, M. Hein, and A. T. Cemgil, "Evaluating the adversarial robustness of adaptive test-time defenses," in *Proc. of the International Conference on Machine Learning (ICML)*, 2022, pp. 4421–4435.

[45] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "A survey of the state of explainable ai for natural language processing," in *Proc. of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (AACL/IJCNLP)*, 2020, pp. 447–459.

[46] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," *CoRR*, vol. abs/2006.11371, 2020.

[47] B. G. Doan, E. Abbasnejad, and D. C. Ranasinghe, "Februus: Input purification defense against trojan attacks on deep neural network systems," in *Proc. of the Annual Computer Security Applications Conference (ACSAC)*, 2020, pp. 897–912.

[48] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel, "Explanations can be manipulated and geometry is to blame," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 13 567–13 578.

[49] A.-K. Dombrowski, C. J. Anders, K.-R. Müller, and P. Kessel, "Towards robust explanations for deep neural networks," *Pattern Recognition*, vol. 121, p. 108194, 2022.

[50] H. Drucker and Y. LeCun, "Improving generalization performance using double backpropagation," *IEEE Transactions on Neural Networks*, vol. 3, no. 6, pp. 991–997, 1992.

[51] M. Du, N. Liu, F. Yang, and X. Hu, "Learning credible deep neural networks with rationale regularization," in *Proc. of the International Conference on Data Mining (ICDM)*, 2019, pp. 150–159.

[52] O. Eberle, J. Büttner, F. Kräutli, K.-R. Müller, M. Valleriani, and G. Montavon, "Building and interpreting deep similarity models," *IEEE Transactions Pattern Analyses and Machine Intelligence*, vol. 44, no. 3, pp. 1149–1161, 2022.

[53] W. Fan, H. Xu, W. Jin, X. Liu, X. Tang, S. Wang, Q. Li, J. Tang, J. Wang, and C. C. Aggarwal, "Jointly attacking graph neural network and its explanations," in *Proc. of the IEEE International Conference on Data Engineering (ICDE)*, 2023, pp. 654–667.

[54] S. Fang and A. Choromanska, "Backdoor attacks on the DNN interpretation system," *Proc. of the National Conference on Artificial Intelligence (AAAI)*, pp. 561–570, 2022.

[55] G. Fidel, R. Bitton, and A. Shabtai, "When explainability meets adversarial learning: Detecting adversarial examples using SHAP signatures," in *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[56] R. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3449–3457, 2017.

[57] H. Gao, Y. Chen, and W. Zhang, "Detection of trojaning attack on neural networks via cost of sample classification," *Secur. Commun. Networks*, vol. 2019, pp. 1 953 839:1–1 953 839:12, 2019.

[58] S. Garg, A. Kumar, V. Goel, and Y. Liang, "Can adversarial weight perturbations inject neural backdoors," in *Proc. of ACM International Conference on Information and Knowledge Management (CIKM)*, 2020, pp. 2029–2032.

[59] J. Geiping, L. H. Fowl, W. R. Huang, W. Czaja, G. Taylor, M. Moeller, and T. Goldstein, "Witches' brew: Industrial scale data poisoning via gradient matching," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.

[60] A. Ghorbani, A. Abid, and J. Y. Zou, "Interpretation of neural networks is fragile," in *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2019.

[61] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 9273–9282.

[62] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2015.

[63] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. D. McDaniel, "On the (statistical) detection of adversarial examples," *CoRR*, vol. abs/1702.06280, 2017.

[64] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," in *Proc. of the International Conference on Learning Representations (ICLR) Workshop Track Proceedings*, 2015.

[65] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[66] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," *CoRR*, vol. abs/1805.10820, 2018.

[67] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 93:1–93:43, 2019.

[68] W. Guo, D. Mu, J. Xu, P. Su, G. Wang, and X. Xing, "LEMNA: Explaining deep learning based security applications," in *Proc. of the ACM Conference on Computer and Communications Security (CCS)*, 2018, pp. 364–379.

[69] S. J. Hanson and L. Y. Pratt, "Comparing biases for minimal network construction with back-propagation," in *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, 1988, pp. 177–185.

[70] Z. He and M. Singhal, "Defense-CycleGAN: A defense mechanism against adversarial attacks using CycleGAN to reconstruct clean images," in *Proc. of the International Conference on Pattern Recognition and Machine Learning (PRML)*, 2022, pp. 173–179.

[71] J. Heo, S. Joo, and T. Moon, "Fooling neural network interpretations via adversarial model manipulation," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 2921–2932.

[72] G. E. Hinton *et al.*, "Learning distributed representations of concepts," in *Proc. of the Annual Conference of the Cognitive Science Society*, vol. 1, 1986, p. 12.

[73] S. Hong, M.-A. Panaitescu-Liess, Y. Kaya, and T. Dumitras, "Qu-ANTI-Zation: Exploiting quantization artifacts for achieving adversarial outcomes," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021, pp. 9303–9316.

[74] X. Huang, M. Alzantot, and M. Srivastava, "NeuronInspect: Detecting backdoors in neural networks via output explanations," *CoRR*, vol. abs/1911.07399, 2019.

[75] A. Ivankay, I. Girardi, C. Marchiori, and P. Frossard, "FAR: A general framework for attributional robustness," in *Proc. of the British Machine Vision Conference (BMVC)*, 2021, p. 24.

[76] A. Ivankay, I. Girardi, P. Frossard, and C. Marchiori, "Fooling explanations in text classifiers," *Proc. of the International Conference on Learning Representations (ICLR)*, p. 13, 2022.

[77] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *Proc. of the IEEE Symposium on Security and Privacy (S&P)*, 2018, pp. 19–35.

[78] M. Jagielski, G. Severi, N. P. Harger, and A. Oprea, "Subpopulation data poisoning attacks," in *Proc. of the ACM Conference on Computer and Communications Security (CCS)*, 2021.

[79] A. Kanehira, K. Takemoto, S. Inayoshi, and T. Harada, "Multimodal explanations by predicting counterfactuality in videos," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8594–8602.

[80] C.-Y. Kao, J. Chen, K. Markert, and K. Böttinger, "Rectifying adversarial inputs using XAI techniques," in *Proc. of the European Signal Processing Conference (EUSIPCO)*, 2022, pp. 573–577.

[81] M. Kendall and J. D. Gibbons, *Rank Correlation Methods*, 5th ed., 1990.

[82] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 80, 2018, pp. 2673–2682.

[83] P.-J. Kindermans, K. Schütt, K.-R. Müller, and S. Dähne, "Investigating the influence of noise and distractors on the interpretation of neural networks," *CoRR*, vol. abs/1611.07270, 2016.

[84] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 70, 2017, pp. 1885–1894.

[85] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, 1991, pp. 950–957.

[86] A. Kuppa and N.-A. Le-Khac, "Black box attacks on explainable artificial intelligence (XAI) methods in cyber security," in *Proc. of the International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[87] G. Laberge, U. Aïvodji, and S. Hara, "Fooling SHAP with stealthily biased sampling," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2023.

[88] H. Lakkaraju and O. Bastani, ""how do i fool you?": Manipulating user trust via misleading black box explanations," in *Proc. of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020, pp. 79–85.

[89] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Faithful and customizable explanations of black box models," in *Proc. of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2019, pp. 131–138.

[90] H. Lakkaraju, N. Arsov, and O. Bastani, "Robust and stable black box explanations," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 119, 2020, pp. 5628–5638.

[91] S. Lapuschkin, A. Binder, G. Montavon, K. Müller, and W. Samek, "Analyzing classifiers: Fisher vectors and deep neural networks," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2016, pp. 2912–2920.

[92] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.

[93] G.-H. Lee, D. Alvarez-Melis, and T. S. Jaakkola, "Towards robust, locally linear deep networks," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2019.

[94] J. R. Lee, S. Kim, I. Park, T. Eo, and D. Hwang, "Relevance-CAM: Your model already knows where to look," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14 944–14 953.

[95] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[96] S. C. Lee, "A black box approach to auditing algorithms," *Issues In Information Systems*, 2022.

[97] A. Levine and S. Feizi, "(de)Randomized smoothing for certifiable defense against patch attacks," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[98] A. Levine, S. Singla, and S. Feizi, "Certifiably robust interpretation in deep learning," *CoRR*, vol. abs/1905.12105, 2019.

[99] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Neural attention distillation: Erasing backdoor triggers from deep neural networks," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.

[100] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 16, no. 10, pp. 36–43, 2018.

[101] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-Pruning: Defending against backdooring attacks on deep neural networks," in *Proc. of the International Symposium Research in Attacks, Intrusions, and Defenses (RAID)*, vol. 11050, 2018, pp. 273–294.

[102] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Proc. of the Network and Distributed System Security Symposium (NDSS)*, 2018.

[103] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. Zhang, "ABS: Scanning neural networks for back-doors by artificial brain stimulation," in *Proc. of the ACM Conference on Computer and Communications Security (CCS)*, 2019.

[104] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, p. 10, 2017.

[105] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2018.

[106] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2018.

[107] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2017.

[108] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable AI systems," *ACM Transactions Interactive Intelligent Systems*, vol. 11, no. 3-4, pp. 24:1–24:45, 2021.

[109] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning – a brief history, state-of-the-art and challenges," in *Proc. of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD Workshops)*, 2020.

[110] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[111] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[112] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: An overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 2019, pp. 193–209.

[113] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 86–94.

[114] S.-M. Moosavi-Dezfooli, A. Fawzi, J. Uesato, and P. Frossard, "Robustness via curvature regularization, and vice versa," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9078–9086.

[115] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proc. of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 071–22 080, 2019.

[116] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 162, 2022, pp. 16 805–16 827.

[117] M. Noppel and C. Wressnegger, "Explanation-aware backdoors in a nutshell." in *Proc. of 46th German Conference on Artificial Intelligence (KI)*, 2023.

[118] M. Noppel, L. Peter, and C. Wressnegger, "Disguising attacks with explanation-aware backdoors," in *Proc. of the IEEE Symposium on Security and Privacy (S&P)*, 2023, pp. 664–681.

[119] N. Papernot, P. D. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. of the IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016, pp. 372–387.

[120] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "SoK: Security and privacy in machine learning," in *Proc. of the IEEE European Symposium on Security and Privacy (EuroS&P)*, 2018, pp. 399–414.

[121] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8779–8788.

[122] G. Plumb, M. Al-Shedivat, Á. A. Cabrera, A. Perer, E. P. Xing, and A. Talwalkar, "Regularizing black-box models for improved interpretability," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[123] N. Puri, P. Gupta, P. Agarwal, S. Verma, and B. Krishnamurthy, "MAGIX: Model agnostic globally interpretable explanations," *CoRR*, vol. abs/1706.07160, 2017.

[124] X. Qi, T. Xie, R. Pan, J. Zhu, Y. Yang, and K. Bu, "Towards practical deployment-stage backdoor attack on deep neural networks," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 337–13 347.

[125] M. Raghu, B. Poole, J. M. Kleinberg, S. Ganguli, and J. Sohl-Dickstein, "On the expressive power of deep neural networks," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 70, 2017, pp. 2847–2854.

[126] A. Raghunathan, J. Steinhardt, and P. Liang, "Certified defenses against adversarial examples," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2018.

[127] D. Rajapaksha, C. Bergmeir, and W. L. Buntine, "LoRMIkA: Local rule-based model interpretability with k-optimal associations," *Information Fusion*, 2020.

[128] A. S. Rakin, Z. He, and D. Fan, "TBT: Targeted neural network attack with bit trojan," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 195–13 204.

[129] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[130] ——, "Anchors: High-precision model-agnostic explanations," in *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2018, pp. 1527–1535.

[131] L. Rieger and L. K. Hansen, "A simple defense against adversarial attacks on heatmap explanations," in *Proc. of the ICML Workshop on Human Interpretability in Machine Learning (WHI)*, 2020.

[132] L. Rieger, C. Singh, W. J. Murdoch, and B. Yu, "Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 119, 2020.

[133] L. Rokach and O. Maimon, *Data Mining with Decision Trees - Theory and Applications. 2nd Edition*, ser. Series in Machine Perception and Artificial Intelligence. WorldScientific, 2014, vol. 81.

[134] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[135] K. Roth, Y. Kilcher, and T. Hofmann, "The odds are odd: A statistical test for detecting adversarial examples," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 97, 2019, pp. 5498–5507.

[136] G. M. Rotskoff and E. Vanden-Eijnden, "Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error," *CoRR*, vol. abs/1805.00915, 2018.

[137] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[138] H. Salman, J. Li, I. P. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang, "Provably robust deep learning via adversarially trained smoothed classifiers," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019, pp. 11 289–11 300.

[139] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Muller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proc. of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.

[140] A. Sarkar, A. Sarkar, and V. N. Balasubramanian, "Enhanced regularizers for attributional robustness," in *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2021, pp. 2532–2540.

[141] T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K.-R. Müller, and G. Montavon, "Higher-order explanations of graph neural networks via relevant walks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7581–7596, 2022.

[142] J. Schneider, C. Meske, and M. Vlachos, "Deceptive AI explanations: Creation and detection," in *Proc. of the International Conference on Agents and Artificial Intelligence (ICAART)*, vol. 2, 2022, pp. 44–55.

[143] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[144] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, 2020.

[145] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison Frogs! targeted clean-label poisoning attacks on neural networks," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018, pp. 6106–6116.

[146] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proc. of the International Conference on Machine Learning (ICML)*, 2017, pp. 3145–3153.

[147] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. of the International Conference on Learning Representations (ICLR) Workshop Track Proceedings*, 2014.

[148] M. Singh, N. Kumari, P. Mangla, A. Sinha, V. N. Balasubramanian, and B. Krishnamurthy, "Attributional robustness training using input-gradient spatial alignment," in *Proc. of the European Conference on Computer Vision (ECCV)*, vol. 12372, 2020, pp. 515–533.

[149] S. Singla, E. Wallace, S. Feng, and S. Feizi, "Understanding impacts of high-order loss approximations and features in deep learning interpretation," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 97, 2019.

[150] S. Sinha, H. Chen, A. Sekhon, Y. Ji, and Y. Qi, "Perturbing inputs for fragile interpretations in deep natural language processing," in *Proc. of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP (BlackboxNLP@EMNLP)*, 2021.

[151] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *Proc. of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2020, pp. 180–186.

[152] D. Slack, S. Hilgard, H. Lakkaraju, and S. Singh, "Counterfactual explanations can be manipulated," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[153] D. Slack, S. Hilgard, S. Singh, and H. Lakkaraju, "Feature attributions and counterfactual explanations can be manipulated," *CoRR*, vol. abs/2106.12563, 2021.

[154] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "SmoothGrad: Removing noise by adding noise," *CoRR*, vol. abs/1706.03825, 2017.

[155] A. Søgaard, "Shortcomings of interpretability taxonomiesfor deep neural networks," *Advances in Interpretable Machine Learning and Artificial Intelligence (AIMLAI)*, 2022.

[156] T. Speith, "A review of taxonomies of explainable artificial intelligence (XAI) methods," in *Proc. of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022, pp. 2239–2250.

[157] J. Steinhardt, P. W. Koh, and P. Liang, "Certified defenses for data poisoning attacks," in *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 3517–3529.

[158] P. Stock and M. Cissé, "ConvNets and ImageNet beyond accuracy: Understanding mistakes and uncovering biases," in *Proc. of the European Conference on Computer Vision (ECCV)*, vol. 11210, 2018, pp. 504–519.

[159] A. Subramanya, V. Pillai, and H. Pirsiavash, "Fooling network interpretation in image classification," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2020–2029.

[160] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. of the International Conference on Machine Learning (ICML)*, 2017, pp. 3319–3328.

[161] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2014.

[162] S. V. Tamam, R. Lapid, and M. Sipper, "Foiling explanations in deep neural networks," *CoRR*, vol. abs/2211.14860, 2022.

[163] R. Tang, N. Liu, F. Yang, N. Zou, and X. Hu, "Defense against explanation manipulation," *Frontiers Big Data*, vol. 5, p. 704203, 2022.

[164] G. Tao, S. Ma, Y. Liu, and X. Zhang, "Attacks meet interpretability: Attribute-steered detection of adversarial samples," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018, pp. 7728–7739.

[165] Y. Tian, F. Suya, F. Xu, and D. Evans, "Stealthy backdoors as compression artifacts," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 1372–1387, 2022.

[166] F. Tramèr, "Detecting adversarial examples is (nearly) as hard as classifying them," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 162, 2022, pp. 21 692–21 702.

[167] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2018, pp. 8011–8021.
[168] A. K. Veldanda, K. Liu, B. Tan, P. Krishnamurthy, F. Khorrami, R. Karri, B. Dolan-Gavitt, and S. Garg, "Nnoculation: Catching badnets in the wild," in *Proc. of the ACM Workshop on Artificial Intelligence and Security (AISEC)*, 2021, pp. 49–60.
[169] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: Challenges revisited," *CoRR*, vol. abs/2106.07756, 2021.
[170] T. J. Viering, Z. Wang, M. Loog, and E. Eisemann, "How to manipulate cnns to make them lie: The gradcam case," in *Proc. of the British Machine Vision Conference (BMVC) Workshop on Interpretable and Explainable Machine Vision*, 2019.
[171] G. Vilone and L. Longo, "Classification of explainable artificial intelligence methods through their output formats," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 615–661, 2021.
[172] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *SSRN Electronic Journal*, 2017.
[173] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural Cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. of the IEEE Symposium on Security and Privacy (S&P)*, 2019, pp. 707–723.
[174] F. Wang and A. W.-K. Kong, "Exploiting the relationship between kendall's rank correlation and cosine similarity for attribution protection," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
[175] ——, "A practical upper bound for the worst-case attribution deviations," *CoRR*, vol. abs/2303.00340, 2023.
[176] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. Liu, "Dual attention suppression attack: Generate adversarial camouflage in physical world," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8565–8574.
[177] J. Wang, J. Tuyls, E. Wallace, and S. Singh, "Gradient-based analysis of NLP models is manipulable," in *Emnlp*, vol. EMNLP 2020, 2020, pp. 247–258.
[178] S. Wang and Y. Gong, "Adversarial example detection based on saliency map features," *Applied Intelligence*, vol. 52, no. 6, pp. 6262–6275, 2022.
[179] X. Wang, C. Liu, X. Hu, Z. Wang, J. Yin, and X. Cui, "Make data reliable: An explanation-powered cleaning on malware dataset against backdoor poisoning attacks," in *Proc. of the Annual Computer Security Applications Conference (ACSAC)*, 2022, pp. 267–278.
[180] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[181] Z. Wang, H. Wang, S. Ramkumar, P. Mardziel, M. Fredrikson, and A. Datta, "Smoothed geometry for robust attribution," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
[182] A. Warnecke, D. Arp, C. Wressnegger, and K. Rieck, "Evaluating explanation methods for deep learning in computer security," in *Proc. of the IEEE European Symposium on Security and Privacy (EuroS&P)*, Sep. 2020.
[183] A. S. Weigend, D. E. Rumelhart, and B. A. Huberman, "Generalization by weight-elimination with application to forecasting," in *Proc. of the Annual Conference on Neural Information Processing Systems (NIPS)*, 1990, pp. 875–882.
[184] E. Weinberger, J. D. Janizek, and S.-I. Lee, "Learning deep attribution priors based on prior knowledge," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, Virtual*, 2020.
[185] D. Wu and Y. Wang, "Adversarial neuron pruning purifies backdoored deep models," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021, pp. 16 913–16 925.
[186] S. Wu, J. Wang, W. Ping, W. Nie, and C. Xiao, "Defending against adversarial audio via diffusion model," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2023.
[187] C. Xiang, A. N. Bhagoji, V. Sehwag, and P. Mittal, "PatchGuard: A provably robust defense against adversarial patches via small receptive fields and masking," in *Proc. of the USENIX Security Symposium*, 2021, pp. 2237–2254.
[188] X. Xu, Q. Wang, H. Li, N. Borisov, C. A. Gunter, and B. Li, "Detecting AI trojans using meta neural analysis," in *Proc. of the IEEE Symposium on Security and Privacy (S&P)*, 2021, pp. 103–120.
[189] M. Xue, Y. Wu, Z. Wu, Y. Zhang, J. Wang, and W. Liu, "Detecting backdoor in deep neural networks via intentional adversarial perturbations," *Inf. Sci.*, vol. 634, pp. 564–577, 2021.
[190] C. Yeh, B. Kim, S. Ö. Arik, C. Li, T. Pfister, and P. Ravikumar, "On completeness-aware concept-based explanations in deep neural networks," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
[191] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "GNNExplainer: Generating explanations for graph neural networks," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019.
[192] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. of the European Conference on Computer Vision (ECCV)*, vol. 8689, 2014, pp. 818–833.
[193] Y. Zeng, S. Chen, W. Park, Z. Mao, M. Jin, and R. Jia, "Adversarial unlearning of backdoors via implicit hypergradient," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2022.
[194] G. Zhang, C. Wang, B. Xu, and R. Grosse, "Three mechanisms of weight decay regularization," in *Proc. of the International Conference on Learning Representations (ICLR) Poster Track Proceedings*, 2018.
[195] H. Zhang, J. Gao, and L. Su, "Data poisoning attacks against outcome interpretations of predictive models," in *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2021, pp. 2165–2173.
[196] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.
[197] R. Zhang, P. Madumal, T. Miller, K. A. Ehinger, and B. I. P. Rubinstein, "Invertible concept-based explanations for CNN models with non-negative concept activation vectors," in *Proc. of the National Conference on Artificial Intelligence (AAAI)*. AAAI Press, 2021, pp. 11 682–11 690. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17389
[198] X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, and T. Wang, "Interpretable deep learning under fire," in *Proc. of the USENIX Security Symposium*, 2020.
[199] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 5, no. 5, pp. 726–742, 2021.
[200] Y. Zhang, A. Albarghouthi, and L. D'Antoni, "BagFlip: A certified defense against data poisoning," in *Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
[201] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.
[202] Y. Zhou, S. Booth, M. T. Ribeiro, and J. Shah, "Do feature attribution methods correctly attribute features?" in *Proc. of the National Conference on Artificial Intelligence (AAAI)*, 2022, pp. 9623–9633.

A. Scalar Robustness Notions

As we discussed in the main part of this paper, robustness can either be considered a boolean or a scalar. The boolean perspective has a clear definition, compared to the scalar case. Nevertheless, to come up with a scalar we do have multiple options:

1) We measure how many pairs satisfy the constraint and how many do not.
2) We measure how close the pairs are to satisfying the constraints on average, or in the worst case.
3) In addition, we can weigh both above suggestions with the probability of the occurrence of the first tuple element.

B. Restrictions and Constraints

In Table 3 we provide an overview on the restrictions and constrains we define in this paper.

TABLE 3: Abbreviations and the corresponding formula for each restriction and constraint, as defined above.

Name	Formula
$\text{LIP}^{d_{\mathcal{E}},d_{\mathcal{X}},K}$	$\exists\gamma\in\mathbb{R}^+\ d_{\mathcal{E}}\big(h_\theta(\mathbf{x}),\gamma h_\theta(\tilde{\mathbf{x}})\big)\leq Kd_{\mathcal{X}}(\mathbf{x},\tilde{\mathbf{x}})$
$\text{EXPLSIM}^{d_{\mathcal{E}},\epsilon}$	$\exists\gamma\in\mathbb{R}^+\ d_{\mathcal{E}}(h_\theta(\mathbf{x}),\gamma h_\theta(\tilde{\mathbf{x}}))\leq\epsilon$
EXPLEQ	$\exists\gamma\in\mathbb{R}^+\ h_\theta(\mathbf{x})=\gamma h_\theta(\tilde{\mathbf{x}})$
CLSEQ	$\mathcal{F}_\theta(\mathbf{x})=\mathcal{F}_\theta(\tilde{\mathbf{x}})$
$\text{LOC}^{d_{\mathcal{X}},\delta}$	$d(\mathbf{x},\tilde{\mathbf{x}})\leq\delta$

C. Meta-Review

C.1. Summary

This paper presents a SoK of explainable machine learning (XAI) techniques in adversarial environments, by (1) summarizing attacks designed to subvert explanations, (2) formalizing notions of adversarial robustness in presence of explanation-aware attacks for attackers with different objectives, (3) presenting the taxonomy of existing defenses against explanation-aware attacks, and (4) pointing out future research directions in this space.

C.2. Scientific Contributions

- Provides a Valuable Step Forward in an Established Field
- Establishes a New Research Direction

C.3. Reasons for Acceptance

1) **Topic:** This paper studies a timely and important topic, whose knowledge has not been sufficiently systemized in existing literature. A SoK paper in this emerging space will help guide future research efforts.
2) **Attack coverage:** This paper covers an extensive spectrum of explanation-aware attacks against existing post-hoc feature-attribution XAI techniques, taxonomized by different aspects (e.g., goal, input, system level) of attacks.
3) **Formulated XAI safety:** This paper introduces a well-grounded formulation and notion framework to analyze the robustness of XAI techniques in presence of explanation-aware attacks, as well as a taxonomy of existing defenses against these attacks. Such formulation and taxonomization helps drive general understandings of XAI safety properties in formal settings.
4) **Insightful open research questions:** Many raised research questions in this paper are worth exploration and can potentially guide future research efforts.

C.4. Noteworthy Concerns

1) **Pre-mature XAI:** We believe an early SoK paper helps drive this emerging space forward. At the same time, however, we would like to see some discussions regarding how the premature nature of this space and its consequential operational issues (e.g., inaccuracy of attribution results) interact with adversarial attempts – such discussion can be qualitative.
2) **Scope:** There are some scope-related claims in the paper that are not completely aligned with the actual paper content. The scope of "XAI" techniques in this paper seems to focus on a specific group (post-hoc feature attribution for non-generative classification models) instead of others.