Adversarial Attacks and Defenses in Explainable Artificial Intelligence: A Survey

Hubert Baniecki [1] **Przemyslaw Biecek** [1 2]

Abstract

Explainable artificial intelligence (XAI) methods are portrayed as a remedy for debugging and trusting statistical and deep learning models, as well as interpreting their predictions. However, recent advances in adversarial machine learning highlight the limitations and vulnerabilities of state-of-the-art explanations, putting their security and trustworthiness into question. The possibility of manipulating, fooling or fairwashing evidence of the model's reasoning has detrimental consequences when applied in high-stakes decision-making and knowledge discovery. This *concise survey* of over 50 papers summarizes research concerning adversarial attacks on explanations of machine learning models, as well as fairness metrics. We discuss how to defend against attacks and design robust interpretation methods. We contribute a list of existing insecurities in XAI and outline the emerging research directions in adversarial XAI (AdvXAI).

1. Introduction

Explainable artificial intelligence (XAI) methods [for a brief overview see Holzinger et al., 2022, and for a comprehensive survey refer to Schwalbe & Finzel, 2023], e.g. post-hoc explanations like PDP (Friedman, 2001), SG (Simonyan et al., 2014), LIME (Ribeiro et al., 2016), IG (Sundararajan et al., 2017), SHAP (Lundberg & Lee, 2017), TCAV (Kim et al., 2018), Grad-CAM (Selvaraju et al., 2020) to name a few, provide various mechanisms to interpret predictions of machine learning models. A popular critique of XAI, in favour of inherently interpretable models, is its inability to faithfully explain the black-box (Rudin, 2019). Nevertheless, explanations find success in applications like autonomous driving (Gu et al., 2020) or drug discovery (Jiménez-Luna et al., 2020), and can be used to better understand the reasoning of large models like AlphaZero (McGrath et al., 2022). Recently, adversarial machine learning (AdvML, Kolter & Madry, 2018; Rosenberg et al., 2021; Machado et al., 2021) became more prevalent in research on XAI, yet vulnerabilities of explanations raise concerns about their trustworthiness and security (Papernot et al., 2018). To assess the scope of these threats, we contribute a systemization of the current state of knowledge concerning *adversarial attacks on model explanations* (Section 2) and *defense mechanisms against these attacks* (Section 3). Figure 1 presents one of such attacks, which is often called *adversarial example*, i.e. a slightly changed image drastically changes the explanation of the class predicted by a model. An aggregation of explanations obtained with different methods shows to be less susceptible to such manipulation. While most related surveys summarize explanation robustness (Mishra et al., 2021), attacks on model predictions (Machado et al., 2021), and the application of XAI in AdvML (Liu et al., 2021), this survey highlights the rapidly emerging cross-domain research in what we call adversarial explainable AI (AdvXAI). We moreover confront it with the closely related work concerning *adversarial attacks on machine learning fairness metrics* (Section 4). A concise overview of over 50 papers allows us to specify the *frontier research directions in AdvXAI* (Section 5).

We acknowledge this is not a systematic review, but rather an approachable outlook to recognize the potential gaps and define future directions. We first included visible papers published in major machine learning conferences (ICML, ICLR, NeurIPS, AAAI) and journals (AIj, NMI) since Ghorbani et al. (2019). We then extensively searched their citation networks for papers related to AdvXAI published in other venues. We purposely exclude a large number of papers focusing primarily on explanation evaluation without relating to the adversarial scenario (refer to Nauta et al., 2023).

2. Adversarial attacks on model explanations

To the best of our knowledge, Ghorbani et al. (2019) is the first contribution to mention[3] and propose an *adversarial attack* against explanation methods, specifically gradient-based saliency maps (Simonyan et al., 2014; Sundararajan

[1] University of Warsaw, Poland [2] Warsaw University of Technology, Poland. Correspondence to: Hubert Baniecki <h.baniecki@uw.edu.pl>.

2^{nd} AdvML-Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[3] Note that we acknowledge papers in order of date published, i.e. presented at a conference, as opposed to the first date appearing online, e.g. as a preprint or final proceedings version.

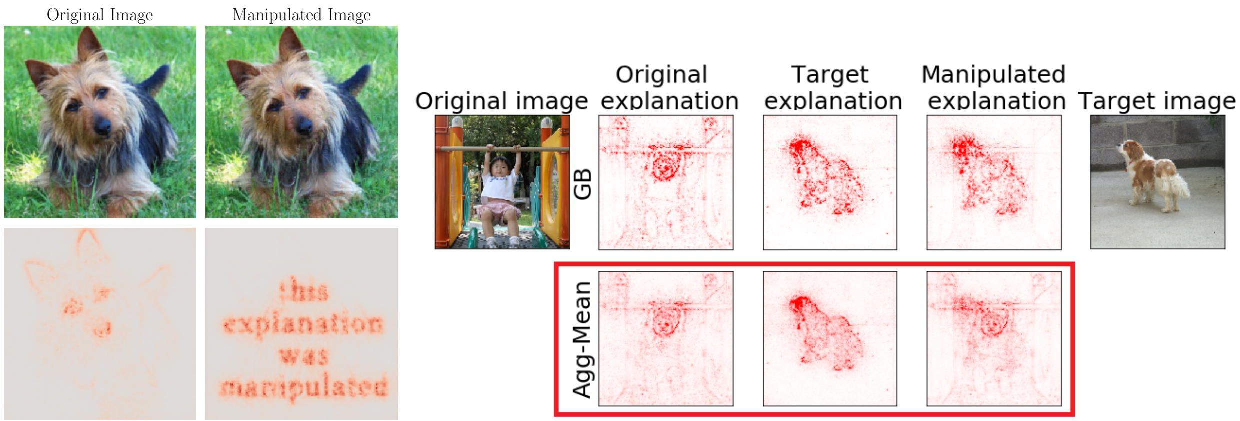

Figure 1. Adversarial example attack on an explanation of the image classifier's prediction (**left**, adapted from Dombrowski et al., 2019, with permission) and an aggregating defense mechanism against this attack (**right**, adapted from Rieger & Hansen, 2020, with permission).

et al., 2017) of (convolutional) neural networks. Previous related work discussed the worst-case (adversarial) notion of explanation robustness (Alvarez Melis & Jaakkola, 2018, page 7) and the notion of explanation sensitivity (Ancona et al., 2018; Kindermans et al., 2019). Crucially, Adebayo et al. (2018) introduced randomization tests showing that a visual inspection of explanations alone can favor methods compelling to humans. It raised to attention the need for evaluating the explanations' quality, especially for deep models, with possible implications in adversarial settings.

Table 1 lists attacks on explanation methods, with the corresponding strategy of changing data, e.g. an adversarial example manipulates the explanation without impacting the prediction (Ghorbani et al., 2019; Dombrowski et al., 2019), changing the model, e.g. fine-tuning or regularizing weights manipulates explanations without impacting the predictive performance (Heo et al., 2019; Dimanov et al., 2020), or changing both data and the model, e.g. in the case when an attacker poisons the training dataset (Zhang et al., 2021). Viering et al. (2019) manipulate Grad-CAM explanations of a convolutional neural network by changing its weights, but also proposes to leave a backdoor in the network (triggered by specific input patterns), which allows retrieving original explanations. Noppel et al. (2023) extend fooling explanations through fine-tuning and backdooring to consider: (i) a red-herring attack that manipulates the explanation to cover an adversarial change in the model's prediction, e.g. a misclassification, and (ii) a fully disguising attack that aims to show the original explanation for a changed prediction.

For each of the attacks in Table 1, we record the mentioned data modalities with the corresponding datasets used in experiments, as well as model algorithms. **Observation 1.** The majority of proposed attacks assume prior knowledge that the explained model is a neural network, e.g. to utilize gradient descent in constructing adversarial examples or changing model parameters, as opposed to black-box approaches that could work with various model algorithms.

In a black-box setting, Slack et al. (2020) manipulate LIME and SHAP explanations for tabular data by exploiting their reliance on perturbing input data for estimation. The proposed attack substitutes a biased black-box with a model surrogate to effectively hide bias, e.g. from auditors. In detail, an out-of-distribution detector is trained to divide input data such that the black-box'es predictions in-distribution remain biased, but its behavior on the perturbed data is controlled, which makes the explanations look fair. Merrer & Trédan (2020) consider a similar adversarial scenario and show that providing explanations cannot prevent a remote service from hiding the true reasons that lead to its predictions. It concludes that an impractically large number of user queries is required to detect explanation manipulation.

While the majority of adversarial attacks are on local methods for interpreting individual predictions; other attacks specifically target global methods explaining the overall model's reasoning (Lakkaraju & Bastani, 2020; Baniecki & Biecek, 2022; Brown & Kvinge, 2022; Baniecki et al., 2022; Laberge et al., 2023). Instead of changing model or data, Lakkaraju & Bastani (2020) introduce misleading rule-based explanations that approximate a model based on the MUSE framework (Lakkaraju et al., 2019). Results of a user study show that various high-fidelity explanations faithful to the black-box considerably affect human judgement. Baniecki & Biecek (2022) and Baniecki et al. (2022) introduce genetic-based algorithms to manipulate SHAP and PDP explanations respectively. The proposed poisoning attack iteratively changes data used in the process of estimating global explanations, and thus can be exploited by an adversary to provide false evidence of feature importance and effects. Laberge et al. (2023) consider a similar adversarial scenario and attack global SHAP using stealthily

Table 1. Summary of adversarial attacks on explanations of machine learning models. We abbreviate the following: data (D), model (M), image (I), tabular (T), language (Lg), neural network (N), black-box (B), local (L), global (G). Appendix A lists other abbreviations.

Attack	**Changes** strategy	**Modality** dataset	**Model** algorithm	**Explanation** algorithm
(Ghorbani et al., 2019)	D adversarial example	I ImageNet, CIFAR-10	N SqueezeNet, InceptionNet	L SG, IG, DeepLIFT
(Kindermans et al., 2019)	D adversarial example	I MNIST, ImageNet	N MLP, CNN, VGG	L SG, GI, IG, LRP, ..
(Viering et al., 2019)	M & D backdooring attack	I ImageNet	N VGG	L Grad-CAM
(Subramanya et al., 2019)	D adversarial example	I ImageNet, VOC2012	N VGG, ResNet, DenseNet	L Grad-CAM
(Heo et al., 2019)	M model manipulation	I ImageNet	N VGG, ResNet, DenseNet	L SG, Grad-CAM, LRP
(Dombrowski et al., 2019)	D adversarial example	I ImageNet, CIFAR-10	N VGG, ResNet, DenseNet, ..	L SG, GI, IG, LRP, ..
(Dimanov et al., 2020)	M model manipulation	T Credit, COMPAS, Adult, ..	N MLP	L SG, GI, IG, SHAP, ..
(Slack et al., 2020)	M surrogate model	T Credit, COMPAS, Crime	B rule set	L SHAP, LIME
(Lakkaraju & Bastani, 2020)	–	T Bail	B rule set	G MUSE
(Anders et al., 2020)	D surrogate model	T credit, I MNIST, CIFAR10, ..	N LR, CNN, VGG	L SG, GI, IG, LRP
(Kuppa & Le-Khac, 2020)	D adversarial example	T PDF, Android, UGR16	N MLP, GAN	L SG, GI, IG, LRP, ..
(Zhang et al., 2020)	D adversarial example	I ImageNet	N ResNet, DenseNet	L SG, CAM, RTS, ..
(Merrer & Trédan, 2020)	M surrogate model	T Credit	B DT, MLP	L custom
(Shokri et al., 2021)	– membership inference	T Adult, Hospital, .. I CIFAR-10, ..	B MLP, CNN	L IG, LRP, LIME, ..
(Sinha et al., 2021)	D adversarial example	Lg IMDB, SST, AG News	B DistilBERT, RoBERTa	L IG, LIME
(Zhang et al., 2021)	D & M data poisoning	T Fracture, I Dogs	N MLP, ResNet	L SG, CAM
(Slack et al., 2021a)	D model manipulation	T Credit, Crime	N MLP	L counterfactual
(Baniecki & Biecek, 2022)	D data poisoning	T Heart, Apartments	B XGBoost	G SHAP, L SHAP
(Brown & Kvinge, 2022)	D data poisoning	I ImageNet, CUB	N InceptionNet, ResNet, ViT, ..	G TCAV, FFV
(Baniecki et al., 2022)	D data poisoning	T Heart, Friedman	B MLP, RF, GBDT, SVM, KNN, ..	G PDP
(Pawelczyk et al., 2023b)	– membership inference	T Adult, Hospital	B LR, NN	L counterfactual
(Laberge et al., 2023)	D data poisoning	T COMPAS, Adult, Bank, Crime	B MLP, RF, XGBoost	G SHAP
(Noppel et al., 2023)	M & D backdooring attack	I CIFAR-10, GTSRB	N ResNet	L SG, Relevance-CAM, ..

biased sampling of the data points used to approximate explanations (an algorithm introduced in Fukuchi et al., 2020). Experiments show an improvement in manipulating SHAP over previous work of Baniecki & Biecek (2022), which further underlines SHAP' vulnerability (Slack et al., 2020).

Explanations might be exploited to breach privacy. Shokri et al. (2021) introduce membership inference attacks that use information from feature attribution explanations to determine whether a data point was present in the training dataset. Pawelczyk et al. (2023b) propose membership inference attacks using counterfactual explanations instead.

Observation 2. Attacking local explanations may impact the model's behaviour globally (Heo et al., 2019; Dimanov et al., 2020; Anders et al., 2020; Noppel et al., 2023), and vice versa, fooling global explanations may manipulate local explanations in the process (Lakkaraju & Bastani, 2020; Baniecki & Biecek, 2022; Laberge et al., 2023). Both of these interactions could improve detectability in practice.

Observation 3. To this date, there are relatively sparse studies concerning adversarial attacks on concept-based explanations, e.g. Brown & Kvinge (2022) attack TCAV and FFV (Goh et al., 2021), counterfactual explanations, e.g. Slack et al. (2021a) attack counterfactuals for neural networks (Guidotti, 2022), and overall explanations for language models, e.g. Sinha et al. (2021) attack IG and LIME.

Observation 4. Research on attacking explanations for image classification relies on a few popular datasets, e.g. ImageNet (Deng et al., 2009) reoccurs in 8 out of 11 studies. In parallel, a larger variety of tabular scenarios is tested.

While many contributions consider the detectability of the attack (Subramanya et al., 2019; Kuppa & Le-Khac, 2020; Zhang et al., 2020), and some propose ways of mitigating the attacks' effects via robustifying mechanisms (Dombrowski et al., 2019; Anders et al., 2020; Zhang et al., 2020), we next systemize contributions that mostly focus on defending.

3. Defense against the attacks on explanations

Whenever a new attack algorithm is introduced in adversarial machine learning, various ways to address the explanation's limitations and fix its insecurities are proposed. Chen et al. (2019b) is one of the first attempts to defend from adversarial examples introduced by Ghorbani et al. (2019) via regularizing a neural network. The proposed robust attribution regularization forces IG explanations to remain unchanged under perturbation attacks. Rieger & Hansen (2020) propose an alternative defense strategy against such

Table 2. Summary of defenses against the attacks on explanations of machine learning models. Each work on explanations' robustness is connected with *up to two* attacks that are mentioned to be potentially addressed by it. We abbreviate the following: data (D), model (M), image (I), tabular (T), language (Lg), neural network (N), black-box (B), local (L), global (G). Appendix A lists other abbreviations.

Defense	**Attack(s)**	**Modality** dataset	**Model** algorithm	**Explanation** algorithm
(Woods et al., 2019)	–	I ImageNet, COCO, ..	N ResNet	L Grad-CAM
(Chen et al., 2019b)	(Ghorbani et al., 2019)	I FashionMNIST, Flower, ..	N CNN, ResNet	L IG
(Rieger & Hansen, 2020)	(Ghorbani et al., 2019) (Dombrowski et al., 2019)	I ImageNet	N VGG	L SG, IG, LRP, GBP
(Boopathy et al., 2020)	(Ghorbani et al., 2019) (Dombrowski et al., 2019)	I MNIST, CIFAR-10, ..	N CNN, ResNet	L IG, CAM, Grad-CAM
(Lakkaraju et al., 2020)	(Ghorbani et al., 2019) (Lakkaraju & Bastani, 2020)	T Bail, Academic, Health	B MLP, RF, GBDT, ..	G MUSE, LIME, SHAP
(Wang et al., 2020)	(Ghorbani et al., 2019) (Dombrowski et al., 2019)	I CIFAR-10, ImageNet, Flower	N ResNet	L SG, IG, SmoothGrad
(La Malfa et al., 2021)	–	Lg IMDB, SST, Twitter	N MLP, CNN	L Anchors
(Ghalebikesabi et al., 2021)	(Slack et al., 2020)	T COMPAS, Adult, Bike, .. I MNIST	B XGBoost, CNN	L SHAP, GradSHAP
(Schneider et al., 2022)	–	Lg IMDB, WoS	B CNN	L Grad-CAM
(Shrotri et al., 2022)	(Slack et al., 2020)	T Credit, COMPAS, Crime, ..	B RF	L LIME
(Dombrowski et al., 2022)	(Dombrowski et al., 2019)	I CIFAR-10, ImageNet	N CNN, VGG, ResNet	L SG, GI, IG, LRP, ..
(Tang et al., 2022)	(Ghorbani et al., 2019) (Dombrowski et al., 2019)	I MNIST, FashionMNIST	N CNN	L SG
(Vreš & Robnik-Šikonja, 2022)	(Slack et al., 2020)	T Credit, COMPAS, Crime	B MLP, RF, SVM, ..	L LIME, SHAP, IME
(Liu et al., 2022)	(Ghorbani et al., 2019)	I VOC2007	N VGG	L SG
(Carmichael & Scheirer, 2023)	(Slack et al., 2020)	T Credit, COMPAS, Crime	B rule set	L SHAP, LIME
(Joo et al., 2023)	(Ghorbani et al., 2019) (Dombrowski et al., 2019)	I CIFAR-10, ImageNet	N ResNet, LeNet	L SG, GI, LRP, GBP
(Virgolin & Fracaros, 2023)	(Slack et al., 2021a)	T Credit, COMPAS, Adult, ..	B MLP, RF	L counterfactual
(Wicker et al., 2023)	(Heo et al., 2019) (Dombrowski et al., 2019)	T Credit, Adult, I MNIST, MedMNIST	N MLP, CNN	L GI, DeepLIFT, GradSHAP
(Pawelczyk et al., 2023a)	(Slack et al., 2021a)	T Credit, COMPAS, Adult	N LR, MLP	L counterfactual

adversarial examples (Ghorbani et al., 2019; Dombrowski et al., 2019), i.e. aggregating multiple explanations created with various algorithms. As the attack targets only a single explanation method, their aggregated mean remains close to the original explanation (shown in Figure 1).

Table 2 lists defenses against the attacks on explanations, where for each, we record the datasets, models and explanation algorithms mentioned in experiments. Excluded from it are works that improve explanation robustness without directly relating to the potential adversarial attack scenario (e.g. see Yeh et al., 2019; Zhou et al., 2021; Zhao et al., 2021; Slack et al., 2021b, and references given there). We link each defense with an attack, but omit to list all attacks potentially addressed by the defense for brevity. The three missing links are worth clarifying here. Woods et al. (2019) is an early work that introduces adversarial explanations, which have improved robustness against adversarial examples targeting model predictions. Similarly, La Malfa et al. (2021) proposes to improve explanations of language models against adversarial perturbations. Unlike most of the contributions that focus on algorithms, Schneider et al. (2022) conduct a user study with artificially manipulated explanations to evaluate if humans can discover the potential attack in practice (related to Lakkaraju & Bastani, 2020; Poursabzi-Sangdeh et al., 2021). Pawelczyk et al. (2023a) and Virgolin & Fracaros (2023) introduce mechanisms to improve the robustness of counterfactual explanations against adversarial perturbations (we link the latter with an otherwise unreferenced attack of Slack et al. (2021a)).

Boopathy et al. (2020) extend the regularization training method of Chen et al. (2019b) to use an l_1-norm 2-class interpretation discrepancy measure. Experiments show an improvement in effectiveness and computation cost when defending explanations over previous work (including Chen et al., 2019b). Moreover, achieving robust explanations alone improves prediction robustness when explanations are compared with the proposed measure. Wang et al. (2020) introduce a smooth surface regularization procedure to force robust attributions by minimizing the difference between explanations for nearby points. Experiments show a tradeoff between regularization performance and computation cost (also with respect to Chen et al., 2019b). Also, models with smoothed geometry become less susceptible to transfer attacks, i.e. where an adversary targeting one explanation method fools other gradient-based explanations as well. Dombrowski et al. (2022); Tang et al. (2022) further compare and extend the in-training techniques to regularize neural networks towards improving explanation robustness (also mentioned in Dombrowski et al., 2019). Dombrowski et al. (2022) use the approximated norm of the Hessian as a regularization term during training to bound the l_2-distance between the gradients of the original and perturbed samples, which benefits gradient-based explanations. Joo et al. (2023) propose to improve this approach by introducing a cosine robust criterion to measure the $cosine$-distance instead. As shown in experiments comparing the two distance measures, it effectively solves issues with normalizing gradient-based attribution values that are used when interpreting predictions in practice. Most recent work in this line of research introduces certifiably robust explanations of neural networks (Liu et al., 2022; Wicker et al., 2023), which reassure that no adversarial explanation exists for a given set of input or model weights.

In parallel to defending adversarial attacks on gradient-based explanations of neural networks, several works address the possibility of fooling model-agnostic LIME and SHAP (Slack et al., 2020). Ghalebikesabi et al. (2021) modifies SHAP estimator by sampling data from a local neighbourhood distribution instead of the marginal or conditional global reference distribution. Experiments show that such constructed on-manifold explainability improves explanations' robustness, i.e. SHAP defends from the attack. Shrotri et al. (2022) modifies LIME estimator to take into account user-specified constraints on the input space that restrict the allowed data perturbations. Alike, experiments show that constrained explanations are less susceptible to out-of-distribution attacks. Moreover, analysing differences between the original and constrained explanations allows for detecting an adversarially discriminative classifier. In contrast, Vreš & Robnik-Šikonja (2022) introduce focused sampling with various data generators to improve the adversarial robustness of both LIME and SHAP. Instead of directly improving perturbation-based explanation methods, Carmichael & Scheirer (2023) propose to "unfool" explanations with conditional anomaly detection. An algorithm based on k-nearest neighbours scores the abnormality of input samples conditioned on their classification labels. Comparing the empirical distribution function of scores between the original and potentially adversarially perturbed samples given a user-defined threshold proves to be effective for attack detection. Removing abnormal samples from the perturbed input set defends an explanation against fooling. We summarize the described XAI failure modes in Figure 2.

Observation 5. Comparing Tables 1 & 2 highlights existing insecurities in XAI methods; namely, not clearly addressed are backdooring attacks (Viering et al., 2019; Noppel et al., 2023), data poisoning attacks (Zhang et al., 2021; Baniecki et al., 2022), attacks specific to language (Sinha et al., 2021) and concept-based explainability (Brown & Kvinge, 2022).

Observation 6. To this date, there are sparse studies concerning defenses against attacks on global explanations, e.g. Lakkaraju et al. (2020) robustify model-agnostic global explanations against a general class of distribution shifts related to adversarial perturbations (Ghorbani et al., 2019).

4. Adversarial attacks on fairness metrics

Closely related to adversarial attacks on explanations are attacks on machine learning fairness metrics, e.g. predictive equality (Corbett-Davies et al., 2017) and (statistical) demographic parity (refer to Mehrabi et al., 2021a, for an introduction to machine learning fairness). Intuitively, algorithms targeting model predictions and accuracy can be applied to manipulate other functions of the model output as well. Table 3 in Appendix B lists a representative set of adversarial attacks on group fairness metrics, with the corresponding strategy of changing data (Fukuchi et al., 2020), the model (Aivodji et al., 2019; 2021), or jointly changing data and the model (Solans et al., 2020; Mehrabi et al., 2021b; Hussain et al., 2022).

Fukuchi et al. (2020) introduce a stealthily biased sampling procedure to adversarially craft an unbiased dataset used to estimate fairness metrics. It is formally defined as a Wasserstein distance minimization problem and solved with an efficient algorithm for a minimum-cost flow problem in practice. Experiments focus on quantifying the trade-off between lowering the perceived model bias and detecting adversarial sampling. Contrary, Solans et al. (2020) introduce a data poisoning attack to increase bias as measured with fairness metrics via adding data points to the training dataset so that the model discriminates against a certain group of individuals. While the approach relies on the differentiation of neural networks for optimization, experiments show the transferability of data poisoning to other algorithms in a black-box setting. Mehrabi et al. (2021b) propose alternative data poisoning attacks, also in both black-box and white-box settings relying on gradient computation to opti-

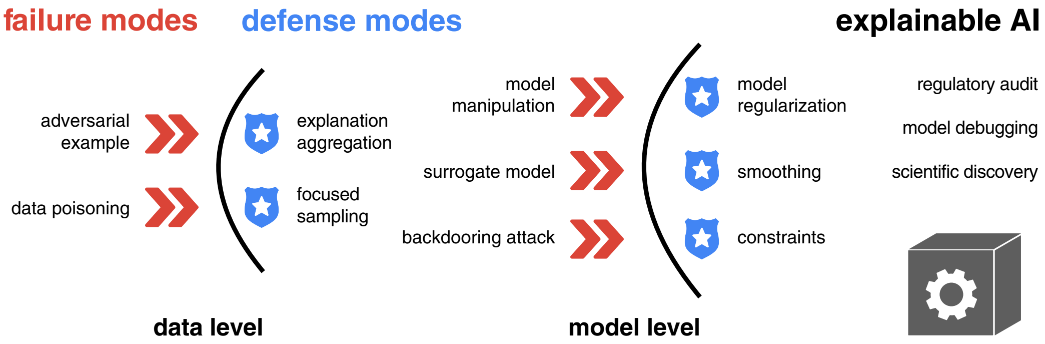

Figure 2. Summary of the possible XAI failure modes on data and model levels.

mize a loss function for data sampling. Experiments show an improvement in manipulating fairness over previous work of Solans et al. (2020). Hussain et al. (2022) extend data poisoning attacks on fairness to the task of node classification with graph convolutional networks. Aivodji et al. (2019; 2021) introduce fairwashing attacks changing the model, i.e. an adversary approximates an unfair black-box model with a faithful surrogate model appearing as fair. Experiments analyse the fidelity-unfairness trade-off and the effect of fairwashing on impacting feature effect model explanations. Further related is work concerning attacking fairness in imaging, where Nanda et al. (2021) introduce the notion of robustness bias, which requires all groups to be equally susceptible to adversarial attacks. Ferry et al. (2023) consider a different adversarial strategy and introduce a dataset extraction attack to retrieve a piece of information about a sensitive attribute based on fairness criteria. Related are membership inference attacks on privacy using explanations (Shokri et al., 2021; Pawelczyk et al., 2023b).

Observation 7. Research on attacking fairness methods focuses on the notion of group fairness, i.e. treating different groups of inputs equally, and omits subgroup or individual fairness, i.e. predicting similarly for similar individuals (see the distinction in Mehrabi et al., 2021a, table 1).

5. Frontier research directions in AdvXAI

We conclude by outlining research directions in AdvXAI.

Attacks. Currently most exploited by the attacks are the first-introduced and most popular XAI methods, e.g. SHAP and Grad-CAM. Future work on adversarial attacks may consider targeting the more recent enhancements that aim to overcome their limitations, e.g. SHAPR that takes into account feature dependence in tabular data (Aas et al., 2021) or Shap-CAM for improved explanations of convolutional neural networks (Zheng et al., 2022). Alike model-specific explanations of neural networks, worth assessing is the vulnerability of explanation methods specific to tree-based models, e.g. TreeSHAP (Lundberg et al., 2020), but also white-box attacks on explanations assuming prior knowledge that the model is an ensemble of decision trees. Beyond post-hoc explainability, adversarial attacks could target vulnerabilities of the interpretable by-design deep learning models like ProtoPNet (Chen et al., 2019a) and its extensions (e.g. see Rymarczyk et al., 2022, and its related work). Finally, there are adversarial attacks on model predictions that actively aim to bypass through a particular defense mechanism (e.g. see Machado et al., 2021, table 4) and such a threat of circumventing the defense in XAI is currently unexplored.

AdvXAI beyond classical models towards transformers. Nowadays, the transformer architecture is the frontier of machine learning research and applications of deep learning in practice. Thus, the adversarial robustness of explanations of large models for various modalities like GPT (Bubeck et al., 2023), ViT (Dosovitskiy et al., 2021), and TabPFN (Hollmann et al., 2023) deserves special attention. For example, Ali et al. (2022) extend LRP explanations to transformers, which might propagate the explanations' vulnerability to adversarial attacks as shown in previous work (Heo et al., 2019; Anders et al., 2020). We acknowledge that the recently proposed transformer-based foundation models, e.g. SAM (Kirillov et al., 2023), more and more frequently include benchmarks specific to evaluating responsibility, e.g. whether segmenting people from images is unbiased with respect to their perceived gender presentation, age group or skin tone (Schumann et al., 2021). The possibility of attacking such fairness measurements by biased sampling becomes a trust issue (Fukuchi et al., 2020).

AdvXAI beyond the image and tabular data modalities. A majority of contributions surveyed here, so as XAI, concern machine learning predictive models trained on imaging and tabular datasets. Further work is required to evaluate which and how severe are adversarial attacks concerning other data modalities like language (La Malfa et al., 2021;

Schneider et al., 2022), graphs (Hussain et al., 2022), time series, multimodal systems, and explanations of reinforcement learning agents (Olson et al., 2021).

Future work. One goal of this survey is to reiterate the apparent insecurities in XAI, i.e. the unaddressed attacks on explanation methods (Viering et al., 2019; Zhang et al., 2021; Brown & Kvinge, 2022; Baniecki et al., 2022; Noppel et al., 2023; Laberge et al., 2023). We also underline that a possibility of manipulating fairness metrics has detrimental consequences when applied in audit and law enforcement, and therefore developing metrics robust against the attacks is desirable. Note that although a particular XAI method is attacked or defended, in fact, the evidence of model predictions is in question here. Our future goal is to categorise the surveyed attack and defense mechanisms in a way to guide practitioners in which scenarios it is secure to use a given model and explanation, e.g. when a researcher uses XGBoost with SHAP instead of logistic regression for scientific discovery.

Ethics, impact on society, and law concerning AdvXAI. Finally, we need to take into account the broader impact adversarial research has on society. How does AdvXAI fit into regulations like AI Act (Floridi, 2021), the four-fifths rule of fairness (Watkins et al., 2022), or the right to explanation (Krishna et al., 2023)? These questions are yet to be answered. For a more philosophical consideration on explanation robustness, we refer the reader to the argument by Hancox-Li (2020) concerning epistemic and ethical reasons for seeking objective explanations.

Acknowledgements

We thank anonymous reviewers for their valuable feedback. This work was financially supported by the Polish National Science Centre grant number 2021/43/O/ST6/00347.

References

Aas, K., Jullum, M., and Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298:103502, 2021.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9505–9515, 2018.

Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., and Tapp, A. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, volume 97, pp. 161–170, 2019.

Aivodji, U., Arai, H., Gambs, S., and Hara, S. Characterizing the risk of fairwashing. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14822–14834, 2021.

Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.-R., and Wolf, L. XAI for Transformers: Better Explanations through Conservative Propagation. In *International Conference on Machine Learning*, volume 162, pp. 435–451, 2022.

Alvarez Melis, D. and Jaakkola, T. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 7775–7784, 2018.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *International Conference on Learning Representations*, 2018.

Anders, C., Pasliev, P., Dombrowski, A.-K., Müller, K.-R., and Kessel, P. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pp. 314–323, 2020.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):1–46, 2015.

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

Baniecki, H., Kretowicz, W., and Biecek, P. Fooling Partial Dependence via Data Poisoning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 121–136, 2022.

Boopathy, A., Liu, S., Zhang, G., Liu, C., Chen, P.-Y., Chang, S., and Daniel, L. Proper Network Interpretability Helps Adversarial Robustness in Classification. In *International Conference on Machine Learning*, volume 119, pp. 1014–1023, 2020.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *Annual Workshop on Computational Learning Theory*, pp. 144–152, 1992.

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

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*. Taylor & Francis, 1984. ISBN 9780412048418.

Brown, D. and Kvinge, H. Making Corgis Important for Honeycomb Classification: Adversarial Attacks on Concept-based Explainability Tools. In *ICML Workshop on New Frontiers in Adversarial Machine Learning*, 2022. Preprint at https://doi.org/10.48550/arXiv.2110.07120.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of Artificial General Intelligence: Early experiments with GPT-4, 2023. Preprint at https://doi.org/10.48550/arXiv.2303.12712.

Carmichael, Z. and Scheirer, W. J. Unfooling Perturbation-Based Post Hoc Explainers. In *AAAI Conference on Artificial Intelligence*, 2023.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8930–8941, 2019a.

Chen, J., Wu, X., Rastogi, V., Liang, Y., and Jha, S. Robust Attribution Regularization. In *Advances in Neural Information Processing Systems*, volume 32, pp. 14302–14312, 2019b.

Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic Decision Making and the Cost of Fairness. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, 2017.

Dabkowski, P. and Gal, Y. Real Time Image Saliency for Black Box Classifiers. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6970–6979, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Dimanov, B., Bhatt, U., Jamnik, M., and Weller, A. You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods. In *European Conference on Artificial Intelligence*, volume 97, pp. 161–170, 2020.

Dombrowski, A.-K., Alber, M., Anders, C., Ackermann, M., Müller, K.-R., and Kessel, P. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13589–13600, 2019.

Dombrowski, A.-K., Anders, C., Müller, K.-R., and Kessel, P. Towards robust explanations for deep neural networks. *Pattern Recognition*, 121:108194, 2022.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.

Ferry, J., Aïvodji, U., Gambs, S., Huguet, M.-J., and Siala, M. Exploiting Fairness to Enhance Sensitive Attributes Reconstruction. In *IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.

Floridi, L. The European Legislation on AI: A brief analysis of its philosophical approach. *Philosophy & Technology*, 34:1–8, 2021.

Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

Fukuchi, K., Hara, S., and Maehara, T. Faking Fairness via Stealthily Biased Sampling. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 412–419, 2020.

Ghalebikesabi, S., Ter-Minassian, L., DiazOrdaz, K., and Holmes, C. C. On Locality of Local Explanation Models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18395–18407, 2021.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of Neural Networks Is Fragile. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019.

Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., and Olah, C. Multimodal Neurons in Artificial Neural Networks. *Distill*, 2021.

Gu, A., Weng, T.-W., Chen, P.-Y., Liu, S., and Daniel, L. Certified Interpretability Robustness for Class Activation Mapping. In *NeurIPS Workshop on Machine Learning for Autonomous Driving*, 2020. Preprint at https://doi.org/10.48550/arXiv.2301.11324.

Guidotti, R. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.

Hancox-Li, L. Robustness in machine learning explanations: Does it matter? In *ACM Conference on Fairness, Accountability, and Transparency*, pp. 640–647, 2020.

Hardt, M., Price, E., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 29, pp. 332–3331, 2016.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 1 edition, 2001.

Heo, J., Joo, S., and Moon, T. Fooling Neural Network Interpretations via Adversarial Model Manipulation. In *Advances in Neural Information Processing Systems*, volume 32, pp. 2925–2936, 2019.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *International Conference on Learning Representations*, 2023.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., and Samek, W. *xxAI - Beyond Explainable AI*, chapter Explainable AI Methods – A Brief Overview, pp. 13–38. Springer, 2022.

Hussain, H., Cao, M., Sikdar, S., Helic, D., Lex, E., Strohmaier, M., and Kern, R. Adversarial Inter-Group Link Injection Degrades the Fairness of Graph Neural Networks. In *EEE International Conference on Data Mining*, pp. 975–980, 2022.

Jiménez-Luna, J., Grisoni, F., and Schneider, G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020.

Joo, S., Jeong, S., Heo, J., Weller, A., and Moon, T. Towards More Robust Interpretation via Local Gradient Alignment. In *AAAI Conference on Artificial Intelligence*, 2023.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and sayres, R. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning*, volume 80, pp. 2668–2677, 2018.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, chapter The (Un)reliability of Saliency Methods, pp. 267–280. Springer, 2019.

Kipf, T. N. and Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*, 2017.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., and Girshick, R. Segment Anything, 2023. Preprint at https://doi.org/10.48550/arXiv.2304.02643.

Kolter, Z. and Madry, A. Adversarial robustness: Theory and practice. *Tutorial at NeurIPS*, 2018.

Krishna, S., Ma, J., and Lakkaraju, H. Towards Bridging the Gaps between the Right to Explanation and the Right to be Forgotten, 2023. Preprint at https://doi.org/10.48550/arXiv.2302.04288.

Kuppa, A. and Le-Khac, N.-A. Black Box Attacks on Explainable Artificial Intelligence(XAI) methods in Cyber Security. In *International Joint Conference on Neural Networks*, pp. 1–8, 2020.

La Malfa, E., Michelmore, R., Zbrzezny, A. M., Paoletti, N., and Kwiatkowska, M. On Guaranteed Optimal Robust Explanations for NLP Models. In *International Joint Conference on Artificial Intelligence*, 2021.

Laberge, G., Aivodji, U., Hara, S., and Mario Marchand, F. K. Fooling SHAP with Stealthily Biased Sampling. In *International Conference on Learning Representations*, 2023.

Lakkaraju, H. and Bastani, O. "How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations. In *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 79–85, 2020.

Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. Faithful and Customizable Explanations of Black Box Models. In *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 131–138, 2019.

Lakkaraju, H., Arsov, N., and Bastani, O. Robust and Stable Black Box Explanations. In *International Conference on Machine Learning*, 2020.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lee, J. R., Kim, S., Park, I., Eo, T., and Hwang, D. Relevance-CAM: Your Model Already Knows Where to Look. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14939–14948, 2021.

Liu, A., Chen, X., Liu, S., Xia, L., and Gan, C. Certifiably robust interpretation via rényi differential privacy. *Artificial Intelligence*, 313:103787, 2022.

Liu, N., Du, M., Guo, R., Liu, H., and Hu, X. Adversarial Attacks and Defenses: An Interpretation Perspective. *ACM SIGKDD Explorations Newsletter*, 23(1):86–99, 2021.

Lundberg, S. M. and Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774, 2017.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

Machado, G. R., Silva, E., and Goldschmidt, R. R. Adversarial Machine Learning in Image Classification: A Survey Toward the Defender's Perspective. *ACM Computing Surveys*, 55(1), 2021.

McGrath, T., Kapishnikov, A., Tomašev, N., Pearce, A., Wattenberg, M., Hassabis, D., Kim, B., Paquet, U., and Kramnik, V. Acquisition of chess knowledge in AlphaZero. *Proceedings of the National Academy of Sciences*, 119 (47):e2206625119, 2022.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021a.

Mehrabi, N., Naveed, M., Morstatter, F., and Galstyan, A. Exacerbating Algorithmic Bias through Fairness Attacks. In *AAAI Conference on Artificial Intelligence*, volume 35, pp. 8930–8938, 2021b.

Merrer, E. L. and Trédan, G. Remote explainability faces the bouncer problem. *Nature Machine Intelligence*, 2: 529–539, 2020.

Mishra, S., Dutta, S., Long, J., and Magazzeni, D. A Survey on the Robustness of Feature Importance and Counterfactual Explanations. In *ICAIF Workshop on Explainable AI in Finance*, 2021. Preprint at https://doi.org/10.48550/arXiv.2111.00358.

Nanda, V., Dooley, S., Singla, S., Feizi, S., and Dickerson, J. P. Fairness Through Robustness: Investigating Robustness Disparity in Deep Learning. In *ACM Conference on Fairness, Accountability, and Transparency*, pp. 466–477, 2021.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, 2023.

Noppel, M., Peter, L., and Wressnegger, C. Disguising Attacks with Explanation-Aware Backdoors. In *IEEE Symposium on Security and Privacy*, pp. 996–1013, 2023.

Olson, M. L., Khanna, R., Neal, L., Li, F., and Wong, W.-K. Counterfactual state explanations for reinforcement learning agents via generative deep learning. *Artificial Intelligence*, 295:103455, 2021.

Papernot, N., McDaniel, P., Sinha, A., and Wellman, M. P. SoK: Security and Privacy in Machine Learning. In *IEEE European Symposium on Security and Privacy*, pp. 399–414, 2018.

Pawelczyk, M., Datta, T., van-den Heuvel, J., Kasneci, G., and Lakkaraju, H. Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. In *International Conference on Learning Representations*, 2023a.

Pawelczyk, M., Lakkaraju, H., and Neel, S. On the privacy risks of algorithmic recourse. In *International Conference on Artificial Intelligence and Statistics*, 2023b.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. Manipulating and Measuring Model Interpretability. In *CHI Conference on Human Factors in Computing Systems*, pp. 237, 2021.

Ribeiro, M. T., Singh, S., and Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. *AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Rieger, L. and Hansen, L. K. A simple defense against adversarial attacks on heatmap explanations. In *ICML Workshop on Human Interpretability in Machine Learning*, 2020. Preprint at https://doi.org/10.48550/arXiv.2007.06381.

Rosenberg, I., Shabtai, A., Elovici, Y., and Rokach, L. Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. *ACM Computing Surveys*, 54(5), 2021.

Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1:206–215, 2019.

Rymarczyk, D., Struski, Ł., Górszczak, M., Lewandowska, K., Tabor, J., and Zieliński, B. Interpretable image classification with differentiable prototypes assignment. In *European Conference on Computer Vision*, pp. 351–368, 2022.

Schneider, J., Meske, C., and Vlachos, M. Deceptive AI Explanations: Creation and Detection. In *International Conference on Agents and Artificial Intelligence*, volume 2, pp. 44–55, 2022.

Schumann, C., Ricco, S., Prabhu, U., Ferrari, V., and Pantofaru, C. A Step Toward More Inclusive People Annotations for Fairness. In *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 916–925, 2021.

Schwalbe, G. and Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 2023.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2): 336–359, 2020.

Shokri, R., Strobel, M., and Zick, Y. On the Privacy Risks of Model Explanations. In *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 231–241, 2021.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning Important Features through Propagating Activation Differences. In *International Conference on Machine Learning*, volume 70, pp. 3145–3153, 2017.

Shrotri, A. A., Narodytska, N., Ignatiev, A., Meel, K. S., Marques-Silva, J., and Vardi, M. Y. Constraint-driven explanations for black-box ml models. In *AAAI Conference on Artificial Intelligence*, volume 36, pp. 8304–8314, 2022.

Simonyan, K., Vedaldi, A., and Zisserman:, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *ICLR Workshops*, 2014.

Sinha, S., Chen, H., Sekhon, A., Ji, Y., and Qi, Y. Perturbing Inputs for Fragile Interpretations in Deep Natural Language Processing. In *EMNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 420–434, 2021.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Slack, D., Hilgard, S., Lakkaraju, H., and Singh, S. Counterfactual Explanations Can Be Manipulated. In *Advances in Neural Information Processing Systems*, volume 34, pp. 62–75, 2021a.

Slack, D., Hilgard, S., Singh, S., and Lakkaraju, H. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9391–9404, 2021b.

Solans, D., Biggio, B., and Castillo, C. Poisoning Attacks on Algorithmic Fairness. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 162–177, 2020.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for Simplicity: The All Convolutional Net. In *ICLR Workshops*, 2015.

Štrumbelj, E. and Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41:647–665, 2014.

Subramanya, A., Pillai, V., and Pirsiavash, H. Fooling Network Interpretation in Image Classification. In *IEEE/CVF International Conference on Computer Vision*, pp. 2020–2029, 2019.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic Attribution for Deep Networks. In *International Conference on Machine Learning*, volume 70, pp. 3319–3328, 2017.

Tang, R., Liu, N., Yang, F., Zou, N., and Hu, X. Defense Against Explanation Manipulation. *Frontiers in Big Data*, 5, 2022.

Viering, T., Wang, Z., Loog, M., and Eisemann, E. How to Manipulate CNNs to Make Them Lie: the GradCAM Case. In *BMVC Workshop on Interpretable and Explainable Machine Vision*, 2019. Preprint at https://doi.org/10.48550/arXiv.1907.10901.

Virgolin, M. and Fracaros, S. On the robustness of sparse counterfactual explanations to adverse perturbations. *Artificial Intelligence*, 316:103840, 2023.

Vreš, D. and Robnik-Šikonja, M. Preventing deception with explanation methods using focused sampling. *Data Mining and Knowledge Discovery*, 2022.

Wang, Z., Wang, H., Ramkumar, S., Mardziel, P., Fredrikson, M., and Datta, A. Smoothed Geometry for Robust Attribution. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13623–13634, 2020.

Watkins, E. A., McKenna, M., and Chen, J. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness, 2022. Preprint at https://doi.org/10.48550/arXiv.2202.09519.

Weerts, H., Dudík, M., Edgar, R., Jalali, A., Lutz, R., and Madaio, M. Fairlearn: Assessing and Improving Fairness of AI Systems, 2023. Preprint at https://doi.org/10.48550/arXiv.2303.16626.

Wicker, M. R., Heo, J., Costabello, L., and Weller, A. Robust Explanation Constraints for Neural Networks. In *International Conference on Learning Representations*, 2023.

Woods, W., Chen, J., and Teuscher, C. Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nature Machine Intelligence*, 1(11):508–516, 2019.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (In)fidelity and Sensitivity of Explanations. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10967–10978, 2019.

Zhang, H., Gao, J., and Su, L. Data Poisoning Attacks Against Outcome Interpretations of Predictive Models. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2165–2173, 2021.

Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., and Wang, T. Interpretable Deep Learning under Fire. In *USENIX Security Symposium*, pp. 1659–1676, 2020.

Zhao, X., Huang, W., Huang, X., Robu, V., and Flynn, D. BayLIME: Bayesian local interpretable model-agnostic explanations. In *Conference on Uncertainty in Artificial Intelligence*, volume 161, pp. 887–896, 2021.

Zheng, Q., Wang, Z., Zhou, J., and Lu, J. Shap-CAM: Visual Explanations for Convolutional Neural Networks Based on Shapley Value. In *European Conference on Computer Vision*, pp. 459–474, 2022.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning Deep Features for Discriminative Localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2921–2929, 2016.

Zhou, Z., Hooker, G., and Wang, F. S-LIME: Stabilized-LIME for Model Explanation. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2429–2438, 2021.

A. List of abbreviations and proper names (with references)

- Anchors: high-precision model-agnostic explanations (Ribeiro et al., 2018)
- CAM: class activation mapping (Zhou et al., 2016)
- CNN: convolutional neural network (Lecun et al., 1998)
- DT: decision tree (Breiman et al., 1984)
- EOdds: equalized odds (Hardt et al., 2016)
- EOpp: equal opportunity (Hardt et al., 2016)
- `fairlearn` software (Weerts et al., 2023)
- FFV: faceted feature visualization (Goh et al., 2021)
- GBDT: gradient boosting decision tree (Friedman, 2001)
- GBP: guided backpropagation (Springenberg et al., 2015)
- GCN: graph convolutional network (Kipf & Welling, 2017)
- GI: gradient input (Shrikumar et al., 2017)
- Grad-CAM: gradient-weighted class activation mapping (Selvaraju et al., 2020)
- IG: integrated gradients (Sundararajan et al., 2017)
- IME: interactions-based method for explanation (Štrumbelj & Kononenko, 2014)
- KNN: k-nearest neighbors (see Hastie et al., 2001, section 13.3)
- LIME: local interpretable model-agnostic explanations (Ribeiro et al., 2016)
- LR: logistic regression (see Hastie et al., 2001, section 4.4)
- LRP: layer-wise relevance propagation (Bach et al., 2015)
- MLP: multi-layer perceptron (see Hastie et al., 2001, chapter 11)
- MUSE: model understanding subspace explanations (Lakkaraju et al., 2019)
- NB: naïve Bayes (see Hastie et al., 2001, section 6.6.3)
- PDP: partial dependence plot (introduced in Friedman, 2001, section 8.2)
- PE: predictive equality (Corbett-Davies et al., 2017)
- Relevance-CAM: relevance-weighted class activation mapping (Lee et al., 2021)
- RF: random forest (Breiman, 2001)
- RTS: real time saliency (Dabkowski & Gal, 2017)
- SG: simple gradient (Simonyan et al., 2014)
- SHAP: Shapley additive explanations (Lundberg & Lee, 2017)
- SP: statistical (demographic) parity (see Mehrabi et al., 2021a, section 4.1)
- SVM: support vector machine (Boser et al., 1992)
- TCAV: testing with concept activation vectors (Kim et al., 2018)
- ViT: vision transformer (Dosovitskiy et al., 2021)
- XGBoost (Chen & Guestrin, 2016)

Table 3. Related work concerning adversarial attacks on fairness metrics of machine learning models. We abbreviate the following: data (D), model (M), image (I), tabular (T), graph (Gr), neural network (N), black-box (B), group (G). Appendix A lists other abbreviations.

Attack	**Changes** strategy	**Modality** dataset	**Model** algorithm	**Fairness** metric
(Aivodji et al., 2019)	M surrogate model	T COMPAS, Adult	B RF	G SP
(Fukuchi et al., 2020)	D data poisoning	T COMPAS, Adult	B RF, LR	G SP
(Solans et al., 2020)	D & M data poisoning	T COMPAS	B LR, RF, SVM, DT, NB	G SP, EOdds
(Mehrabi et al., 2021b)	D & M data poisoning	T Credit, COMPAS, Drug	B MLP	G SP, EOdds
(Nanda et al., 2021)	D, M adversarial example	I Adience, UTKFace, ..	N VGG, ResNet, DenseNet, ..	G robustness bias
(Aivodji et al., 2021)	M surrogate model	T Credit, COMPAS, Adult, ..	B MLP, RF, AdaBoost, XGBoost	G SP, PE, EOdds, EOpp
(Hussain et al., 2022)	D & M data poisoning	Gr Pokec, DBLP	N GCN	G SP, EOdds, EOpp
(Ferry et al., 2023)	– data reconstruction	T ACSIncome, ACSPublicCoverage	B DT+`fairlearn`	G SP, PE, EOdds, EOpp

B. Adversarial attacks on fairness metrics

See Table 3.