SoK: Unintended Interactions among Machine Learning Defenses and Risks

Vasisht Duddu*, Sebastian Szyller†, N. Asokan*‡
*University of Waterloo, †Intel Labs, ‡Aalto University
vasisht.duddu@uwaterloo.ca, contact@sebszyller.com, asokan@acm.org

2024 IEEE Symposium on Security and Privacy (SP) | 979-8-3503-3130-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/SP54263.2024.00243

***Abstract*—Machine learning (ML) models cannot neglect risks to security, privacy, and fairness. Several defenses have been proposed to mitigate such risks. When a defense is effective in mitigating one risk, it may correspond to increased or decreased susceptibility to *other risks*. Existing research lacks an effective framework to recognize and explain these *unintended interactions*. We present such a framework, based on the conjecture that *overfitting* and *memorization* underlie unintended interactions. We survey existing literature on unintended interactions, accommodating them within our framework. We use our framework to conjecture two previously unexplored interactions, and empirically validate them.**

1. Introduction

Potential *risks* to security, privacy, and fairness of machine learning (ML) models have led to various proposed *defenses* [42], [65], [96], [103], [106], [115], [124], [162]. When a defense is effective against a specific risk, it may correspond to an increased or decreased *susceptibility to other risks* [61], [153]. For instance, adversarial training for increasing robustness also increases susceptibility to *membership inference* [70], [151]. We refer to these as *unintended interactions* between defenses and risks.

Foreseeing such unintended interactions is challenging. A unified framework enumerating different defenses and risks can help (a) researchers identify unexplored interactions and design algorithms with better trade-offs, and (b) practitioners to account for such interactions before deployment. However, prior works are limited to a specific risk (e.g., membership inference [153]), defense (e.g., fairness with privacy risks [32]) or interaction (e.g., explanations with security risks [121]) and do not systematically study the underlying causes (e.g., Gittens et al. [61]). A comprehensive framework spanning multiple defenses and risks, to systematically identify potential unintended interactions is currently absent.

We aim to fill this gap by systematically examining various unintended interactions spanning multiple defenses and risks. We conjecture that *overfitting* and *memorization* of training data are the potential causes underlying these unintended interactions. An effective defense may induce, reduce or rely on overfitting or memorization which in turn, impacts the model's susceptibility to other risks. We identify different factors like the characteristics of the training dataset, objective function, and the model, that collectively influence a model's propensity to overfit or memorize. These factors help us get a nuanced understanding of the susceptibility to different risks when a defense is in effect. We claim the following main contributions:

1) the *first systematic framework* for understanding unintended interactions, by their underlying causes and factors that influence them. (Section 3)
2) a literature *survey* to identify different unintended interactions, situating them within our framework, and a guideline to use our framework to conjecture unintended interactions. (Section 4)
3) identifying *unexplored unintended interactions* for future research, using our guideline to *conjecture two such interactions*, and *empirically validating* them. (Section 5)

2. Background

We introduce common ML concepts (Section 2.1), different risks to ML (Section 2.2) and corresponding defenses proposed in the literature (Section 2.3).

2.1. ML Classifiers

Let X be the space of input attributes, Y, the corresponding labels, and $P(X,Y)$ the underlying probability distribution of all data points in the universe $X \times Y$. A training dataset $\mathcal{D}_{tr} = \{x_i, y_i\}_{i=1}^{n}$ is sampled from the underlying distribution, i.e., $\mathcal{D}_{tr} \sim P(X,Y)$. The i^{th} data record in $\mathcal{D}_{tr}$ contains input attributes $x_i \in X$ and the classification labels $y_i \in Y$. An ML classification model f_θ is a parameterized function mapping: $f_\theta : X \rightarrow Y$ where θ represents f_θ's parameters. f_θ's capacity is indicated by the number of parameters which varies with different hyperparameters such as the number, size, and types of different layers.

Training. For each data record $(x,y) \in \mathcal{D}_{tr}$, we compute the loss $l(f_\theta(x), y)$ using the difference in model's prediction $f_\theta(x)$ and the true label y. θ is then iteratively updated to minimize the cost function $\mathcal{C}$ defined as: $\min_\theta L(\theta) + \lambda R(\theta)$ where $\mathrm{L}(\theta) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{tr}}[l(f_\theta(x), y)]$. Here, $R(\theta)$ is the regularization function which is used to restrict the range of θ to reduce overfitting to $\mathcal{D}_{tr}$. λ is the regularization hyperparameter which controls the extent of regularization. The loss function is minimized using gradient descent algorithm (e.g, Adam) where θ is updated as $\theta := \theta - \frac{\partial \mathcal{C}}{\partial \theta}$ where $\frac{\partial \mathcal{C}}{\partial \theta}$ is the gradient of the cost with

DOI 10.1109/SP54263.2024.00182

respect to θ. Training stops when algorithm converges to one of $\mathcal{C}$'s local minima.

Inference. Once f_θ is trained, clients can query f_θ with their inputs x and obtain corresponding predictions $f_\theta(x)$ (blackbox setting as in ML as a service) or intermediate layer activation $f^k_\theta(x)$ and θ (whitebox setting as in federated learning). We refer to $f_\theta(x)$, $f^k_\theta(x)$, $\frac{\partial \mathcal{C}}{\partial \theta}$ corresponding to an input, and θ as *observables*. We evaluate the performance of f_θ using accuracy on an unseen test dataset ($\mathcal{D}_{te}$).

2.2. Risks to ML

We identify three categories of risk:

- **Security** includes *evasion*, *poisoning*, and *theft*:

R1 Evasion forces f_θ to misclassify on specific inputs with carefully crafted perturbations [62], [106]. Given an input x, an adversary $\mathcal{A}dv$ generates an adversarial example x_{adv} by adding a perturbation δ (bounded by ϵ_{rob}), i.e., $x_{adv} = x+\delta$ which maximizes f_θ's loss: $\max_{\|\delta\| \leq \epsilon_{rob}} \ell(f_\theta(x+\delta), y)$.

R2 Poisoning modifies f_θ's decision boundary by adding malicious data records to $\mathcal{D}_{tr}$ (referred as *poisons*). This is done to force f_θ to misclassify specific inputs, or make f_θ's accuracy worse. Backdoor attacks [64] are a specific case when $\mathcal{A}dv$ trains f_θ with some inputs with a "trigger" corresponding to an incorrect label. f_θ maps any input with the trigger to the incorrect label but behaves normally on all data records.

R3 Unauthorized model ownership is when $\mathcal{A}dv$ obtains a derived model (f^{der}_θ) mimicking f_θ's functionality. It may be an identical copy of f_θ (*whitebox model theft*), or derived by querying f_θ and using the outputs to train f^{der}_θ (*blackbox model extraction)*) [37], [123], [164]. There are two types of $\mathcal{A}dv$: (a) *malicious suspects* perform model theft but try to evade ownership verification by modifying f^{der}_θ to look different from f_θ, and (b) *malicious accusers* falsely accuse an independently trained model as being f^{der}_θ [99].

- **Privacy** covers *inference attacks* and *data reconstruction*:

P1 Membership inference exploits the difference in f_θ's behaviour on data records in inside and outside $\mathcal{D}_{tr}$ to infer the membership status of a specific data record [25], [191].

P2 Data reconstruction recovers entire data records in $\mathcal{D}_{tr}$ using observables [59], [187], [193], [207], [211]. Furthermore, generative models can be attacked to directly output memorized data records in $\mathcal{D}_{tr}$ [23], [24], [44].

P3 Attribute inference infers values of specific input attributes by exploiting distinguishability in observables. Specifically, observables are distinguishable for different values of the input attribute. This attack violates attribute privacy when the inferred attribute is sensitive [56], [81], [112], [113], [150], [192].

P4 Distribution inference (aka property inference) infers properties of the distribution from which $\mathcal{D}_{tr}$ was sampled [9], [58], [117], [156], [204]. $\mathcal{A}dv$ exploits the difference in model behaviour when trained on datasets with different properties.

- **Fairness** includes f_θ's discriminatory behaviour for different sensitive subgroups.

F Discriminatory behaviour is observed when f_θ's behaviour varies for different sensitive subgroups (e.g., ethnicity or gender) with respect to different metrics such as accuracy, false positive/negative rates. This arises from bias in $\mathcal{D}_{tr}$ or the algorithm [115], [129]. Identifying such biases is challenging as f_θ's behaviour is *incomprehensible*, i.e., difficult to explain why f_θ made a particular prediction.

2.3. Defenses for ML

We identify defenses for the same categories, beginning with security (**RD1**, **RD2**, **RD3**, and **RD4**).

RD1 Adversarial training safeguards against **R1** by training f_θ with adversarial examples [96], [106]. For empirical robustness without any theoretical guarantees, we minimize the maximum loss from the worst case adversarial examples: $\min_\theta \frac{1}{|D|} \sum_{x,y \in D} \max_{\|\delta\| \leq \epsilon_{rob}} \ell(f_\theta(x+\delta), y)$ [107], [201]. Certified robustness gives an upper bound on the loss for adversarial examples within a perturbation budget [93], [96]. f^{rob}_θ is the adversarially trained, robust, variant of f_θ.

RD2 Outlier removal safeguards against **R2** by treating poisoning examples and backdoors as being closer to outliers of $\mathcal{D}_{tr}$'s distribution [19]. While these are empirical defences against **R2**, certified robustness provides an upper bound on the prediction loss for poisoned data records in $\mathcal{D}_{tr}$ [177], [182]. We refer to the model obtained after retraining after outlier removal as f^{outrem}_θ.

RD3 Watermarking safeguards against **R3** by embedding out-of-distribution data (e.g., backdoors), referred to as "watermarks" in $\mathcal{D}_{tr}$ [3], [10], [137], [159]) which can be extracted later to demonstrate ownership. f^{wm}_θ is the watermarked variant of f_θ.

RD4 Fingerprinting safeguards against **R3** by comparing the fingerprints of f_θ and f^{der}_θ to detect if f^{der}_θ was derived from f_θ. Unlike watermarking which requires retraining a model with watermarks, fingerprints are extracted from an already trained f_θ [22], [102], [128], [173], [210]. Fingerprints can be generated from model characteristics (e.g., observables [173], [210]), transferable adversarial examples [22], [102], [128] or identifying overlapping $\mathcal{D}_{tr}$ [111].

For privacy (**PD1** and **PD2**):

PD1 Differential privacy [2] (DP) says that given two neighboring training datasets $\mathcal{D}_{tr}$ and $\mathcal{D}'_{tr}$ differing by one record, a mechanism $\mathcal{M}$ preserves $(\epsilon_{dp}, \delta_{dp})$-DP for all $O \subseteq Range(\mathcal{M})$ if $Pr[\mathcal{M}(\mathcal{D}_{tr}) \in O] \leq Pr[\mathcal{M}(\mathcal{D}'_{tr}) \in O] \times e^{\epsilon}_{dp} + \delta$ where ϵ_{dp} is the privacy budget and δ_{dp} is probability mass of events where the privacy loss is larger than e^{ϵ}_{dp}. For ML models, we add carefully computed noise to gradients during stochastic gradient descent updates, referred to as DPSGD [2]. This reduces the influence of any single data record in $\mathcal{D}_{tr}$ and thereby mitigates **P1** and **P2** [190]. We refer to differentially private variant of f_θ obtained from DPSGD as f^{dp}_θ.

PD2 Attribute privacy and PD3 distribution privacy have been proposed as formal frameworks to preserve privacy in the context of database queries [34], [205]. Adapting this framework to ML is an open problem [155], [156]. Hence, we do not include these defenses in our discussions.

Finally, for fairness (**FD1** and **FD2**):

FD1 Group fairness ensures that the behaviour of f_θ is equitable across different subgroups. Formally, given the set of inputs $\mathcal{X}$, sensitive attributes $\mathcal{S}$ and true labels $\mathcal{Y}$, we can train f_θ to satisfy different fairness metrics (e.g., demographic parity, equality of odds and equality of opportunity [67]) such that $f_\theta(X)$ is statistically independent of S. Group fairness can be achieved by pre-processing $\mathcal{D}_{tr}$, in-processing by modifying training objective function or post-processing $f_\theta(X)$ [115], [129]. We refer to fair variant of f_θ obtained from group fairness constraints as f_θ^{fair}.

FD2 Explanations increase transparency in f_θ's computation by releasing additional information along with $f_\theta(x)$. These help identify biases in f_θ by checking if relevant attributes are being memorized [90], [139]. None of the explanations require retraining. We present three types of explanations used in our paper:

- Attribution-based: For an input x, explanations $\phi(x)$ are a vector $x = (x_1, \cdots x_n)$ indicating the influence of each of the n attributes to $f_\theta(x)$.
- Influence-based: For an input x, explanations $\phi(x)$ are data records in $\mathcal{D}_{tr}$ which resulted in $f_\theta(x)$ [91], [132].
- Recourse-based (counterfactual): For an input x, these explanations output a data record x_c which indicates the minimal change from x to get the favourable outcome.

We summarize the notations in Appendix A (Table 9).

3. Framework

We introduce a framework to examine how defenses interact with various risks. Hence, we conjecture that when a defence is effective in f_θ, *overfitting* (Section 3.1) and *memorization* (Section 3.2) are the underlying causes affecting other risks (validated in Section 4). We further explore underlying factors that influence them. We also establish the relationship between overfitting and memorization (Section 3.3).

3.1. Overfitting

ML classifiers are optimized to improve the *accuracy* on $\mathcal{D}_{tr}$ while ensuring that it generalizes to $\mathcal{D}_{te}$, i.e., high accuracy on both $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$. However, failing to generalize to $\mathcal{D}_{te}$ despite high accuracy on $\mathcal{D}_{tr}$ is *overfitting*.

Measuring Overfitting. We use *generalization error* ($\mathcal{G}_{err}$) to measure overfitting which is the gap between the expected loss on $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$ [68]: $\mathcal{G}_{err} = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{te}}[l(f_\theta(x), y)] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{tr}}[l(f_\theta(x), y)]$. Since expected loss is inversely proportional to model accuracy, henceforth, we rewrite $\mathcal{G}_{err}$ as the difference in accuracy between $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$.

Factors Influencing Overfitting. We present two factors which influence overfitting: i) size of $\mathcal{D}_{tr}$, i.e., $|\mathcal{D}_{tr}|$ (D1), ii) f_θ's capacity (M1) [15], [41], [60], [189]. They can be understood from the perspective of bias and variance: bias is an error from poor hyperparameter choices for f_θ and variance is an error resulting from sensitivity to changes in $\mathcal{D}_{tr}$ [60]. High bias can prevent f_θ from adequately learning relevant relationships between attributes and labels, while high variance leads to f_θ fitting noise in $\mathcal{D}_{tr}$.

D1 **Size of $\mathcal{D}_{tr}$ ($|\mathcal{D}_{tr}|$).** Increasing $|\mathcal{D}_{tr}|$ reduces overfitting: during inference, the likelihood that f_θ has encountered a similar data record in $\mathcal{D}_{tr}$ is higher. Lower variance results in better generalization.

M1 **f_θ's capacity.** Models with large capacity (large number of parameters) can perfectly fit $\mathcal{D}_{tr}$ [8]. Smaller models have higher bias and cannot capture all patterns in $\mathcal{D}_{tr}$.

3.2. Memorization

Memorization is measured with respect to a *single data record* in $\mathcal{D}_{tr}$. We categorize memorization into:

1) **Data Record Memorization.** For classifiers, f_θ tends to memorize entire data records and their corresponding labels, especially for outliers, mislabeled records, or those belonging to long-tail of $\mathcal{D}_{tr}$'s distribution [8], [50], [198]. This *label-only memorization* is essential for f_θ to achieve high accuracy on $\mathcal{D}_{te}$ [20]. Furthermore, memorized data records exhibit a higher influence on observables.

For generative models, which have low overfitting, memorization of entire data records results in their verbatim replication as seen in language and text-to-image diffusion models [23], [24], [28], [44], [44]. This is influenced by the presence of duplicate data records in $\mathcal{D}_{tr}$ and generative model's large capacity [23], [28], [87]. Several metrics such as *exposure*, *k-Eidetic memorization*, and *counterfactual memorization* have been used for generative models.

2) **Attribute Memorization.** f_θ can selectively memorize specific attributes from $\mathcal{D}_{tr}$, which is essential for accurate predictions [108], [163]. There are two types of attributes which can be memorized by f_θ: *stable attributes*, which remain consistent despite changes in data distributions or *spurious attributes* which change f_θ's prediction with change in distribution. f_θ's generalization is better when it prioritizes learning stable attributes [163].

Measuring Memorization. As we focus on classifiers, we use *label-only memorization* of a data record $z_i = (x_i, y_i) \in S$, where $S \subseteq \mathcal{D}_{tr}$, as the difference in probability $\mathcal{P}(f_\theta(x_i) = y_i)$ with and without z_i in $\mathcal{D}_{tr}$ [27], [50], [199]. It is given by $\texttt{mem}(\mathcal{A}, S, i) := \mathcal{P}_{f_\theta \leftarrow \mathcal{A}(S)}[f_\theta(x_i) = y_i] - \mathcal{P}_{f_\theta \leftarrow \mathcal{A}(S\setminus z_i)}[f_\theta(x_i) = y_i]$. If mem is close to one, it is likely that f_θ has memorized z_i.

Factors Influencing Memorization. Different characteristics of a) $\mathcal{D}_{tr}$, b) training algorithm, c) f_θ, influence memorization of data records. For $\mathcal{D}_{tr}$, we consider i) tail length of $\mathcal{D}_{tr}$'s distribution (D2), ii) number of input attributes (D3), iii) f_θ's priority of learning stable attributes (D4).

D2 **Tail length of $\mathcal{D}_{tr}$'s distribution.** Generally, an increase in the tail length, indicated by having more classes with fewer data records than the head of the distribution. This corresponds to higher memorization of atypical data records from the tail classes [165].

D3 **Number of input attributes**. The number of input attributes inversely correlates with the extent of memorization of individual attributes by f_θ.

D4 **Priority of learning stable attributes.** When f_θ prioritizes learning stable attributes over spurious attributes, it generalizes better, making it invariant to changes in other attributes, dataset, or distribution [69], [108], [163]. On the contrary, prioritizing spurious attributes increases f_θ's memorization. Measuring stable attributes is challenging as there are no ground truths. Prior works have used *mean rank metric* to measure the distance between any two inputs over their learnt representation by f_θ [108], [109]. Its value is small if some attributes between two inputs are the same which corresponds to stable attributes. Hence, a low mean rank means f_θ prioritized learning stable attributes.

For characteristics of training algorithms, we identify i) *curvature smoothness* (O1), ii) *distinguishability of observables (between data records, subgroups, datasets, and models)* (O2), iii) *distance of $\mathcal{D}_{tr}$ data records to decision boundary* (O3) as potential factors underlying memorization.

O1 **Curvature smoothness.** The convergence of the loss on different data records towards a minima depends on the smoothness of the objective function. Since different data records have varying levels of difficulty in terms of learning, a less smooth curvature can lead to local minima around difficult records [26], [54], [165]. Consequently, these data records demonstrate distinct memorization patterns, evident by different influence on the observables.

O2 **Distinguishability of observables.** Due to the varying levels of difficulty in learning different records, and thus variable memorization, f_θ exhibits distinguishability in observables. We identify three subtypes based on distinguishability in observables between O2.1 datasets, O2.2 subgroups, O2.3 models.

O3 **Distance of $\mathcal{D}_{tr}$ data records to decision boundary.** Due to the varying difficulty to learn different data records, they demonstrate variable memorization [36], [166], [167]. Memorized data records are closer to the decision boundary [54].

For model characteristics, we consider *capacity*. This is the same as M1. Increasing f_θ's capacity generally increases memorization by fitting a more complex decision boundary [24], [28], [94], [151]. While larger model sizes may reduce overfitting when the model is too simple to effectively generalize to the dataset, over-parameterized models generalize well. Hence, $\mathcal{G}_{err}$ does not change and an increase in model size correlates with higher memorization.

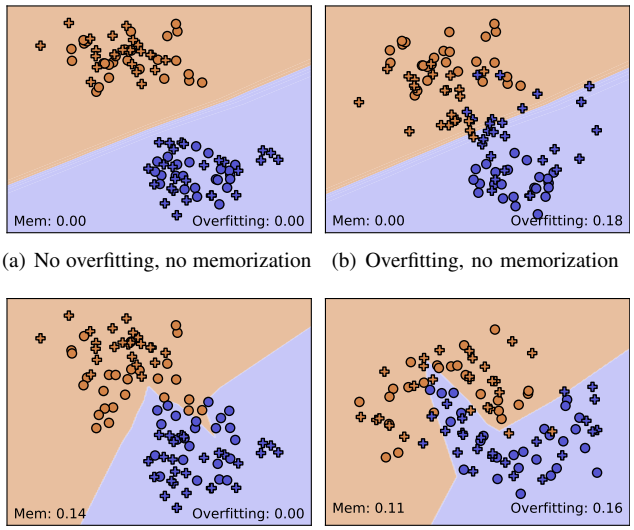
We summarize the different factors in Table 1.

3.3. Overfitting and Memorization

Memorization and overfitting have been shown as distinct phenomenon which can occur simultaneously as seen for generative models [23], [44]. We present four cases based on the presence or absence of overfitting and memorization in a model to illustrate the relation between them (Figure 1). We use a synthetic dataset for binary classification and train a simple neural network with two hidden layers with ten neurons each. We measure overfitting using $\mathcal{G}_{err}$ and memorization using average mem across all $\mathcal{D}_{tr}$ data records.

TABLE 1. SUMMARY AND DESCRIPTION OF DIFFERENT FACTORS.

Factor	Description
D1	Size of $\mathcal{D}_{tr}$
D2	Tail length of $\mathcal{D}_{tr}$'s distribution
D3	Number of input attributes
D4	Priority of learning stable attributes
O1	Curvature smoothness of the objective function
O2 (O2.1)	Distinguishability of observables between datasets
O2 (O2.2)	Distinguishability of observables across subgroups
O2 (O2.3)	Distinguishability of observables across models
O3	Distance of $\mathcal{D}_{tr}$ data records to the decision boundary
M1	f_θ's capacity



(a) No overfitting, no memorization (b) Overfitting, no memorization

(c) Memorization, no overfitting (d) Overfitting, memorization

Figure 1. Relationship between overfitting and memorization: Average mem across all data records in $\mathcal{D}_{tr}$ and overfitting ($\mathcal{G}_{err}$) are at the bottom. Circles indicate $\mathcal{D}_{tr}$ data records, and crosses indicate $\mathcal{D}_{te}$ data records.

Figure 1(a) is when data records in $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$ are linearly separable. Hence, f_θ perfectly fits $\mathcal{D}_{tr}$ and generalizes to $\mathcal{D}_{te}$ ($\mathcal{G}_{err} = 0$). Further, data records in $\mathcal{D}_{tr}$ are far from the decision boundary (mem=0). Figure 1(b) is when $\mathcal{D}_{tr}$ is linearly separable but $\mathcal{D}_{te}$ is not ($\mathcal{G}_{err} > 0$). However, data records in $\mathcal{D}_{tr}$ are far from the decision boundary (mem=0). Figure 1(c) is when data records in $\mathcal{D}_{tr}$ and $\mathcal{D}_{te}$ are reasonably separable resulting in $\mathcal{G}_{err} = 0$. However, with $\mathcal{D}_{tr}$ data records being close to the decision boundary, we observe a high mem. Figure 1(d) is where f_θ both overfits to $\mathcal{D}_{tr}$, i.e., f_θ fits $\mathcal{D}_{tr}$ perfectly but does not generalize to $\mathcal{D}_{te}$, and memorizes some data records in $\mathcal{D}_{tr}$, i.e., they are not leave-one-out stable. This is generally seen in real-world applications, where $\mathcal{D}_{tr}$ and f_θ are complex. Hence, hereafter we focus only on this setting.

4. Understanding Unintended Interactions

We now show how factors influencing overfitting and memorization relate to defenses (Section 4.1) and risks

(Section 4.2), summarizing in Table 2. We then survey unintended interactions explored in prior work, and situate them within our framework (Section 4.3). Finally, we suggest a guideline to help conjecture such interactions using our framework (Section 4.4).

4.1. Revisiting Defenses for ML

We explore how the effectiveness of different defenses correlates with different factors. We also indicate whether a defense degrades accuracy (▼) or not (≈).

RD1 (Adversarial Training). f_θ learns spurious attributes which are essential for achieving high accuracy and low $\mathcal{G}_{err}$. However, this makes f_θ susceptible to evasion [76]. f_θ^{rob} is compelled to learn stable attributes [76], [169]. This improves robustness but with higher $\mathcal{G}_{err}$ and an accuracy drop (▼) [76]. More training data can reduce overfitting [169]. Increasing the tail length, lowers robustness for the tail classes [16], [74]. Also, **RD1** includes a regularization term to minimize the maximum loss from adversarial examples which results in a smoother curvature [107]. This creates an explicit trade-off between generalization and robustness [158]. This is accompanied with an increase in the distance of $\mathcal{D}_{tr}$ data records to the decision boundary [184]. Further, f_θ^{rob} exhibits higher distinguishability between predictions on data records inside and outside $\mathcal{D}_{tr}$ [151]. Finally, f_θ^{rob} should have high capacity to learn adversarial examples [107].

RD2 (Outlier Removal). Outliers are identified by measuring the influence of memorized data records on observables [84]. A longer tail increases the number of memorized outliers making their detection easy [174]. Some of these outliers contribute to f_θ's accuracy [20], [50], [51]. Thus, removing such outliers increases overfitting in f_θ^{outrem} which results in an accuracy drop (▼) [47], [82]–[84].

RD3 (Watermarking). Ideally, f_θ^{wm} should maintain $\mathcal{G}_{err}$ similar to f_θ. In practice, f_θ^{wm} incurs an accuracy drop (▼) [3], [103], [159]. A long tail enables stealthier watermarks from tail classes [101]. Lower distinguishability in observables for watermarks between f_θ and f_θ^{der}, but distinct from independently trained models, helps identifying f_θ^{der} [3]. High model capacity aids memorizing the watermarks [3].

RD4 (Fingerprinting). There is no explicit impact on overfitting, memorization, or accuracy (≈) as **RD4** does not require retraining. A lower distinguishability in fingerprints between f_θ and f_θ^{der}, but distinct from independently trained models, is preferred for effectively identifying f_θ^{der} [102], [173]. The effectiveness of fingerprints relies on f_θ's overfitting and memorization. For instance, dataset inference [111] hinges on the distinguishability of observables between data records inside and outside $\mathcal{D}_{tr}$, for both f_θ and f_θ^{der}. This helps identify if they share overlapping $\mathcal{D}_{tr}$. Similarly, fingerprints based on adversarial examples rely on f_θ's generalization for transferring adversarial examples to f_θ^{der} [22], [102], [111], [128], [175]. For effective transfer, a lower distance to the boundary is preferred [22], [102], [128].

PD1 (Differential Privacy). Memorization of data records is reduced by lowering their influence on observables with **PD1**. f_θ^{dp} is compelled to learn stable attributes as evident by better privacy guarantees in causal models [163]. Also, adding noise to gradients regularizes f_θ^{dp} which curbs overfitting but with an accuracy drop (▼) [2], [80], [135]. This reduces distinguishability in observables for data records inside and outside $\mathcal{D}_{tr}$ [2], and decreases the distance of $\mathcal{D}_{tr}$ data records to the boundary [170]. Also, **PD1** reduces the curvature smoothness [18], [170], and lower privacy for tail classes [11], [157]. Low capacity leads to reasonable privacy, without significant accuracy drop [142].

FD1 (Group Fairness). For fairness via in-processing algorithms, the regularization term explicitly creates a trade-off between fairness and accuracy [172], [195], exacerbating overfitting with a drop in f_θ^{fair}'s accuracy (▼) [130], [136], [172], [195]. This reduces distinguishability in observables across subgroups [1], [197]. Further, f_θ^{fair} prioritizes learning stable attributes which do not overlap with sensitive attributes [1], [119]. For equitable behaviour, f_θ^{fair} memorizes some data records in minority subgroups, thereby decreasing their distance to the boundary [30], [168].

FD2 (Explanations). There is no explicit impact on overfitting, memorization, or accuracy (≈) as **FD2** does not require retraining. More attributes increase the dimensions of the explanations, and reduce the influence of individual attributes which affects the quality of **FD2** [144]. Also, **FD2** exhibits distinguishability across subgroups [46]. The quality of explanations is tied to f_θ's overfitting: they exhibit better approximation for data records in $\mathcal{D}_{tr}$ compared to $\mathcal{D}_{te}$ [144]. Therefore, **FD2** exhibits higher distinguishability between data records inside and outside $\mathcal{D}_{tr}$ [144].

4.2. Revisiting Risks in ML

We now discuss different factors and how it correlates with the susceptibility on different risks.

R1 (Evasion). Longer tails increase **R1**: adversarial examples from tail classes are more effective [95], [181]. Lower curvature smoothness makes it easier to generate adversarial examples [107]. Further, data records close to f_θ's decision boundary require smaller perturbations to induce misclassification [33], [88]. Hence, **R1** risk decreases for data records farther away from the boundary [170]. When overfitting and memorization are pronounced, the decision boundary is more complex and bigger, making **R1** easier [33].

R2 (Poisoning). Longer tail increases **R2** as the poisons from tail classes are more discreet [17], [101], [126]. Following watermarking literature which use poisons, the effectiveness of poisons is higher with larger f_θ capacity due to their better memorization [3].

R3 (Unauthorized Model Ownership). Whitebox model theft does not explicitly change overfitting or memorization. Henceforth, we focus only on blackbox model extraction. **R3** is less successful if f_θ overfits or memorizes: accuracy of f_θ is lower and consequently, labels from f_θ for training f_θ^{der} are of a bad quality. Hence, f_θ^{der} has low

TABLE 2. SUMMARY OF CORRELATION BETWEEN EFFECTIVENESS OF A DEFENSE <d> AND SUSCEPTIBILITY TO A RISK <r> WITH A FACTOR <f>: ↑ INDICATES <d> OR <r> POSITIVELY CORRELATES WITH <f>; ↓ INDICATES A NEGATIVE CORRELATION.

Defences (<↑ or ↓>, <f>)	Risks (<↑ or ↓>, <f>)
RD1 (Adversarial Training): • D1 ↑, $\|\mathcal{D}_{tr}\|$ [169] • D2 ↓, tail length [16], [74] • D4 ↑, priority for learning stable attributes [169] • O1 ↑, curvature smoothness [107] • O2.1 ↑, distinguishability in data records inside and outside $\mathcal{D}_{tr}$ [151] • O3 ↑, distance to boundary for most $\mathcal{D}_{tr}$ data records [184] • M1 ↑, model capacity [107] **RD2 (Outlier Removal):** • D2 ↑, tail length [174] **RD3 (Watermarking):** • D2 ↑, tail length [101] • O2.3 ↓, distinguishability in observables for watermarks between f_θ and f_θ^{der}, but distinct from independent models [3] • M1 ↑, model capacity [3] **RD4 (Fingerprinting):** • O2.3 ↓, distinguishability in observables for fingerprints between f_θ and f_θ^{der}, but distinct from independent models [102], [173] • O3 ↓, distance of $\mathcal{D}_{tr}$ data records to boundary [22], [102], [128] **PD1 (Differential privacy):** • D2 ↓, tail length [11], [157] • D4 ↑, priority for learning stable attributes [163] • O1 ↓, curvature smoothness [18], [170] • O2.1 ↓, distinguishability for data records inside and outside $\mathcal{D}_{tr}$ [2] • O3 ↓, distance of $\mathcal{D}_{tr}$ data records to decision boundary [170] • M1 ↓, model capacity [142] **FD1 (Group Fairness):** • D4 ↑, priority for learning stable attributes [1], [57], [119] • O2.2 ↓, distinguishability in observables across subgroups [1], [197] • O3 ↓, distance to decision boundary for most $\mathcal{D}_{tr}$ data records [168] **FD2 (Explanations):** • D3 ↓, number of input attributes [144] • O2.1 ↑, distinguishability in data records inside and outside $\mathcal{D}_{tr}$ [144] • O2.2 ↑, distinguishability in observables across subgroups [46]	**R1 (Evasion):** • D2 ↑, tail length [95], [181] • O1 ↓, curvature smoothness [107] • O3 ↓, distance of $\mathcal{D}_{tr}$ data records to boundary [170] **R2 (Poisoning):** • D2 ↑, tail length [17], [101], [126] • M1 ↑, model capacity [3] **R3 (Unauthorized Model Ownership):** • M1 ↓, model capacity [92], [123] **P1 (Membership Inference):** • D1 ↓, $\|\mathcal{D}_{tr}\|$ [143], [192] • D2 ↑, tail length [25], [26] • D4 ↓, priority for learning stable attributes [108], [163] • O2.1 ↑, distinguishability for data records inside and outside $\mathcal{D}_{tr}$ [143] • O3 ↓, distance to decision boundary [144] • M1 ↑, model capacity [48], [151] **P2 (Data Reconstruction):** • D2 ↑, tail length [179] • D3 ↓, number of input attributes [55] • O2.1 ↑, distinguishability for data records inside and outside $\mathcal{D}_{tr}$ [120], [207] • O2.2 ↑, distinguishability in observables across subgroups [188] **P3 (Attribute Inference):** • D2 ↑, tail length [81] • D4 ↓, priority for learning stable attributes [108], [150] • O2.2 ↑, distinguishability in observables across subgroups [1] **P4 (Distribution Inference):** • D2 ↑, tail length [31], [110] • D4 ↑, priority for learning stable attributes [155] • O2.1 ↑, distinguishability in observables between datasets [155], [156] • M1 ↑, model capacity [31] **F (Discriminatory behaviour):** • D2 ↑, tail length [115] • O2.2 ↑, distinguishability in observables across subgroups [197]

accuracy [100]. Also, for a given f_θ^{der}, increasing f_θ's capacity will decrease **R3** [92], [123].

P1 (Membership Inference). Overfitting is correlated with an increase in **P1**. Hence, **P1** decreases with larger $|\mathcal{D}_{tr}|$ as it increases generalization [143], [192]. Further, when f_θ prioritizes learning stable attributes, **P1** decreases due to better generalization [108], [163]. Also, a higher distinguishability between data records inside and outside $\mathcal{D}_{tr}$, increases **P1** [73], [143], [192]. Memorized data records exhibit a higher influence on observables, rendering them more susceptible to **P1** [25], [47]. Hence, longer tail increases memorized outliers thereby increasing their susceptibility [25], [26]. Also, data records located farther from the decision boundary, with low memorization, are less susceptible to **P1** [144]. Finally, increasing f_θ's capacity amplifies the memorization of $\mathcal{D}_{tr}$, which increases **P1** [48], [151].

P2 (Data Reconstruction). Overfitting improves reconstruction of data records in $\mathcal{D}_{tr}$ [120]. Higher distinguishability between records inside and outside $\mathcal{D}_{tr}$, increases **P2** [207]. Memorization amplifies the influence of data records on observables, leading to better reconstruction [149], [179], [180], [208]. **P2** is higher for outliers from tail classes [179]. Further, more input attributes reduces accurate reconstruction of inputs [55]. Finally, higher distinguishability in observables across subgroups increases **P2** [188].

P3 (Attribute Inference). Memorization of sensitive attributes is likely to increase **P3** [165]. Hence, longer tail increases memorized outliers, increasing their susceptibility [81]. Prioritizing learning stable attributes decreases **P3** [108], [150]. Higher distinguishability across subgroups increases **P3** [1]. Notably, overfitting has no discernible impact on this risk [100].

P4 (Distribution Inference). Empirical evidence suggests that overfitting reduces **P4**, although the reason remains unexplained [31], [155]. Tail length correlates with **P4** [31], [110]. Prioritizing stable attributes amplifies the risk because stable attributes are the same as the distributional property being inferred [155]. Higher distinguishability in observables across datasets with different distributional properties, increases **P4** [155], [156]. Finally, increasing f_θ's capacity, without overfitting, increases **P4** [31].

F (Discriminatory Behaviour). Overfitting makes bias prominent as evident by the accuracy disparity between minority and majority subgroups [185]. Increased memorization of minority subgroups contributes to their discrimination [30]. Longer tails increase memorized outliers, worsening **F** [115]. Increase in distinguishability of observables across subgroups correlates with increase in **F** [1], [197].

4.3. Surveying Unintended Interactions

We now survey various unintended interactions reported in the literature and situate them within our framework. We selected papers on various interactions using Google Scholar, focusing on those published in top-tier venues. Next, we examined their citations and related work to uncover additional works. We use existing surveys (e.g., Gittens et al. [61]) to confirm that our coverage is reasonable.

We present the different interactions explored in prior works in Table 3. We mark the type of interaction (indicated as "**I**") between <d> and <r> as ● if "an increase (decrease) in effectiveness of <d> correlates with an increase (decrease) in <r>" and ● if "an increase (decrease) in effectiveness of <d> correlates with a decrease (increase) in <r>". When **I** is ●, prior work has not studied this particular interaction and the middle columns are also left empty. When **I** is ● or ●, empty middle columns correspond to cases where prior work merely identifies the type of interaction without evaluating the influence of the factors.

We use the different factors from Table 1 to gain a more nuanced understanding of the interactions. We mark a factor with ● if there are empirical results regarding its influence, ⊙ if there are theoretical results and ○ if it is only conjectured. Exceptions to any interactions are listed under "References" and discussed in the text.

RD1 (Adversarial Training)

● **R1 (Evasion)** is less effective with **RD1** which is specifically designed to resist adversarial examples [96], [107], [201]. Using the min-max objective function increases the curvature smoothness which reduces the possibility of generating adversarial examples [107]. Further, **RD1** pushes the decision boundary away from the data records [201]. Also, f_θ^{rob} requires a high capacity to learn a complex decision boundary to fit adversarial examples [107]. Longer tail increases the effectiveness of adversarial examples from the tail classes despite using **RD1** [95], [181].

● **R2 (Poisoning)** can be effective with **RD1** [178]. Recall that f_θ^{rob} is forced to learn stable attributes. Hence, it might seem that poisons, generated by manipulating spurious attributes, does not affect f_θ^{rob}'s predictions [160]. However, Wen et al. [178] design poisons to degrade f_θ^{rob}'s accuracy by manipulating the stable attributes.

● **R3 (Unauthorized model ownership)** can be more prevalent as model theft of f_θ^{rob} is easier than f_θ [89]. However, no reasons were explored. We conjecture that since f_θ^{rob} has uniform predictions on neighboring data records [85], [96], [107], [164], hence, f_θ^{der} requires fewer queries compared to f_θ. The above interaction assumes $\mathcal{A}dv$ is a malicious suspect. For a malicious accuser, **RD1** can mitigate false claims [99], which decreases **R3** (●).

● **P1 (Membership inference)** is easier with **RD1** [70], [151]. Recall that **RD1** modifies f_θ^{rob}'s decision boundary to memorize adversarial examples in $\mathcal{D}_{tr}$ [151]. This increases the influence of some records on observables. Consequently, the distinguishability between data records inside and outside $\mathcal{D}_{tr}$ is higher [151]. Hence, **P1** increases. Hayes et al. [70] provably show that increasing $|\mathcal{D}_{tr}|$ lowers overfitting and hence, reduces **P1**. Further, **P1** increases with f_θ^{rob}'s capacity due to higher memorization.

● **P2 (Data reconstruction)** is easier with **RD1** [116], [203]: f_θ^{rob} has interpretable gradients [134], which enhances the quality of reconstruction. Further, consistent attributes across inputs, are easier to reconstruct.Finally, lower f_θ^{rob}'s capacity increases **P2** [203].

● **P4 (Distribution inference)** is easier as f_θ^{rob} tends to learn stable attributes which matches with distributional properties being inferred [155]. However, increasing ϵ_{rob} reduces **P3** as f_θ^{rob}'s accuracy decreases.

● **F (Discriminatory behaviour)** increases with **RD1** as evident from the accuracy disparity of f_θ^{rob} across different classes [16], [38], [74]. Longer tail contributes to the tail classes having lower robustness [16], [74]. Also, increasing ϵ_{rob}, increases the class-wise disparity [104]. Empirically, $|\mathcal{D}_{tr}|$, f_θ^{rob}'s capacity, type of $\mathcal{D}_{tr}$ (e.g., images, tabular), and f_θ^{rob}'s architecture and optimization algorithms *do not* influence the interaction [16], [74].

RD2 (Outlier Removal)

● **R1 (Evasion)** is less effective with **RD2** [63]. Adversarial examples are outliers compared to $\mathcal{D}_{tr}$'s distribution which can detected before passing them as inputs [63].

● **R2 (Poisoning)** is less effective with **RD2** [126], [162]. Poisons have a different distribution from $\mathcal{D}_{tr}$ and can be detected and removed as outliers during pre-processing [126]. During training, the influence of outliers, often correspond to poisons, on observables can be minimized [17].

● **P1 (Membership inference)** is arguably worse with **RD2** followed by retraining [26]. Memorized outliers are more susceptible to membership inference. However, after removing these outliers, f_θ^{outrem} memorizes other data records making them more susceptible than before [26]. Varying $|\mathcal{D}_{tr}|$, f_θ^{outrem}'s capacity, or the number of duplicates in $\mathcal{D}_{tr}$ does *not* have any impact on the risk [26].

● **P3 (Attribute inference)** is more effective as retraining after removal of memorized outliers constituting the long-tail of $\mathcal{D}_{tr}$, makes new data records susceptible [81].

● **F (Discriminatory behaviour)** increases as outliers could constitute a statistical minority. This exacerbates bias against minorities by excessively identifying them as outliers [141]. On increasing $|\mathcal{D}_{tr}|$, more minority data records are classified as outliers further exacerbating the bias [141].

RD3 (Watermarking)

● **R2 (Poisoning)** is used by most watermarking schemes thereby requiring f_θ to be susceptible to it [3], [97], [140], [202]. These schemes inject data records closer to outliers for memorization by f_θ. Hence, longer tails can help generate effective and discreet poisons as watermarks.

TABLE 3. UNINTENDED INTERACTIONS BETWEEN DEFENSES WITH DIFFERENT RISKS. **TYPE OF INTERACTION (I)**: 🔴→ RISK INCREASES; 🟢→ RISK DECREASES; ⚫→ UNEXPLORED. WE INDICATE EXCEPTIONS TO THESE UNDER "REFERENCES" AND DISCUSS THEM IN THE TEXT. **UNDERLYING FACTORS:** ● → EMPIRICAL; ⊙ → THEORETICAL AND ○ → CONJECTURED. UNDER O2, WE WRITE O2.1, O2.2, AND O2.3 AS 1, 2, AND 3 RESPECTIVELY. WE MARK THE FACTORS EVALUATED FOR CONJECTURED INTERACTIONS AS ∗ (SEE SECTION 5).

Defenses	Risks	I	OVFT	Memorization					Both		References
			D1	D2	D3	D4	O1	O2	O3	M1	
RD1 (Adversarial Training)	R1 (Evasion)	🟢		●			●		●	●	[95], [107], [181], [201]
	R2 (Poisoning)	🔴									[160], [178]
	R3 (Unauthorized Model Ownership)	🔴	○								[89] ([100]: 🟢)
	P1 (Membership Inference)	🔴	⊙, ●					1: ●		●	[70], [151]
	P2 (Data Reconstruction)	🔴				○				●	[116], [203]
	P3 (Attribute Inference)	⚫									
	P4 (Distribution Inference)	🔴				○					[155]
	F (Discriminatory Behaviour)	🔴		⊙, ●							[16], [38], [74], [104]
RD2 (Outlier Removal)	R1 (Evasion)	🟢									[63]
	R2 (Poisoning)	🟢									[162]
	R3 (Unauthorized Model Ownership)	⚫									
	P1 (Membership Inference)	🔴		●							[26], [47]
	P2 (Data Reconstruction)	⚫									
	P3 (Attribute Inference)	🔴		●							[81]
	P4 (Distribution Inference)	⚫									
	F (Discriminatory Behaviour)	🔴	●	○							[141]
RD3 (Watermarking)	R1 (Evasion)	⚫									
	R2 (Poisoning)	🔴		○							[3], [97], [140], [202]
	R3 (Unauthorized Model Ownership)	🟢		○				3: ●	●		[3], [103], [159]
	P1 (Membership Inference)	🔴		○				1: ●	●		[35], [165]
	P2 (Data Reconstruction)	🔴		○				1: ●	●		[165]
	P3 (Attribute Inference)	🔴		○				2: ●	●		[165]
	P4 (Distribution Inference)	🔴	⊙, ●	○				1: ●	●	●	[31], [110]
	F (Discriminatory Behaviour)	🔴		○							[101]
RD4 (Fingerprinting)	R1 (Evasion)	🔴						3: ●	●		[22], [102], [128]
	R2 (Poisoning)	⚫									
	R3 (Unauthorized Model Ownership)	🟢		●				3: ●	●		[102], [111], [173], [210]
	P1 (Membership Inference)	🔴							●		[111]
	P2 (Data Reconstruction)	⚫									
	P3 (Attribute Inference)	⚫									
	P4 (Distribution Inference)	⚫									
	F (Discriminatory Behaviour)	⚫									
PD1 (Differential Privacy)	R1 (Evasion)	🔴					●		●		[18], [170] ([21]:🟢)
	R2 (Poisoning)	🔴									[72], [79], [105]
	R3 (Unauthorized Model Ownership)	⚫									
	P1 (Membership Inference)	🟢				⊙, ●		1: ●			[75], [80], [100], [135]
	P2 (Data Reconstruction)	🟢						1: ●			[152], [190]
	P3 (Attribute Inference)	🟢						2: ○			[100]
	P4 (Distribution Inference)	🟢						●			[9], [155] ([155]: 🔴)
	F (Discriminatory Behaviour)	🔴		○					●		[11], [49], [54], [166], [196]
FD1 (Group Fairness)	R1 (Evasion)	🔴					●		⊙, ●		[168]
	R2 (Poisoning)	🔴									[29], [114], [148], [171]
	R3 (Unauthorized Model Ownership)	⚫									
	P1 (Membership Inference)	🔴	●						●	●	[30], [161]
	P2 (Data Reconstruction)	⚫			∗			2: ∗			Our conjecture: 🟢
	P3 (Attribute Inference)	🟢						2:⊙,●			[1] ([53]: 🔴)
	P4 (Distribution Inference)	🔴	●								[155]
	F (Discriminatory Behaviour)	🟢	●					2:●			[4], [86], [131], [197]
FD2 (Explanations)	R1 (Evasion)	🔴									[133], [206]
	R2 (Poisoning)	🔴			●		⊙, ●			●	[14], [71], [146], [200]
	R3 (Unauthorized Model Ownership)	🔴									[5], [133], [176]
	P1 (Membership Inference)	🔴	●		●			1: ●	●	●	[98], [127], [133], [144]
	P2 (Data Reconstruction)	🔴		⊙, ●	●						[144], [209]
	P3 (Attribute Inference)	🔴						2: ○			[46]
	P4 (Distribution Inference)	⚫			∗			1: ∗		∗	Our Conjecture: 🔴
	F (Discriminatory Behaviour)	🔴		●	●					●	[12], [40] ([90], [139]: 🟢)

🟢 **R3 (Unauthorized model ownership)** can be detected with **RD3** [3], [103], [159]. f_θ^{wm} should have enough capacity to memorize watermarks. Therefore, increasing capacity can improve effectiveness of watermarking [103]. Also, longer tails help generate more discreet watermarks.

🔴 **P1 (Membership inference)**, **P2 (Data reconstruction)**, **P3 (Attribute inference)**, **P4 (Distribution inference)** do not have any prior evaluations with watermarks. However, prior work has explored stealthy poisoning where mislabeled data records are added to $\mathcal{D}_{tr}$ to increase different privacy risks without an accuracy drop [31], [35], [110], [165]. We liken this form of poisoning to **RD3** [3], [97], [140], [202], and use their observations to justify different interactions.

We conjecture that adding watermarks from tail classes, similar to poisoning, will increase privacy risks. Hence, a longer tail is likely to increase different privacy risks. Further, memorization of watermarks increases their influence on observables. This increases the distinguishability in observables across datasets, subgroups, and models. Moreover, the decision boundary is more complex, consequently

increasing the susceptibility of outliers closer to the decision boundary [35], [165]. For distribution inference, the risk increases with lower capacity, and higher $|\mathcal{D}_{tr}|$ [31], [110].

● **F (Discriminatory behaviour)** could worsen with watermarking schemes which rely on adding bias to $\mathcal{D}_{tr}$ [101]. Watermarks (with incorrect labels) belonging to the tail classes are more effectively memorized for ownership verification [29], [78], [101]. Hence, increasing the tail length is likely to increase the effectiveness of watermarks.

RD4 (Fingerprinting)

● **R1 (Evasion)** can be used for fingerprinting with adversarial examples, necessitating f_θ's susceptibility to this risk [22], [102], [128]. Hence, when f_θ is robust against evasion, fingerprints are less effective [102]. Such fingerprints require $\mathcal{D}_{tr}$ records to be closer to class boundaries [22]. Further, overfitting and memorization make the decision boundary complex, which helps generate fingerprints. To transfer these fingerprints from f_θ to f_θ^{der}, the distinguishability between their observables should be low but distinct from independent models [22], [102], [128].

● **R3 (Unauthorized model ownership)** can be detected using fingerprints. Fingerprints based on adversarial examples are closer to the outliers in the tail classes of $\mathcal{D}_{tr}$'s distribution. Unique intermediate activations for a given input can also be used as a fingerprint which have low distinguishability between f_θ and f_θ^{der} but distinct from independent models [173].

● **P1 (Membership inference)** can be used to construct fingerprints which relies f_θ being susceptible to it. Dataset inference [111] uses a weak membership inference attack to identify f_θ's decision boundary, that constitute its fingerprint. Specifically, it distinguishes between the distribution of distances of data records in $\mathcal{D}_{tr}$ and any other independent dataset to the decision boundary, and *not* any individual data records.

PD1 (Differential Privacy)

● **R1 (Evasion)** is easier for f_θ^{dp}. DPSGD reduces the distance between $\mathcal{D}_{tr}$ records and decision boundaries making it easier to generate adversarial examples [170]. Also, training with DPSGD results in a less smooth curvature which can can potentially increase **R1** [18], [43], [118], [170]. f_θ^{dp} creates a false sense of robustness via gradient masking, i.e., modifying the gradients to be useless for gradient based evasion (as observed in [170]). However, f_θ^{dp} is still susceptible to non-gradient evasion. Bu and Zhang [21] show that training with DPSGD can result in adversarial robustness depending on the choice of hyperparameters (●). Further exploration is required to understand the conditions under which DPSGD and adversarial robustness align.

● **R2 (Poisoning)** can be more effective against f_θ^{dp}. Prior works have suggested that **R2** decreases as **PD1** reduces their influence on the observables [72], [105]: DPSGD constrains gradient magnitudes which enhances f_θ^{dp}'s robustness against poisons. However, it has been shown that the choice of hyperparameters (e.g., noise multiplier, gradient clipping) are the reason for an improvement in robustness [79]. More investigation is required to understand the root cause of this interaction.

● **P1 (Membership inference)** is less effective on f_θ^{dp} due to **PD1**'s regularizing effect as evident from low $\mathcal{G}_{err}$ [2]. This reduces the distinguishability between data records inside and outside $\mathcal{D}_{tr}$ [2], [80], [100]. Further, **PD1** provably bounds **P1** for i.i.d. data [80], [135]. Prioritizing causal attributes gives a higher privacy guarantee while reducing the risk to **P1** [163]. However, **PD1**'s guarantee does not hold for **P1** on non-i.i.d data [75]. Also, duplicates in $\mathcal{D}_{tr}$ weaken the bound and the achievable ϵ_{dp} is too high for **P1**.

● **P2 (Data reconstruction)** is less effective against f_θ^{dp} as **PD1** reduces the distinguishability between data records inside and outside $\mathcal{D}_{tr}$ [190]. The lower quality of information available to $\mathcal{A}dv$ reduces **P2**.

● **P3 (Attribute inference)** was less effective against f_θ^{dp} [100], but no explanation was presented. We conjecture that **PD1**, by reducing the memorization of sensitive attributes and their influence on observables, reduces the distinguishability across subgroups [84].

● **P4 (Distribution inference)** is less effective assuming $\mathcal{A}dv$ does not know f_θ^{dp}'s training hyperparameters to adapt the attack [155]. **PD1** has a regularizing effect which reduces f_θ^{dp}'s ability to learn $\mathcal{D}_{tr}$'s distribution. This reduces the distinguishability of observables across datasets with different distributional properties thereby reducing **P4** [155]. However, if $\mathcal{A}dv$ is aware of training hyperparameters, they can adapt their attack to increase **P4** (●) [155].

● **F (Discriminatory behaviour)** increases for f_θ^{dp} as evident by accuracy disparity between minority and majority subgroups [11]. This effect varies with f_θ^{dp}'s architecture. Also, training hyperparameters (e.g., clipping and gradient noise) disproportionately affect underrepresented classes [49], [166]. Further, longer tail results in poor privacy protection for the tail classes [157]. These data records have a high gradient norm and are closer to the decision boundary, which exacerbates the disparity [166]. Finally, lowering ϵ_{dp} makes the disparity worse [11].

FD1 (Group Fairness)

● **R1 (Evasion)** is easier against f_θ^{fair}. Data records in $\mathcal{D}_{tr}$ have a smaller distance to the decision boundary making it easier to generate adversarial examples [168]. Robustness, on the other hand, is higher for data records with a larger distance to the decision boundary. Hence, the optimization for **FD1** is contrary to the robustness requirement [168].

● **R2 (Poisoning)** is more effective against f_θ^{fair} [29], [114], [148], [171]. It aims to increase f_θ^{fair}'s disparity, and differs in ways to craft poisons which include: using f_θ^{fair}'s gradients [148], distorting $\mathcal{D}_{tr}$'s distribution [29], skewing the decision boundary [114], and maximizing the covariance between the sensitive attributes and labels [114].

● **P1 (Membership inference)** is more effective against f_θ^{fair} [30], [161]. Memorization of some data records lowers their distance to decision boundary thereby increasing their influence on observables which increases **P1** [30]. Further, **P1** increases with f_θ^{fair}'s capacity and $|\mathcal{D}_{tr}|$ assuming the fraction of minority subgroup remains the same [30]. This is assuming equalized odds as a fairness metric. It is not clear

how the nature of the interaction changes for a different fairness metric (e.g., demographic parity).

● **P3 (Attribute inference)** is less effective. f_θ^{fair} has lower distinguishability in predictions across subgroups [1]. Hence, **FD1** specifically with demographic parity as a constraint, can provably and empirically reduce **P3** [1]. On the contrary, Ferry et al. [53] show that it is possible to infer sensitive attributes from f_θ^{fair} which suggests an increase in **P3** (●). However, this is under a restrictive threat model where $\mathcal{A}dv$ knows the group fairness algorithm and hyperparameters to adapt their attack.

● **P4 (Distribution inference)** is less effective with over-sampling and under-sampling data records. However, the disparity between different subgroups increases [155]. Hence, the resistance to **P4** and **FD1** are in conflict.

● **F (Discriminatory behaviour)** is less for f_θ^{fair}. This can be achieved by *pre-processing* $\mathcal{D}_{tr}$ to remove bias [86]. Also, *in-processing* using regularization reduces the distinguishability of predictions across subgroups to reduce disparity [4], [197]. Finally, *post-processing* can help calibrate the predictions to satisfy some fairness metric [131].

FD2 (Explanations)

● **R1 (Evasion)** is easier as **FD2** uses gradients which captures f_θ's decision boundary. This helps generating adversarial examples, which in addition to forcing a misclassification, produces $\mathcal{A}dv$'s chosen explanation [133], [206].

● **R2 (Poisoning)** is easier as **FD2** provides a new attack surface – f_θ is forced to output $\mathcal{A}dv$'s target explanations to hide its discriminatory behaviour [14], [45], [71], [200]. Poisons shift the local optima, skew θ and f_θ's decision boundary [45], [71], [146], [200]. Further, **R2** is more effective with more capacity and smoother curvature [14], [71], [146]. Furthermore, explanations such as saliency maps and counterfactuals, can be similarly manipulated [71], [146]. Less number of attributes reduces **R2** by reducing the information available to $\mathcal{A}dv$.

● **R3 (Unauthorized model ownership)** is easier as **FD2** provides information about f_θ's decision boundary [5], [133], [176]. For instance, exploiting counterfactuals results in f_θ^{der} with high accuracy and fidelity with fewer queries than exploiting predictions [5]. The variability in **R3** increases with $\mathcal{A}dv$'s dataset size and imbalanced dataset [5]. The query efficiency can be improved by using a "counterfactual of a counterfactual" to train f_θ^{der} [176].

● **P1 (Membership inference)** is easier as **FD2** leaks membership status [98], [127], [133], [144]: explanations reflect the distance of data records to the boundary which makes them distinguishable for records inside and outside $\mathcal{D}_{tr}$. Further, **P1** decreases when there are fewer attributes but increases with $|\mathcal{D}_{tr}|$, and f_θ's capacity [127].

● **P2 (Data reconstruction)** is easier with releasing **FD2** [209]: explanations are a proxy for f_θ's sensitivity to different attribute values which allow for accurate reconstruction – the risk increases with the quality of explanations. Moreover, some algorithms release influential $\mathcal{D}_{tr}$ data records for a given prediction on some input. This allows for perfect reconstruction of $\mathcal{D}_{tr}$ specially for data with a longer tail and with fewer attributes [144].

● **P3 (Attribute inference)** is more effective as **FD2** leaks the values of sensitive attributes [46]. The higher distinguishability across subgroups is a potential contributor.

● **F (Discriminatory behaviour)** can be detected using **FD2** by identifying whether the relevant attributes are learnt by f_θ [90], [139]. This suggests that **FD2** reduces **F** (●). However, explanations are flawed, i.e., they exhibit discriminatory behaviour across subgroups(●) [12], [40]. This is more pronounced for large capacity [40]. Also, explanation quality varies with the number of attributes [12].

4.4. Guideline: Exploring Unintended Interactions

We now present a guideline to identify the nature of an unintended interaction and the underlying factors influencing it. In Table 2, we summarized statements about how a change in: 1) effectiveness of <d> correlates with a change in <f>, 2) <f> correlates with <r>. We used ↑ to identify positive correlation and ↓ for negative correlation. Hence, for a given <d> and <r> with a common factor <f_c>, we can infer how a change in effectiveness of <d> correlates with <r> using the direction of the pair of arrows that describe how each of <d> and <r> correspond to <f_c>.

Creating an exact algorithm for predicting unexplored interactions is difficult due to the intricate interplay of various factors. Therefore, we outline a guideline to make such conclusions with some expert knowledge about the nature of unintended interactions and underlying factors:

G1 For a given <d> and <r>, identify <f_c>s from Table 2.
G2 For each <f_c>, if change in <f> on using <d> aligns with change in <r>, i.e., (↑,↑) or (↓,↓), then effectiveness of <d> corresponds to an increase in <r> (●) else <d> negatively correlates with <r> (●) for (↑,↓) or (↓,↑).
G3 Consider the suggestions from all <f_c>s to conjecture the nature of the interaction. The final conclusion is: a) common suggestion when all the <f_c> agree on it, b) the suggestion from the dominant <f_c> determined using expert knowledge for conflicting suggestions from <f_c>s.
G4 Empirically evaluate to see if other non-dominant <f>s influence the interaction.

Choosing Dominant <f> with Expert Knowledge. Some factors, namely, O1, O2, O3, are exploited by different attacks. For instance, most inference attacks rely on O2 while **R1** (evasion) and **R2** (poisoning) rely on O1, and O3. Using a defense affects these factors, which has a significant effect on the attack success. Therefore, we identify them as dominant compared to other factors which remain constant while evaluating if the risk changes on using a defense.

Guideline's Coverage of Interactions. To see how well our guideline covers different interactions, we compare conclusions from our guideline with corresponding observations from our survey as the ground truth (Section 4). The guideline and the ground truth are consistent in most cases. We discuss a few exceptions in Section 7. Next, we use our guideline to conjecture about unexplored interactions.

5. Unexplored Interactions and Conjectures

We enumerate different unexplored interactions (● in Table 3) in Table 4. As all defenses and risks we surveyed hinge on factors related to overfitting and memorization (Section 4.1 and 4.2), we conjecture that overfitting and memorization will influence these unexplored interactions. To demonstrate the applicability of our framework, we apply our guideline from Section 4.4 to present conjectures for two of these unexplored interactions and validate them empirically [1] i) Section 5.1: **FD2** (explanations) and **P4** (distribution inference), ii) Section 5.2: **FD1** (group fairness) and **P2** (data reconstruction). Among the unexplored interactions in Table 3, we selected two examples for empirical evaluation based on ease of implementation – we already had convenient implementations for both attacks and defenses involved in these two examples. We mark the factors evaluated for these interactions with $*$ in Table 3. These examples are intended for illustrating the applicability of our framework. We leave a systematic empirical exploration of unintended interactions as future work. In Section 5.3, we speculate on the remaining interactions which we could not evaluate empirically.

TABLE 4. UNEXPLORED INTERACTIONS (● FROM TABLE 3) FROM A DEFENSE TO A RISK (INDICATED WITH "→" BETWEEN THEM).

Defense → Risk
RD1 (Adversarial Training) → **P3** (Attribute Inference)
RD2 (Outlier Removal) → **R3** (Unauthorized Model Ownership)
RD2 (Outlier Removal) → **P2** (Data Reconstruction)
RD2 (Outlier Removal) → **P4** (Distribution Inference)
RD3 (Watermarking) → **R1** (Evasion)
RD4 (Fingerprinting) → **R2** (Poisoning)
RD4 (Fingerprinting) → **P2** (Data Reconstruction)
RD4 (Fingerprinting) → **P3** (Attribute Inference)
RD4 (Fingerprinting) → **P4** (Distribution Inference)
RD4 (Fingerprinting) → **F** (Discriminatory Behaviour)
PD1 (Differential Privacy) → **R3** (Unauthorized Model Ownership)
FD1 (Group Fairness) → **R3** (Unauthorized Model Ownership)
FD1 (Group Fairness) → **P2** (Data Reconstruction)
FD2 (Explanations) → **P4** (Distribution Inference)

5.1. FD1 → P2

Conjecture. From Table 2, we identify `O2.2` as a common factor (<f_c>). As the arrows are in opposite directions, we conjecture that **FD1** reduces **P2** →●. Now, we check for other factors: `D3` suggests that **P2** decreases anyway when there are more input attributes. Thus, we expect that the decrease due to **FD1** is only evident when the number of attributes is small. Consequently, the difference in attack success between f_θ^{fair} and f_θ becomes negligible. We describe the experimental setup below.

Threat Model. We assume x is sensitive and confidential. $\mathcal{A}dv$ only observes $f_\theta^{fair}(x)$ for some unknown confidential x which are to be reconstructed. $\mathcal{A}dv$ also has access to auxiliary dataset which is split into $\mathcal{D}_{tr}^{adv}$ to train an attack decoder model (f_{att}) and $\mathcal{D}_{te}^{adv}$ for evaluation.

Attack Methodology. $\mathcal{A}dv$ uses $\mathcal{D}_{tr}^{adv}$ to train f_{att} to map $f_\theta^{fair}(x)$ to x, minimizing the error between $f_{att}(\bar{f}_\theta^{fair}(x))$ and the ground truth x. Here, $\bar{f}_\theta^{fair}$ is a "shadow model" which is trained on $\mathcal{A}dv$'s auxiliary dataset to have similar functionality as f_θ^{fair}. We give the advantage to $\mathcal{A}dv$ and assume $f_\theta^{fair} == \bar{f}_\theta^{fair}$. The attack is evaluated on $\mathcal{D}_{te}^{adv}$.

Setup. We use the CENSUS dataset which consists of attributes like age, race, education level, to predict whether an individual's annual income exceeds 50K. We select the top 10 attributes and consider both race and sex attributes. We use $|\mathcal{D}_{tr}|$=15470, $|\mathcal{D}_{tr}^{adv}|$=7735, and $|\mathcal{D}_{te}^{adv}|$=7735.

We train f_θ^{fair} with adversarial debiasing which uses an adversarial network f_{inf} to infer sensitive attribute values from $f_\theta^{fair}(x)$ [197]. The loss from f_{inf} is used to train f_θ^{fair} such that $f_\theta^{fair}(x)$ is indistinguishable.

We measure fairness by (a) checking if p%-rule is $> 80\%$ as required by the regulations [194], and (b) AUC score of f_{inf} to correctly infer the sensitive attribute values (fair model should have a score of 0.50). To measure attack success, we use reconstruction loss (l_2-norm) between x and $f_{att}(\bar{f}_\theta^{fair}(x))$ (higher the loss, less effective the attack). We use a multi-layered perceptron (MLP) for f_θ^{fair} with hidden layer dimensions [32, 32, 32, 1], f_{inf} with [32, 32, 32, 2], and f_{att} with [18, 32, 64, 128, 10].

We compare the attack success with the baseline of data reconstruction against f_θ with the same architecture as f_θ^{fair} but without group fairness.

Empirical Validation. We validate our conjecture, show the effect on distinguishability across subgroups (`O2.2`), and evaluate the influence of the number of input attributes (`D3`).

Nature of Interaction. We compare the reconstruction loss between f_θ^{fair} and f_θ in Table 5. We confirm that training with adversarial debiasing is indeed fair (p%-rule $> 80\%$ and AUC score ~ 0.50). We observe that f_θ^{fair} has a higher reconstruction loss validating our conjecture.

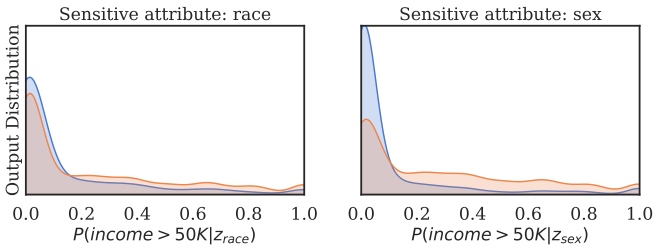
TABLE 5. **NATURE OF INTERACTION.** COMPARISON OF $\mathcal{D}_{te}$ ACCURACY, FAIRNESS (AUC SCORES, P%-RULE VALUES) AND RECONSTRUCTION LOSS FOR f_θ AND f_θ^{fair} ACROSS TEN RUNS. ↑: HIGHER VALUE IS DESIRABLE AND ↓: LOWER VALUE IS DESIRABLE.

Metric	f_θ	f_θ^{fair}
Acc ↑	84.40 ± 0.09	77.96 ± 0.58
AUC ↓	0.66 ± 0.00	0.52 ± 0.01
p%-rule **Race** ↑	46.51 ± 0.59	93.69 ± 5.60
p%-rule **Sex** ↑	29.70 ± 1.15	90.50 ± 6.87
ReconLoss ↑	0.85 ± 0.01	0.95 ± 0.02

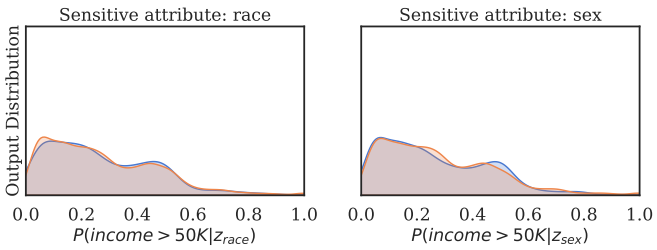
Distinguishability in observables across subgroups (`O2.2`). We plot f_θ^{fair}'s prediction distribution for race (whites as red vs non-whites as blue) and sex (males as red vs females as blue) (Figure 2).

We notice that there is no income disparity between whites-non-whites and males-females for f_θ^{fair}. Further, f_θ^{fair}'s predictions are normalized which reduces the information available to $\mathcal{A}dv$ for successful reconstruction.

1. Code: https://github.com/ssg-research/sok-unintended-interactions.



(a) f_θ's Prediction Distribution



(b) f_θ^{fair}'s Prediction Distribution

Figure 2. Distinguishability across subgroups (`O2.2`) decreases for f_θ^{fair}.

Number of Input Attributes (`D3`). We present the impact of the number of input attributes in Table 6. Firstly, we note that the accuracy is similar across different number of input attributes for f_θ^{fair} regardless of the difference in reconstruction loss. Hence, the accuracy does not impact the risk, and only the number of input attributes is the variable.

TABLE 6. **NUMBER OF INPUT ATTRIBUTES (`D3`).** RECONSTRUCTION LOSS AND $\mathcal{D}_{te}$ ACCURACY FOR DIFFERENT NUMBER OF INPUT ATTRIBUTES FOR f_θ AND f_θ^{fair} (AVERAGED ACROSS TEN RUNS).

# Input Attributes	f_θ		f_θ^{fair}	
	ReconLoss	Acc	ReconLoss	Acc
10	0.85 ± 0.00	84.40 ± 0.09	0.95 ± 0.02	78.96 ± 0.58
20	0.93 ± 0.03	84.72 ± 0.22	0.93 ± 0.00	80.32 ± 1.12
30	0.95 ± 0.02	84.41 ± 0.39	0.94 ± 0.00	79.50 ± 0.91

We present the reconstruction loss for a different number of input attributes $\{10, 20, 30\}$. As expected, for 10 attributes, group fairness reduces attack success significantly (as also seen in Table 5). However, increasing the number of input attributes has no significant difference in reconstruction loss between f_θ and f_θ^{fair}, validating our conjecture.

5.2. FD2 → P4

Conjecture. We identify `O2.1` as <f_c>. As the arrows are the same direction, we conjecture that releasing **FD2** increases **P4** (●). Now, we check for other factors: the number of input attributes (`D3`) and model capacity (`M1`). For `D3`, increasing the number of input attributes should reduce **P4** by making it harder to capture distributional properties. For `M1`, increasing f_θ's capacity should increase memorization of distributional properties, thereby increasing **P4**. We describe the experimental setup below.

Threat Model. We assume that on input x, f_θ outputs both explanations $\phi(x)$ and $f_\theta(x)$. $\mathcal{A}dv$ differentiates between f_θ trained on $\mathcal{D}_{tr}$ with ratio α_1 vs. α_2 (e.g., 50% vs. 10% women in $\mathcal{D}_{tr}$) by distinguishing between their respective explanations $\phi(x)^{\alpha_1}$, and $\phi(x)^{\alpha_2}$ using an ML attack model f_{att}. $\mathcal{A}dv$ has an auxiliary dataset which is split into $\mathcal{D}_{tr}^{adv}$ and $\mathcal{D}_{te}^{adv}$ to train and evaluate f_{att}.

Attack Methodology. Given two ratios, α_1 or α_2, $\mathcal{A}dv$ samples $\mathcal{D}_{tr}^{adv}$ to get ${}^{\alpha_1}\mathcal{D}_{tr}^{adv}$ and ${}^{\alpha_2}\mathcal{D}_{tr}^{adv}$ satisfying α_1 or α_2. $\mathcal{A}dv$ then trains multiple *shadow models* on these datasets which try to mimic f_θ; $\mathcal{A}dv$ uses $\phi'(x)^{\alpha_1}$ and $\phi'(x)^{\alpha_2}$ from the shadow models to train f_{att}. Given f_θ which is trained on $\mathcal{D}_{tr}$ satisfying some unknown ratio, $\mathcal{A}dv$ uses $\phi(x)$ where $x \in \mathcal{D}_{te}^{adv}$ from f_θ as input to f_{att} to infer if $\mathcal{D}_{tr}$'s property was α_1 or α_2.

Setup. We use the CENSUS dataset from Section 5.1 with 42 input attributes. We consider different ratios of women in $\mathcal{D}_{tr}$ as our property of interest in the range $0.1 - 0.9$ incremented by 0.1. Following prior work [155], [156], we set $\alpha_1 = 0.5$ while varying α_2. For each value of α_2, $\mathcal{A}dv$ trains f_{att} to differentiate between $\alpha_1 = 0.5$ and the specific value of α_2. We measure the attack success using (distribution) inference accuracy, where 50% is random guess. We use an MLP with the following hidden layer dimensions: [1024, 512, 256, 128]. We use three explanation algorithms: IntegratedGradients [154], DeepLift [7], [145], and SmoothGrad [147]. For explanations, we want to evaluate if releasing explanations increases **P4**. Following the approach by Pawelczyk et al. [127], assuming random guess (50%) as the baseline, we want to show that an attack that *only* makes use of explanations, can do better than the baseline. Our goal is to see if explanations leak distribution information, and not whether explanations can improve the state-of-the-art distribution inference attacks.

Empirical Validation. We validate our conjecture and evaluate how it is influenced by the number of input attributes (`D3`), and model capacity (`M1`).

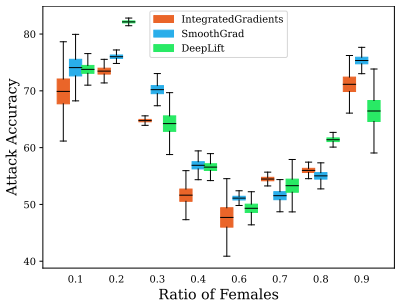


Figure 3. **Nature of Interaction.** Accuracy to differentiate explanations from model trained on $\mathcal{D}_{tr}$ with α_1=0.5 and $\alpha_2 \in \{0.1 - 0.9\}$.

Nature of Interaction. We present the attack accuracy across different ratios and explanation algorithms in Figure 3. We use all 42 attributes for evaluation. For the three explanation algorithms, the attack accuracy is higher than a random guess of 50% for most of the ratios, validating our conjecture – explanations indeed leak distribution properties of $\mathcal{D}_{tr}$, and increase **P4**.

Number of input attributes (D3). We use an attribute selection algorithm to select the top 35, 25 and 15 attributes from the original 42 attributes. We evaluate the impact of varying the top-k relevant attributes on the interaction. We fix α_1=0.5 and α_2=0.1 as it shows the highest attack accuracy (Figure 3). We present the results in Table 7.

TABLE 7. **NUMBER OF INPUT ATTRIBUTES (D3).** ATTACK ACCURACY FOR $\alpha_1 = 0.5$ AND $\alpha_2 = 0.1$ WHILE VARYING THE NUMBER OF INPUT ATTRIBUTES (RESULTS AVERAGED ACROSS TEN RUNS).

# Attributes	IntegratedGradients	DeepLift	SmoothGrad
15	81.07 ± 2.13	78.74 ± 1.66	65.40 ± 1.39
25	66.09 ± 0.95	73.64 ± 1.38	59.42 ± 1.09
35	50.43 ± 0.59	59.93 ± 2.81	56.78 ± 1.93

We confirm that increasing the number of most relevant input attributes, decreases the attack success: more number of attributes decreases the memorization of each attribute thereby reducing the quality of explanations.

Model Capacity (M1). We consider four MLPs with different model capacities: "Model 1" with dimensions [128], "Model 2" with [256, 128], "Model 3" with [256,512,256], and "Model 4" is the same as before. We present the attack accuracy in Table 8.

TABLE 8. **MODEL CAPACITY (M1).** ATTACK ACCURACY ON EXPLANATIONS CORRESPONDING TO α_1=0.5 AND α_2=0.1 FOR DIFFERENT MODELS (AVERAGED ACROSS TEN RUNS).

# Parameters	IntegratedGradients	DeepLift	SmoothGrad
Model 1 (5762)	47.57 ± 4.25	49.19 ± 2.75	53.26 ± 0.10
Model 2 (44162)	53.29 ± 3.65	50.86 ± 3.24	62.40 ± 0.95
Model 3 (274434	62.60 ± 2.74	67.73 ± 1.69	70.21 ± 0.73
Model 4 (733314)	69.90 ± 3.24	73.78 ± 1.03	74.09 ± 2.17

All models have an accuracy of $82.10 \pm 1.17\%$ and close to perfect $\mathcal{D}_{tr}$ accuracy. Hence, neither accuracy nor $\mathcal{G}_{err}$ impact **P4**. We observe that **P4** increases with capacity: a higher capacity correlates with greater memorization of $\mathcal{D}_{tr}$'s distributional properties.

5.3. Other Unintended Interactions

RD1 (Adversarial Training) and P3 (Attribute Inference). Here, D2 and D4 are the common factors. As the arrows are in the opposite directions, both suggest a decrease in **P3** (●). However, O2.2 is likely to be the dominant factor. Hence, the nature of this interaction will be determined by whether the distinguishability in observables across subgroups increases or decreases with **RD1**.

RD2 (Outlier Removal) and P2 (Data Reconstruction). Here, D1 is the common factor and the arrows are in the same direction, suggesting an increase in **P2** (●). Similar to membership and attribute inference [26], [47], [81], we expect retraining without outliers will make f_θ^{outrem} memorize a new set of data records, thus making them more susceptible to **P2**.

RD2 (Outlier Removal) and P4 (Distribution Inference). D2 is the common factor which has arrows in the same direction which suggests an increase in **P4** (●). Prior work has shown that adding more outliers is likely to increase **P4** [31], [110]. **RD2** removes outliers belonging from tail classes. We can assume that these outliers belong to minority subgroup, and outlier removal will increase the bias thereby increasing the distinguishability in observables for datasets with different properties. Hence, **P4** is likely to increase.

6. Related Work

Surveys/SoKs. There are several surveys covering defenses and risks separately [42], [65], [73], [96], [106], [115], [124], [138]. None of them explore the interactions between defenses and risks. Some prior works cover interactions among defenses and risks. Strobel and Shokri [153] illustrate the limited case of increased membership inference with robustness, fairness, and explanations. Noppel and Wressnegger [121] cover specific interaction between explanations and evasions/poisoning. Chen et al. [32] present conflicts between fairness and different privacy risks. Gittens et al. [61] study the interactions of different risks with DP, fairness and robustness but do not cover the factors underlying these interactions. Ferry et al. [52] evaluate the interactions among fairness, privacy and interpretability but do not cover the factors underlying these interactions.

Interactions within Defenses. An orthogonal line of work evaluates trade-offs within defenses or designs algorithms to improve them. For instance, it is impossible to achieve group fairness and DP together [6], [39], [54]. Some prior works attempt to reconcile them by offering pareto-optimal solutions with an accuracy drop [77], [186]. Further, privacy and transparency are inherently conflicting and designing explanations with DP degrades the quality of explanations [66], [122], [125]. Conflicts also arise between watermarking with adversarial training and DP [158]. Gradients from models with adversarial training are interpretable which suggest potential alignment between transparency and robustness [134]. Furthermore, prior works have designed objective functions to reconcile adversarial training and group fairness [183]. Finally, DP over image pixels can provide certified robustness against evasion [93].

7. Discussion and Conclusions

Completeness of the framework. We identified overfitting and memorization (and their underlying factors) as the causes for unintended interactions based on prior work. These are sufficient to explain all the unintended interactions explored in the literature as seen in our survey (Sections 3–4). We do not claim them to be complete considering the complexity of ML models. Whether there are indeed other possible underlying causes, remains an open question. Regardless, our framework is flexible enough to accommodate emerging causes and factors. To incorporate a new cause, we can use the same methodology as in Section 3–4: identify

factors that influence the cause, describe how they relate with defenses and risks, and use them to conjecture unintended interactions. Further, the framework can be extended to cover variants of current risks and defenses as well as new ones by expanding Tables 2 and 3. Finally, the framework can be extended to account for any exceptions in some interactions stemming from a difference in threat model.

Interactions not covered by our Guideline. There are a few interactions which are not covered by our guideline which can be addressed by extending our framework:

- **RD2** (Outlier Removal), **R2** (Poisoning), and **R3** (Unauthorized Model Ownership) have too few (non-dominant) factors to confidently identify the nature of some interactions: **RD1** (Adversarial training) ⟶ **R2** (Poisoning) and **R3** (Unauthorized Model Ownership); **RD2** (Outlier Removal) ⟶ **R1** (Evasion) and **R2** (Poisoning); **PD1** (Differential Privacy) ⟶ **R2** (Poisoning); **FD1** (Group fairness) ⟶ **R2** (Poisoning); and **FD2** (Explanations) ⟶ **R2** (Poisoning), **R3** (Unauthorized Model Ownership). Additionally, **FD2** (Explanations) ⟶ **R1** (Evasion) do not have any common factors to determine the interaction using our guideline.
- **RD1** (Adversarial Training) and **PD1** (Differential Privacy) ⟶ **F** (Discriminatory Behaviour): D2 is a common factor but it covers distribution over different classes but it does not account for subgroups. That is, data records from the tail classes may not belong to a minority subgroup.

Minimizing unintended interactions. Given the factors influencing unintended interactions, practitioners can mitigate unintended interactions by exploring: 1) data augmentation (D1), 2) balanced classes with undersampling and oversampling (D2), 3) adding proxy attributes or removing redundant attributes (D3), 4) domain generalization algorithms (e.g., invariant risk minimization) [69], [108] (D4), 5) adjusting training hyperparameters (O1, O2, and O3), 6) experimenting with more layers, lower precision, or pruning (M1).

Applicability beyond classifiers. Our framework extends to other models by: 1) identifying unintended interactions, 2) evaluating the relation between defenses, risks, and underlying factors, 3) including them in our framework for insights using our guideline. For example, in generative models where overfitting and memorization manifest as the replication of $\mathcal{D}_{tr}$ [23], [24], [28], D1 and M1 are relevant.

Future work. In on-going work, we are creating a software framework to facilitate systematic empirical exploration of unintended interactions. In addition to empirically evaluating different unexplored interactions (Table 4), a more comprehensive evaluation of how factors correlate with effectiveness of defenses and susceptibility to risks is needed (i.e., expanding Table 2). Similarly, understanding how different variants of risks and defenses (e.g., types of fairness algorithms and metrics) affect the interaction, needs further study. Also, our framework suggests correlation and not causation. *Causal reasoning* can establish causation between different factors with unintended interactions (e.g., [13]). Finally, the interactions are empirically evaluated, both in existing literature and our experiments in Section 5. Formally modeling these interactions and predicting their nature mathematically is for future work.

Takeaways. We highlight the following key takeaways: i) Current literature lacks a comprehensive understanding of unintended interactions between defenses (intended to protect against some risks), and *other unrelated risks*. But understanding such interactions is critical for deployment. We showed that identifying common factors (for overfitting and memorization) between a defense and a risk can shed light on whether these factors can induce unintended interactions, ii) an important open problem is: how to design effective defenses and select the right combination of defenses to minimize increases in susceptibility to other risks?

Acknowledgements

This work is supported in part by Intel (Private AI consortium) and the Government of Ontario. Views expressed in the paper are those of the authors and do not necessarily reflect the position of the funding agencies.

References

[1] J. Aalmoes *et al.*, "Leveraging algorithmic fairness to mitigate blackbox attribute inference attacks," *arXiv:2211.10209*, 2022.

[2] M. Abadi *et al.*, "Deep learning with differential privacy," in *CCS*, 2016, p. 308–318.

[3] Y. Adi *et al.*, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in *USENIX Security*, 2018, pp. 1615–1631.

[4] A. Agarwal *et al.*, "A reductions approach to fair classification," in *ICML*, vol. 80, 2018, pp. 60–69.

[5] U. Aïvodji *et al.*, "Model extraction from counterfactual explanations," *arXiv:2009.01884*, 2020.

[6] G. Alves *et al.*, "Survey on fairness notions and related tensions," *EURO Journal on Decision Processes*, 2023.

[7] M. Ancona *et al.*, "Towards better understanding of gradient-based attribution methods for deep neural networks," in *ICLR*, 2018.

[8] D. Arpit *et al.*, "A closer look at memorization in deep networks," in *ICML*, 2017, p. 233–242.

[9] G. Ateniese *et al.*, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," *Int. J. Secur. Netw.*, vol. 10, no. 3, p. 137–150, Sep. 2015.

[10] B. G. Atli Tekgul and N. Asokan, "On the effectiveness of dataset watermarking," in *IWSPA*, 2022, p. 93–99.

[11] E. Bagdasaryan *et al.*, "Differential privacy has disparate impact on model accuracy," in *NeurIPS*, 2019, pp. 15 479–15 488.

[12] A. Balagopalan *et al.*, "The road to explainability is paved with bias: Measuring the fairness of explanations," in *FaccT*, 2022, p. 1194–1206.

[13] T. Baluta *et al.*, "Membership inference attacks and generalization: A causal perspective," in *CCS*, 2022, p. 249–262.

[14] H. Baniecki *et al.*, "Fooling partial dependence via data poisoning," in *Machine Learning and Knowledge Discovery in Databases*, 2023, pp. 121–136.

[15] M. Belkin *et al.*, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proc. of National Academy of Sciences*, vol. 116, no. 32, pp. 15 849–15 854, 2019.

[16] P. Benz *et al.*, "Robustness may be at odds with fairness: An empirical study on class-wise accuracy," in *NeurIPS Workshop on Pre-registration in ML*, 2021, pp. 325–342.

[17] P. Blanchard *et al.*, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *NeurIPS*, 2017.

[18] F. Boenisch *et al.*, "Gradient masking and the underestimated robustness threats of differential privacy in deep learning," *arXiv:2105.07985*, 2021.

[19] E. Borgnia *et al.*, "Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff," in *ICASSP*, 2021, pp. 3855–3859.

[20] G. Brown *et al.*, "When is memorization of irrelevant training data necessary for high-accuracy learning?" in *STOC*, 2021, p. 123–132.

[21] Z. Bu and Y. Zhang, "Differentially private optimizers can learn adversarially robust models," *TMLR*, 2023.

[22] X. Cao *et al.*, "Ipguard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary," in *AsiaCCS*, 2021, p. 14–25.

[23] N. Carlini *et al.*, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *USENIX Security*, 2019, pp. 267–284.

[24] ——, "Extracting training data from large language models," in *USENIX Security*, 2021, pp. 2633–2650.

[25] ——, "Membership inference attacks from first principles," in *S&P*, 2022, pp. 1897–1914.

[26] ——, "The privacy onion effect: Memorization is relative," in *NeurIPS*, 2022.

[27] ——, "Quantifying memorization across neural language models," *arXiv:2202.07646*, 2022.

[28] ——, "Extracting training data from diffusion models," in *USENIX Security*, 2023.

[29] H. Chang *et al.*, "On adversarial bias and the robustness of fair machine learning," in *arXiv:2006.08669*, 2020.

[30] H. Chang and R. Shokri, "On the privacy risks of algorithmic fairness," in *EuroS&P*, 2021, pp. 292–303.

[31] H. Chaudhari *et al.*, "Snap: Efficient extraction of private properties with poisoning," in *S&P*, 2023, pp. 400–417.

[32] H. Chen, T. Zhu, T. Zhang, W. Zhou, and P. S. Yu, "Privacy and fairness in federated learning: On the perspective of tradeoff," *ACM Comput. Surv.*, vol. 56, no. 2, sep 2023.

[33] J. Chen *et al.*, "Hopskipjumpattack: A query-efficient decision-based attack," in *2020 S&P*, 2020, pp. 1277–1294.

[34] M. Chen and O. Ohrimenko, "Protecting global properties of datasets with distribution privacy mechanisms," 2022.

[35] Y. Chen *et al.*, "Amplifying membership exposure via data poisoning," in *NeurIPS*, 2022.

[36] C. A. Choquette-Choo *et al.*, "Label-only membership inference attacks," in *ICML*, 2021, pp. 1964–1974.

[37] J. R. Correia-Silva *et al.*, "Copycat cnn: Stealing knowledge by persuading confession with random non-labeled data," in *IJCNN*, 2018, pp. 1–8.

[38] F. Croce *et al.*, "Robustbench: a standardized adversarial robustness benchmark," in *NeurIPS Datasets and Benchmarks Track*, 2021.

[39] R. Cummings *et al.*, "On the compatibility of privacy and fairness," in *Adjunct Publication of UMAP*, 2019, p. 309–315.

[40] J. Dai *et al.*, "Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations," in *AIES*, 2022, p. 203–214.

[41] Y. Dar *et al.*, "A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning," *arXiv:2109.02355*, 2021.

[42] E. De Cristofaro, "An overview of privacy in machine learning," *arXiv:2005.08679*, 2020.

[43] A. Demontis *et al.*, "Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks," in *USENIX Security*, 2019, pp. 321–338.

[44] G. J. V. den Burg and C. Williams, "On memorization in probabilistic deep generative models," in *NeurIPS*, 2021.

[45] A.-K. Dombrowski *et al.*, "Explanations can be manipulated and geometry is to blame," *NeurIPS*, vol. 32, 2019.

[46] V. Duddu and A. Boutet, "Inferring sensitive attributes from model explanations," in *CIKM*, 2022, p. 416–425.

[47] V. Duddu *et al.*, "Shapr: An efficient and versatile membership privacy risk metric for machine learning," *arXiv:2112.02230*, 2021.

[48] ——, "Towards privacy aware deep learning for embedded systems," in *SAC*, 2022, p. 520–529.

[49] M. S. Esipova, A. A. Ghomi, Y. Luo, and J. C. Cresswell, "Disparate impact in differential privacy from gradient misalignment," in *ICLR*, 2023.

[50] V. Feldman, "Does learning require memorization? a short tale about a long tail," in *STOC*, 2020, pp. 954–959.

[51] V. Feldman and C. Zhang, "What neural networks memorize and why: Discovering the long tail via influence estimation," *NeurIPS*, pp. 2881–2891, 2020.

[52] J. Ferry, U. Aïvodji, S. Gambs, M.-J. Huguet, and M. Siala, "Sok: Taming the triangle–on the interplays between fairness, interpretability and privacy in machine learning," *arXiv:2312.16191*, 2023.

[53] J. Ferry *et al.*, "Exploiting fairness to enhance sensitive attributes reconstruction," in *SaTML*, 2023, pp. 18–41.

[54] F. Fioretto *et al.*, "Differential privacy and fairness in decisions and learning tasks: A survey," in *IJCAI*, 2022, pp. 5470–5477.

[55] M. Fredrikson *et al.*, "Model inversion attacks that exploit confidence information and basic countermeasures," in *CCS*, 2015, p. 1322–1333.

[56] ——, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *USENIX Security*, 2014, p. 17–32.

[57] S. Galhotra *et al.*, "Causal feature selection for algorithmic fairness," in *SIGMOD*, 2022, p. 276–285.

[58] K. Ganju *et al.*, "Property inference attacks on fully connected neural networks using permutation invariant representations," in *CCS*, 2018, p. 619–633.

[59] J. Geiping *et al.*, "Inverting gradients - how easy is it to break privacy in federated learning?" in *NeurIPS*, 2020, pp. 16 937–16 947.

[60] S. Geman *et al.*, "Neural networks and the bias/variance dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992.

[61] A. Gittens *et al.*, "An adversarial perspective on accuracy, robustness, fairness, and privacy: Multilateral-tradeoffs in trustworthy ml," *IEEE Access*, vol. 10, pp. 120 850–120 865, 2022.

[62] I. J. Goodfellow *et al.*, "Explaining and harnessing adversarial examples," *arXiv:1412.6572*, 2014.

[63] K. Grosse *et al.*, "On the (statistical) detection of adversarial examples," *arXiv:1702.06280*, 2017.

[64] T. Gu *et al.*, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, vol. 7, pp. 47 230–47 244, 2019.

[65] R. Guidotti *et al.*, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, 2018.

[66] F. Harder *et al.*, "Interpretable and differentially private predictions," in *AAAI*, 2020, pp. 4083–4090.

[67] M. Hardt *et al.*, "Equality of opportunity in supervised learning," *NeurIPS*, 2016.

[68] ——, "Train faster, generalize better: Stability of stochastic gradient descent," in *ICML*, 2016, p. 1225–1234.

[69] V. Hartmann *et al.*, "Distribution inference risks: Identifying and mitigating sources of leakage," in *SaTML*, 2023, pp. 136–149.

[70] J. Hayes, "Trade-offs between membership privacy & adversarially robust learning," in *arXiv:2006.04622*, 2020.

[71] J. Heo *et al.*, "Fooling neural network interpretations via adversarial model manipulation," in *NeurIPS*, 2019.

[72] S. Hong *et al.*, "On the effectiveness of mitigating data poisoning attacks with gradient shaping," *arXiv:2002.11497*, 2020.

[73] H. Hu *et al.*, "Membership inference attacks on machine learning: A survey," *ACM Comput. Surv.*, 2022.

[74] Y. Hu *et al.*, "Understanding the impact of adversarial robustness on accuracy disparity," in *ICML*, 2023.

[75] T. Humphries *et al.*, "Investigating membership inference attacks under data dependencies," in *CSF*, 2023, pp. 473–488.

[76] A. Ilyas *et al.*, "Adversarial examples are not bugs, they are features," *NeurIPS*, vol. 32, 2019.

[77] M. Jagielski *et al.*, "Differentially private fair learning," in *ICML*, 2019, pp. 3000–3008.

[78] ——, "Subpopulation data poisoning attacks," in *CCS*, 2021, pp. 3104–3122.

[79] M. Jagielski and A. Oprea, "Does differential privacy defeat data poisoning?" *Workshop on Distributed and Private Machine Learning, ICLR*, 2021.

[80] B. Jayaraman and D. Evans, "Evaluating differentially private machine learning in practice," in *USENIX Security*, 2019.

[81] ——, "Are attribute inference attacks just imputation?" in *CCS*, 2022, p. 1569–1582.

[82] R. Jia *et al.*, "Efficient task-specific data valuation for nearest neighbor algorithms," *Proc. VLDB Endow.*, vol. 12, no. 11, p. 1610–1623, 2019.

[83] ——, "Towards efficient data valuation based on the shapley value," in *AISTATS*, 2019, pp. 1167–1176.

[84] ——, "Scalability vs. utility: Do we have to sacrifice one for the other in data importance quantification?" in *CVPR*, 2021.

[85] M. Juuti *et al.*, "Prada: protecting against dnn model stealing attacks," in *EuroS&P*, 2019, pp. 512–527.

[86] F. Kamiran and T. Calders, "Data pre-processing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, 2011.

[87] N. Kandpal *et al.*, "Deduplicating training data mitigates privacy risks in language models," in *ICML*, 2022.

[88] H. Karimi *et al.*, "Characterizing the decision boundary of deep neural networks," *arXiv:1912.11460*, 2019.

[89] K. Khaled *et al.*, "Careful what you wish for: on the extraction of adversarially trained models," in *PST*, 2022, pp. 1–10.

[90] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *ICML*, 2018, pp. 2668–2677.

[91] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *ICML*, 2017, pp. 1885–1894.

[92] K. Krishna *et al.*, "Thieves of sesame street: Model extraction on bert-based apis," in *ICLR*, 2020.

[93] M. Lecuyer *et al.*, "Certified robustness to adversarial examples with differential privacy," in *S&P*, 2019, pp. 656–672.

[94] K. Lee *et al.*, "Deduplicating training data makes language models better," in *ACL*, 2022, pp. 8424–8445.

[95] G. Li *et al.*, "Adversarial training over long-tailed distribution," *arXiv:2307.10205*, 2023.

[96] L. Li *et al.*, "Sok: Certified robustness for deep neural networks," in *S&P*, 2023, pp. 1289–1310.

[97] Y. Li *et al.*, "Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection," *NeurIPS*, vol. 35, pp. 13 238–13 250, 2022.

[98] H. Liu, Y. Wu, Z. Yu, and N. Zhang, "Please tell me more: Privacy impact of explainability through the lens of membership inference attack," in *SP*, 2024, pp. 120–120.

[99] J. Liu *et al.*, "False claims against model ownership resolution," *arXiv:2304.06607*, 2023.

[100] Y. Liu *et al.*, "{ML-Doctor}: Holistic risk assessment of inference attacks against machine learning models," in *USENIX Security*, 2022, pp. 4525–4542.

[101] S. Lounici *et al.*, "Blindspot: Watermarking through fairness," in *IH&MMSec*, 2022, p. 39–50.

[102] N. Lukas *et al.*, "Deep neural network fingerprinting by conferrable adversarial examples," in *ICLR*, 2021.

[103] ——, "Sok: How robust is image classification deep neural network watermarking?" in *S&P*, 2022, pp. 787–804.

[104] X. Ma *et al.*, "On the tradeoff between robustness and fairness," in *NeurIPS*, 2022.

[105] Y. Ma *et al.*, "Data poisoning against differentially-private learners: Attacks and defenses," in *IJCAI*, 2019, p. 4732–4738.

[106] G. R. Machado *et al.*, "Adversarial machine learning in image classification: A survey toward the defender's perspective," *ACM Comput. Surv.*, vol. 55, no. 1, 2021.

[107] A. Madry *et al.*, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[108] D. Mahajan *et al.*, "The connection between out-of-distribution generalization and privacy of ml models," in *Workshop on Privacy Preserving Machine Learning, NeurIPS*, 2020.

[109] ——, "Domain generalization using causal matching," in *ICML*, 2021, pp. 7313–7324.

[110] S. Mahloujifar *et al.*, "Property inference from poisoning," in *S&P*, 2022, pp. 1120–1137.

[111] P. Maini *et al.*, "Dataset inference: Ownership resolution in machine learning," in *ICLR*, 2021.

[112] M. Malekzadeh *et al.*, "Honest-but-curious nets: Sensitive attributes of private inputs can be secretly coded into the classifiers' outputs," in *CCS*, 2021, p. 825–844.

[113] S. Mehnaz *et al.*, "Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models," in *USENIX Security*, 2022.

[114] N. Mehrabi *et al.*, "Exacerbating algorithmic bias through fairness attacks," *AIES*, pp. 8930–8938, 2021.

[115] ——, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, 2021.

[116] F. A. Mejia *et al.*, "Robust or private? adversarial training makes models more vulnerable to privacy attacks," in *arXiv:1906.06449*, 2019.

[117] L. Melis *et al.*, "Exploiting unintended feature leakage in collaborative learning," in *S&P*, 2019, pp. 691–706.

[118] S.-M. Moosavi-Dezfooli *et al.*, "Robustness via curvature regularization, and vice versa," in *CVPR*, 2019, pp. 9078–9086.

[119] R. Nabi and I. Shpitser, "Fair inference on outcomes," in *AAAI*, 2018.

[120] N.-B. Nguyen *et al.*, "Re-thinking model inversion attacks against deep neural networks," in *CVPR*, 2023, pp. 16 384–16 393.

[121] M. Noppel and C. Wressnegger, "Sok: Explainable machine learning in adversarial environments," in *SP*, 2024, pp. 21–21.

[122] H. Nori *et al.*, "Accuracy, interpretability, and differential privacy via explainable boosting," in *ICML*, 2021, pp. 8227–8237.

[123] T. Orekondy *et al.*, "Knockoff nets: Stealing functionality of black-box models," in *CVPR*, 2019, pp. 4954–4963.

[124] N. Papernot *et al.*, "Sok: Security and privacy in machine learning," in *EuroS&P*, 2018, pp. 399–414.

[125] N. Patel *et al.*, "Model explanations with differential privacy," in *FaccT*, 2022, p. 1895–1904.

[126] A. Paudice *et al.*, "Detection of adversarial training examples in poisoning attacks through anomaly detection," *arXiv:1802.03041*, 2018.

[127] M. Pawelczyk *et al.*, "On the privacy risks of algorithmic recourse," in *AISTATS*, 2023, pp. 9680–9696.

[128] Z. Peng *et al.*, "Fingerprinting deep neural networks globally via universal adversarial perturbations," in *CVPR*, 2022, pp. 13 430–13 439.

[129] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Comput. Surv.*, vol. 55, no. 3, feb 2022.

[130] C. Pinzon *et al.*, "On the incompatibility of accuracy and equal opportunity," *Machine Learning*, pp. 1–30, 2023.

[131] G. Pleiss *et al.*, "On fairness and calibration," *NeurIPS*, 2017.

[132] G. Pruthi *et al.*, "Estimating training data influence by tracing gradient descent," *NeurIPS*, pp. 19 920–19 930, 2020.

[133] P. Quan *et al.*, "On the amplification of security and privacy risks by post-hoc explanations in machine learning models," *arXiv:2206.14004*, 2022.

[134] A. Raghunathan *et al.*, "Understanding and mitigating the tradeoff between robustness and accuracy," in *ICML*, 2020.

[135] M. A. Rahman *et al.*, "Membership inference attack against differentially private deep learning model." *Trans. Data Priv.*, vol. 11, no. 1, pp. 61–79, 2018.

[136] K. Rodolfa *et al.*, "Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy," *Nature Machine Intelligence*, vol. 3, no. 10, pp. 896–904, 2021.

[137] A. Sablayrolles *et al.*, "Radioactive data: tracing through training," in *ICML*, 2020, pp. 8326–8335.

[138] A. Salem *et al.*, "Sok: Let the privacy games begin! a unified treatment of data inference privacy in machine learning," in *S&P*, 2023, pp. 327–345.

[139] R. R. Selvaraju *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[140] A. Shafahi *et al.*, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *NeurIPS*, 2018, p. 6106–6116.

[141] S. Shekhar *et al.*, "Fairod: Fairness-aware outlier detection," in *AIES*, 2021, p. 210–220.

[142] Y. Shen *et al.*, "Towards understanding the impact of model size on differential private classification," in *arXiv:2111.13895*, 2021.

[143] R. Shokri *et al.*, "Membership inference attacks against machine learning models," in *S&P*, 2017, pp. 3–18.

[144] ——, "On the privacy risks of model explanations," in *AIES*, 2021, p. 231–241.

[145] A. Shrikumar *et al.*, "Learning important features through propagating activation differences," in *ICML*, 2017, p. 3145–3153.

[146] D. Z. Slack *et al.*, "Counterfactual explanations can be manipulated," in *NeurIPS*, 2021.

[147] D. Smilkov *et al.*, "Smoothgrad: removing noise by adding noise," in *ICML Workshop on Visualization for Deep Learning*, 2017.

[148] D. Solans *et al.*, "Poisoning attacks on algorithmic fairness," in *ECML PKDD*, 2020, p. 162–177.

[149] C. Song *et al.*, "Machine learning models that remember too much," in *CCS*, 2017, p. 587–601.

[150] C. Song and V. Shmatikov, "Overlearning reveals sensitive attributes," in *ICLR*, 2020.

[151] L. Song *et al.*, "Privacy risks of securing machine learning models against adversarial examples," in *CCS*, 2019, p. 241–257.

[152] P. Stock *et al.*, "Defending against reconstruction attacks using rényi differential privacy," 2023.

[153] M. Strobel and R. Shokri, "Data privacy and trustworthy machine learning," *IEEE Security & Privacy*, no. 01, pp. 2–7, 5555.

[154] M. Sundararajan *et al.*, "Axiomatic attribution for deep networks," in *ICML*, 2017, p. 3319–3328.

[155] A. Suri *et al.*, "Dissecting distribution inference," in *SaTML*, 2023, pp. 150–164.

[156] A. Suri and D. Evans, "Formalizing and estimating distribution inference risks," *PETS*, 2022.

[157] V. M. Suriyakumar *et al.*, "Chasing your long tails: Differentially private prediction in health care settings," in *FaccT*, 2021, p. 723–734.

[158] S. Szyller and N. Asokan, "Conflicting interactions among protection mechanisms for machine learning models," in *AAAI*, 2023, pp. 15 179–15 187.

[159] S. Szyller *et al.*, "Dawn: Dynamic adversarial watermarking of neural networks," in *MM*, 2021, p. 4417–4425.

[160] L. Tao *et al.*, "Better safe than sorry: Preventing delusive adversaries with adversarial training," in *NeurIPS*, 2021.

[161] H. Tian *et al.*, "Fd-mia: Efficient attacks on fairness-enhanced models," *arXiv:2311.03865*, 2023.

[162] Z. Tian *et al.*, "A comprehensive survey on poisoning attacks and countermeasures in machine learning," *ACM Comput. Surv.*, vol. 55, no. 8, 2022.

[163] S. Tople *et al.*, "Alleviating privacy attacks via causal learning," in *ICML*, 2020.

[164] F. Tramèr *et al.*, "Stealing machine learning models via prediction apis," in *USENIX Security*, 2016, p. 601–618.

[165] ——, "Truth serum: Poisoning machine learning models to reveal their secrets," in *CCS*, 2022.

[166] C. Tran *et al.*, "Differentially private empirical risk minimization under the fairness lens," in *NeurIPS*, 2021.

[167] ——, "A fairness analysis on private aggregation of teacher ensembles," *arXiv:2109.08630*, 2021.

[168] ——, "Fairness increases adversarial vulnerability," in *arXiv:2211.11835*, 2022.

[169] D. Tsipras *et al.*, "Robustness may be at odds with accuracy," in *ICLR*, 2019.

[170] N. Tursynbek *et al.*, "Robustness threats of differential privacy," *NeurIPS Privacy-Preserving Machine Learning Workshop*, 2020.

[171] M.-H. Van *et al.*, "Poisoning attacks on fair machine learning," in *DASFAA*, 2022, pp. 370–386.

[172] A. K. Veldanda *et al.*, "Fairness via in-processing in the over-parameterized regime: A cautionary tale with mindiff loss," *TMLR*, 2023.

[173] A. Waheed *et al.*, "Grove: Ownership verification of graph neural networks using embeddings," in *S&P*, 2024.

[174] H. Wang *et al.*, "Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition," in *ICML*, 2022, pp. 23 446–23 458.

[175] Y. Wang and F. Farnia, "On the role of generalization in transferability of adversarial examples," in *UAI*, 2023, pp. 2259–2270.

[176] Y. Wang *et al.*, "Dualcf: Efficient model extraction attack from counterfactual explanations," in *FaccT*, 2022, p. 1318–1329.

[177] M. Weber *et al.*, "Rab: Provable robustness against backdoor attacks," in *S&P*, 2023, pp. 640–657.

[178] R. Wen *et al.*, "Is adversarial training really a silver bullet for mitigating data poisoning?" in *ICLR*, 2023.

[179] Y. Wen *et al.*, "Fishing for user data in large-batch federated learning via gradient magnification," in *ICML*, 2022.

[180] F. Wu *et al.*, "Linkteller: Recovering private edges from graph neural networks via influence analysis," in *S&P*, 2022.

[181] T. Wu *et al.*, "Adversarial robustness under long-tailed distribution," in *CVPR*, 2021, pp. 8659–8668.

[182] C. Xie *et al.*, "Crfl: Certifiably robust federated learning against backdoor attacks," in *ICML*, 2021, pp. 11 372–11 382.

[183] H. Xu *et al.*, "To be robust or to be fair: Towards fairness in adversarial training," in *ICML*, 2021, pp. 11 492–11 501.

[184] Y. Xu *et al.*, "Exploring and exploiting decision boundary dynamics for adversarial robustness," *ArXiv*, vol. abs/2302.03015, 2023.

[185] M. Yaghini *et al.*, "Disparate vulnerability: on the unfairness of privacy attacks against machine learning," in *PETS*, 2022.

[186] ——, "Learning with impartiality to walk on the pareto frontier of fairness, privacy, and utility," *arXiv:2302.09183*, 2023.

[187] Z. Yang *et al.*, "Neural network inversion in adversarial setting via background knowledge alignment," in *CCS*, 2019, p. 225–240.

[188] ——, "Purifier: defending data inference attacks via transforming confidence scores," in *AAAI*, 2023, pp. 10 871–10 879.

[189] ——, "Rethinking bias-variance trade-off for generalization of neural networks," in *ICML*, 2020.

[190] D. Ye *et al.*, "One parameter defense—defending against data inference attacks via differential privacy," *IEEE TIFS*, vol. 17, pp. 1466–1480, 2022.

[191] J. Ye *et al.*, "Enhanced membership inference attacks against machine learning models," in *CCS*, 2022, pp. 3093–3106.

[192] S. Yeom *et al.*, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *CSF*, 2018, pp. 268–282.

[193] H. Yin *et al.*, "See through gradients: Image batch recovery via gradinversion," in *CVPR*, 2021, pp. 16 337–16 346.

[194] M. B. Zafar *et al.*, "Fairness Constraints: Mechanisms for Fair Classification," in *AISTATS*, vol. 54, 2017, pp. 962–970.

[195] R. Zhai *et al.*, "Understanding why generalized reweighting does not improve over erm," in *ICLR*, 2023.

[196] B. Zhang *et al.*, "Privacy for all: Demystify vulnerability disparity of differential privacy against membership inference attack," *arXiv:2001.08855*, 2020.

[197] B. H. Zhang *et al.*, "Mitigating unwanted biases with adversarial learning," in *AIES*, 2018, p. 335–340.

[198] C. Zhang *et al.*, "Understanding deep learning requires rethinking generalization," in *ICLR*, 2017.

[199] ——, "Counterfactual memorization in neural language models," in *NeurIPS*, 2023.

[200] H. Zhang *et al.*, "Data poisoning attacks against outcome interpretations of predictive models," in *KDD*, 2021, p. 2165–2173.

[201] ——, "Theoretically principled trade-off between robustness and accuracy," in *ICML*, 2019, pp. 7472–7482.

[202] J. Zhang *et al.*, "Protecting intellectual property of deep neural networks with watermarking," in *AsiaCCS*, 2018, p. 159–172.

[203] ——, "Privacy leakage of adversarial training models in federated learning systems," in *CVPR Workshops*, 2022, pp. 108–114.

[204] W. Zhang *et al.*, "Leakage of dataset properties in Multi-Party machine learning," in *USENIX Security*, 2021, pp. 2687–2704.

[205] ——, "Attribute privacy: Framework and mechanisms," in *FaccT*, 2022, pp. 757–766.

[206] X. Zhang *et al.*, "Interpretable deep learning under fire," in *USENIX Security*, 2020.

[207] Y. Zhang *et al.*, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *CVPR*, 2020.

[208] Z. Zhang *et al.*, "Model inversion attacks against graph neural networks," *IEEE TKDE*, 2022.

[209] X. Zhao *et al.*, "Exploiting explanations for model inversion attacks," in *CVPR*, 2021, pp. 682–692.

[210] Y. Zheng *et al.*, "A dnn fingerprint for non-repudiable model ownership identification and piracy detection," *IEEE TIFS*, vol. 17, pp. 2977–2989, 2022.

[211] L. Zhu *et al.*, "Deep leakage from gradients," in *NeurIPS*, 2019.

Appendix A. Summary of Notations

TABLE 9. SUMMARY OF NOTATIONS AND THEIR DESCRIPTIONS.

Notation	Description
Standard Model and Metrics	
f_θ	standard ML model (no defense)
θ	f_θ's parameters
$\mathcal{A}$	Training algorithm
$R(\theta)$	Regularization function over θ
λ	Regularization hyperparameter
$l(f_\theta, y)$	Loss function
$f_\theta^k(x)$	k^{th} layer's output (activation)
$\mathcal{C}$	Cost function to be optimized during training
f_{att}	Adversary's attack model
$\mathcal{G}_{err}$	Generalization error (train-test accuracy gap)
Data	
$\mathcal{D}_{tr}$	Training dataset for f_θ
$z_i = (x_i, y_i)$	Data record with attributes x_i and label y_i
$\mathcal{S}$	Random subset of $\mathcal{D}_{tr}$
$\lvert\mathcal{D}_{tr}\rvert$	Size of $\mathcal{D}_{tr}$
$\mathcal{D}_{te}$	Test dataset
$\mathcal{D}_{tr}^{adv}$	Adversary's dataset to train f_{att}
$\mathcal{D}_{te}^{adv}$	Adversary's dataset to evaluate f_{att}
Defenses and Risks	
f_θ^{rob}	Robust model from adversarial training
f_θ^{dp}	Private model from DPSGD
f_θ^{fair}	Fair model with group fairness constrains
f_θ^{der}	Stolen model derived from f_θ
x_{adv}	Adversarial Example
f_θ^{outrem}	Model after retraining on $\mathcal{D}_{tr}$ without outliers
f_θ^{wm}	Model trained with watermarks
ϵ_{rob}	Perturbation budget for adversarial examples
δ_{rob}	Perturbation added to generate x_{adv}
ϵ_{dp}	Privacy budget
δ_{dp}	probability where privacy loss $> e_{dp}^{\epsilon}$
Evaluating Conjectures	
α_1, α_2	Distributional properties of $\mathcal{D}_{tr}$
$\alpha_1\mathcal{D}_{tr}^{adv}$, $\alpha_2\mathcal{D}_{tr}^{adv}$	Datasets satisfying α_1, α_2
$\phi(x)^{\alpha_1}$, $\phi(x)^{\alpha_2}$	Explanations corresponding to α_1 and α_2

Appendix B.
Meta-Review

The following meta-review was prepared by the program committee for the 2024 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

B.1. Summary

The paper study a broad range of risks and their (unintended) induction by defenses. Specifically, the authors observe that the risks associated with machine learning are vast, and while there have been many proposed defenses to address their shortcomings, papers have (informally) observed that defenses may induce other deficiencies explored in other threat models. To this end, the authors propose a systematic framework that categories risks across security and privacy, while simultaneously linking them where empirical or theoretical evidence has been found. With this broad perspective, the authors identify many opportunities for future research, including the lack of evidence for some interactions as well as some interactions that have yet to be measured. Using their framework, the authors perform a set of experiments to validate some conjectures made apparent by the framework and validate their hypotheses empirically. Finally, the authors also provide general guidelines for how unintended interactions can be identified.

B.2. Scientific Contributions

- Independent Confirmation of Important Results with Limited Prior Research
- Provides a Valuable Step Forward in an Established Field
- Establishes a New Research Direction

B.3. Reasons for Acceptance

1) The core scientific contribution of this work is the framework that unifies vast bodies of work across machine learning security and privacy risks. The framework offers convincing evidence on why interactions happen, what has left to be identified, and how future work can systematically explore if a defense induces an unintended interaction.
2) The paper did an excellent job surveying, collecting, and extrapolating the interactions between ML defenses. The bibliography is comprehensive. This paper can greatly impact the implementation of ML defenses and might open new research possibilities for the exploitation of these interactions.
3) The presentation of this paper is exceptional. The ideas are concise, the definitions are rooted in well-grounded concepts of learning theory as well as modern definitions commonly observed in security and privacy literature.