TRUSTWORTHY AI MUST ACCOUNT FOR INTERSECTIONALITY

Jesse C. Cresswell
Layer 6 AI
Toronto, Canada
jesse@layer6.ai

ABSTRACT

Trustworthy AI encompasses many aspirational aspects for aligning AI systems with human values, including fairness, privacy, robustness, explainability, and uncertainty quantification. However, efforts to enhance one aspect often introduce unintended trade-offs that negatively impact others, making it challenging to improve all aspects simultaneously. In this paper, we review notable approaches to five aspects and systematically consider every pair, detailing the negative interactions that can arise. For example, applying differential privacy to model training can amplify biases in the data, undermining fairness. Drawing on these findings, we take the position that addressing trustworthiness along each axis in isolation is insufficient. Instead, to achieve better alignment between humans and AI, efforts in Trustworthy AI must account for intersectionality between aspects and adopt a holistic view across all relevant axes at once. To illustrate our perspective, we provide guidance on how researchers can work towards integrated trustworthiness, and a case study on how intersectionality applies to the financial industry.

1 INTRODUCTION

Artificial intelligence (AI) systems have become widespread for automated decision making across industries, and as productivity aids for consumers. Industries such as banking and insurance increasingly rely on predictive AI models that directly impact customers, while the healthcare sector explores AI-driven advancements in patient care. Increased scrutiny by regulators and concerns around the *trustworthiness* of these systems call for a more measured approach to AI development with considerations beyond raw performance. In response, the field of Trustworthy AI (TAI) has blossomed, with the general goal of aligning AI to human values. Chief among the tenets of TAI are fairness, privacy, robustness, explainability, and uncertainty quantification – each of which is a noble pursuit, but all of which must be harmonized to promote deep trust.

These five core concepts are well-studied individually in machine learning (ML). However, as we will extensively discuss, isolated study fails to account for the complex interactions between TAI aspects. Layering multiple methodologies designed for individual aspects tends to produce unforeseen consequences and negative intersectionalities[1], ultimately undermining trust rather than reinforcing it. Many such negative interactions have been documented, but their prevalence and severity may not yet be fully realized due to the diverse and clustered nature of research on TAI. By collating them in one place, we aim to bring these interactions to light and highlight that a more holistic approach to TAI is needed to achieve the goal of aligning AI with human values. To this end, we broaden the definition of alignment to account for intersectionality between TAI aspects.

2 TRUSTWORTHY AI ASPECTS

Throughout the discussions below we will typically consider a classification model $F_\theta : \mathcal{X} \to \mathcal{Y}$ parameterized by θ, which maps a feature space $\mathcal{X}$ to a discrete set of labels $\mathcal{Y}$. We use capital letters (e.g. X and Y) to denote random variables, while lower case (e.g. $x \in \mathcal{X}$ and $y \in \mathcal{Y}$) indicates data instances. F_θ is trained to minimize a loss function $\mathcal{L}$ over its training set $\mathcal{D}_{\text{train}}$. Softmax outputs are denoted as $f_\theta : \mathcal{X} \to [0,1]^m$, such that $F_\theta(x) = \arg\max_{i \in \mathcal{Y}} f_\theta(x)_i$.

[1] In social sciences, "intersectionality" describes how overlapping social identities create unique experiences (Crenshaw, 1989). In this spirit, we use "intersectionality" to describe how overlapping TAI aspects interact in positive or negative ways.

In the following subsections we provide a brief overview of five TAI aspects, whose interactions we consider in Section 3. These reviews are intentionally selective, not comprehensive, and focus on a limited set of topics to highlight that the negative interactions we examine are commonplace, not arbitrarily chosen from a wide-ranging body of literature.

2.1 FAIRNESS

Fairness is a foundational pillar in the development of TAI, ensuring systems treat diverse populations equally or equitably. Since fairness is a nuanced and highly contextual topic, it cannot be boiled down into a single set of guidelines to follow in all cases. Instead, TAI researchers and practitioners must consider the appropriate fairness definitions and methodologies to use in each circumstance.

We consider the case where the data is partitioned into n_{g} groups based on an attribute $a \in \mathcal{A} = \{1, ..., n_{\text{g}}\}$ (e.g. age bins). The objective of fair ML is to create a model F_θ which treats all groups fairly by equalizing its prediction behaviour. Exactly how this is defined varies from one application to the next. We present several common viewpoints.

Consider the dichotomy between *procedural* and *substantive fairness*. Procedural fairness emphasizes treating all individuals equally (Grgić-Hlača et al., 2016). A model which does not have access to group identifiers cannot treat individuals differently based on that information, leading to *fairness through unawareness* (Zemel et al., 2013; Kusner et al., 2017). *Disparate Treatment* can result when systems are not procedurally fair. Meanwhile, substantive fairness aligns with the concept of equity, and encourages treating individuals differently to achieve comparable outcomes, so-called *fairness through awareness* (Dwork et al., 2012). If individuals do not receive the same beneficial outcomes from a model, there is *Disparate Impact*. Both Disparate Treatment and Impact are commonly used concepts in legal and regulatory frameworks (OCC, 2025). Disparate Impact can be measured in terms of the target outcome of the model, for instance through *accuracy disparity*

$$\Delta_{\text{acc}} = \max_{a,b \in \mathcal{A}} [\text{acc}(F_\theta, \mathcal{D}_a) - \text{acc}(F_\theta, \mathcal{D}_b)], \tag{1}$$

where $\mathcal{D}_a$ denotes the subset of $\mathcal{D}$ belonging to group a.

Minimizing Δ_{acc} is one example of a fairness goal, and can be pursued at several stages including pre-processing (e.g. balancing data across groups before training), in-processing (e.g. adding fairness regularization terms to the loss $\mathcal{L}$), or post-processing (e.g. using model scores differently across groups when making decisions). These *fairness interventions* are examples of intentionally treating groups differently so that outcomes will be more similar.

2.2 PRIVACY

Privacy has become a crucial area of research in TAI as systems increasingly rely on sensitive data for training (Liu et al., 2021). The use of personal information, such as health records, financial transactions, and social media activity, has led to growing concerns about privacy breaches, unauthorized data access, and the risk of re-identification.

In the context of ML, privacy concerns are usually demonstrated adversarially, where an attack is employed to extract as much private information as possible from the model itself, or its outputs. The standard example is a *membership inference attack* (MIA) (Shokri et al., 2017; Ye et al., 2022b) in which the adversary tries to determine if a test datapoint x_{test} was included in $\mathcal{D}_{\text{train}}$. MIAs help to demonstrate when a system can fail to preserve privacy, but an unsuccessful MIA does not indicate the system is safe; there could always exist a stronger adversarial attack that would succeed. Hence, privacy researchers rely on future-proof frameworks that provide statistical guarantees on privacy protection.

Differential privacy (DP) (Dwork et al., 2006) is the primary framework for quantifying how much private information could be exposed by an ML model. Formally, let M be a probabilistic function acting on datasets $\mathcal{D}$. We say that M is (ϵ, δ)-*differentially private* if for all subsets of possible outputs $S \subseteq \text{Range}(M)$, and for all pairs of datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ by the addition or removal of one element,

$$\Pr[M(\mathcal{D}) \in S] \leq \exp(\epsilon) \Pr[M(\mathcal{D}') \in S] + \delta. \tag{2}$$

This inequality guarantees that the function M cannot strongly depend on any one datapoint, and hence the amount of information that can be extracted about any datapoint is bounded. Strong DP

guarantees (i.e. ϵ and δ both close to 0) have been empirically shown to be effective defenses against MIAs and other privacy attacks (Rahman et al., 2018; Ye et al., 2022a). Importantly, no amount of post-processing on the outputs on M can weaken its guarantee.

DP is typically achieved in ML through DPSGD (Abadi et al., 2016), a stochastic gradient descent method that satisfies Equation (2). It first computes per-sample gradients and clips them, then aggregates them before adding noise. For samples x_i, y_i in a batch B, and per-sample gradients $g_i = \nabla_\theta \mathcal{L}(\theta_t; x_i, y_i)$, the DPSGD gradient update is

$$\theta_{t+1} = \theta_t - \lambda \left[\frac{1}{|B|} \sum_{i \in B} \text{clip}_C \left(g_i \right) + \frac{\sigma C}{|B|} \xi \right], \tag{3}$$

where λ is the learning rate, C is the clipping bound, σ is the noise level, and $\xi \sim \mathcal{N}(0, \mathbf{I})$ is Gaussian noise. As training with DPSGD progresses, more *privacy budget* is consumed (ϵ and δ increase), which is usually accounted numerically (Mironov et al., 2019; Yousefpour et al., 2021).

2.3 Robustness

Robustness broadly refers to the ability of a model to maintain its performance and reliability under a variety of conditions. Since stable performance is desired even in unforeseen situations, researchers commonly test robustness adversarially. An attacker will actively try to produce unintended behaviour by perturbing the input of a model, often in ways that are imperceptible to humans (Biggio et al., 2013; Szegedy et al., 2013). *Adversarial examples* can be created through optimization by maximizing the model's loss on the correct answer rather than minimizing:

$$x^\dagger(\theta, x_i, y_i) = \underset{x \in \mathcal{B}(x_i, \varepsilon)}{\arg\max}\, \mathcal{L}(\theta; x, y_i), \tag{4}$$

where $x^\dagger$ is constrained to be "close" to x_i, for instance within a ball $\mathcal{B}(x_i, \varepsilon)$ of radius ε around x_i. Such attacks can be defended against by exposing the model to adversarially perturbed inputs during training (Goodfellow et al., 2014):

$$\theta_{t+1} = \theta_t - \lambda \frac{1}{|B|} \sum_{i \in B} \nabla_\theta \mathcal{L}(\theta_t; x^\dagger(\theta_t, x_i, y_i), y_i). \tag{5}$$

2.4 Explainability

Explainability enables researchers and practitioners to understand, validate, and trust decisions made by complex models, and gives the ability to audit those decisions retroactively. When humans manually accept or reject a model's prediction, explanations help them understand the reasoning behind the decision and compare it to their own expertise.

Some ML models are inherently more interpretable, such as shallow decision trees and linear models, but deep neural networks are not. Methods to explain complex models may focus on *global explanations*, providing an overarching understanding of model behavior, or *local explanations*, which shed light on individual predictions (Linardatos et al., 2021). We will focus on local explanations, especially model-agnostic, feature-based explanations due to their applicability across ML algorithms (Islam et al., 2021). These methods interpret behavior by analyzing the importance of input features for a given prediction, regardless of the underlying architecture. For an input x, in addition to the model's output $F_\theta(x)$, a local explainability method also provides some form of *feature importance* $E_\theta(x)$, a quantification of how important each element of x is for predicting $F_\theta(x)$.

We recount one popular method as a typical example, Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016). LIME aims to provide local explanations that preserve local *fidelity* – that the explanations correspond to the model's actual behaviour in the vicinity of x. LIME's explanations take the form of a model, one that is inherently more interpretable than $F_\theta(x)$, namely a sparse linear model that locally approximates F_θ at x. The weights of the linear model are returned as $E_\theta(x)$ and communicate how important each corresponding feature is.

2.5 Uncertainty Quantification

Typical ML models output a point prediction (e.g. a single label for classification) and are not designed to quantify confidence in those predictions. We note that softmax outputs $f_\theta(x)$ are unreliable because of miscalibration (Guo et al., 2017; Minderer et al., 2021). ML models that quantify

their uncertainty are more trustworthy, as the user can judge when to ignore the model in favor of alternatives (Soize, 2017). While there are several major approaches to uncertainty quantification, we focus on one increasingly popular method, *conformal prediction* (CP) (Vovk et al., 2005; Angelopoulos & Bates, 2021). The idea of CP is to output sets of predictions (e.g. several class labels) where larger sets indicate greater model uncertainty. CP takes a heuristic notion of uncertainty, like $f_\theta(x)$, and calibrates it using a held-out dataset $\mathcal{D}_{\text{cal}}$ to give statistically grounded uncertainty quantification. CP defines a *conformal score* function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where larger values indicate worse agreement between $f_\theta(x)$ and y. After computing s on the n_{cal} calibration datapoints, one finds the $\frac{\lceil (n_{\text{cal}}+1)(1-\alpha) \rceil}{n}$ quantile q of the conformal scores, using a free parameter $\alpha \in (0, 1)$. For a new datapoint x_{test}, prediction sets $\mathcal{C}_q$ are generated by including all output values for which the conformal score is below the threshold q,

$$\mathcal{C}_q(x_{\text{test}}) = \{y \in \mathcal{Y} \mid s(x_{\text{test}}, y) < q\}. \tag{6}$$

Notably, CP provides a *coverage* guarantee over the true label y_{test},

$$\mathbb{P}[y_{\text{test}} \in \mathcal{C}_q(x_{\text{test}})] \geq 1 - \alpha, \tag{7}$$

as long as x_{test} is exchangeable with the calibration data drawn from $\mathbb{P}$. Hence, the user can specify a maximum error rate, α, ensuring that the sets generated during testing will exclude the ground truth no more often than α. For equal coverage levels, the usefulness of prediction sets is judged by their size, with smaller average set sizes $\mathbb{E}|\mathcal{C}_q|$ indicating more confident predictions.

2.6 Other Aspects

Our discussion focuses on the five aspects recounted above. There are, however, *many* additional aspects one may strive to achieve when building TAI, including **safety** (Amodei et al., 2016; Hendrycks, 2024), **alignment** (Russell, 2019; Gabriel, 2020; Sorensen et al., 2024), **diversity** (Buolamwini & Gebru, 2018; Fazelpour & De-Arteaga, 2022), **reproducibility** (Pineau et al., 2021), **accountability** (Cooper et al., 2022), and **human agency** (Fanni et al., 2023). While intersectional TAI strives to encompass as many aspects as possible, examining five aspects in detail is sufficient for us to motivate the need for an intersectional approach.

3 Negative Interactions Between Trustworthy AI Aspects

To demonstrate that negative interactions are commonplace, not the exception, we exhaustively consider every pairwise combination of our five aspects of Fairness (F), Privacy (P), Robustness (R), Explainability (E), and Uncertainty Quantification (UQ). For each pair we give examples of negative implications on one aspect from the application of the other, and cover both directions. For two TAI aspects A and B, we use the shorthand A $\rightharpoonup$ B to indicate that applying a concept or method from A has a negative impact on B. While there are also examples of positive interactions, we focus on negative interactions to demonstrate the potential harms of failing to consider intersectionality in TAI.

3.1 Fairness and Privacy

F $\rightharpoonup$ P: At the most basic level, evaluating or correcting the fairness of an ML model with respect to some group usually necessitates collecting information on the group identifier $\mathcal{A}$. These identifiers, like age, gender, or race, are often sensitive personal information – exactly the type of information that should be afforded privacy. Collecting, storing, and using this information for fairness purposes exposes individuals to greater risk of conventional data leaks or hacks.

Beyond conventional privacy leaks, Section 2.2 discussed how trained models can leak private information through MIAs which exploit the differences in model behaviour between populations. Fairness interventions during training can reduce the differences between populations, and hence better protect against standard MIAs (Tonni et al., 2020). However, such techniques actually increase vulnerability to specialized MIAs which are also group-aware (Tian et al., 2024). Generally, fairness-aware ML algorithms tend to memorize from underrepresented groups, improving model accuracy, but weakening privacy (Chang & Shokri, 2021).

P $\rightharpoonup$ F: Some individuals or groups in the data can be more vulnerable to privacy attacks than others (Long et al., 2020). When vulnerability is unequal, applying privacy-enhancing techniques can improve the privacy of some groups more than others, an example of Disparate Impact

(Kulynych et al., 2022). Protecting a vulnerable group by removing it from the training set is counterproductive, as the vulnerability merely shifts to a different group (Carlini et al., 2022).

While DPSGD is the *de facto* standard method for achieving privacy guarantees on ML models, it is well-known to cause Disparate Impact by increasing accuracy disparity (Equation (1)) as compared to ordinary SGD (Bagdasaryan et al., 2019). Suppose some group $a \in \mathcal{A}$ in the data is underrepresented, or is otherwise more difficult to correctly predict on. In ordinary SGD, data points from this group would have higher loss, and hence larger gradients, which would increase their relative influence on the optimization. In DPSGD (Equation (3)), large gradients with $\|g_i\| > C$ are clipped, making them relatively less influential on the optimization process. This uneven clipping introduces bias into the gradients which is the primary source of Disparate Impact (Tran et al., 2021; Esipova et al., 2023).

3.2 Fairness and Robustness

F $\rightharpoonup$ R: When groups in the dataset are underrepresented, fairness interventions to reduce model bias (Section 2.1) can increase the relative influence of those groups. Unfortunately, this increased influence can make the very same groups more susceptible to adversarial attacks (Chang et al., 2020; Xu et al., 2021). Tran et al. (2024) show that fairness interventions can reduce the average distance from training samples to the decision boundary, which makes them more vulnerable to adversarial examples (Madry et al., 2018).

R $\rightharpoonup$ F: The main method to improve adversarial robustness, namely adversarial training, adds perturbed versions $x^\dagger$ of inputs x to training batches (Equation (5)). The side-effects of adversarial training include decreased overall accuracy on unperturbed samples, but more importantly for fairness, larger disparities in class-wise performance (Nanda et al., 2021; Xu et al., 2021; Benz et al., 2021). This robustness bias has been attributed to properties of the data distribution like feature distributions across groups (Benz et al., 2021), differences in the intrinsic difficulty of classes (Xu et al., 2021), and biased representations learned during pre-training (Nanda et al., 2021).

3.3 Fairness and Explainability

F $\rightharpoonup$ E: Fairness interventions can inadvertently alter the relative importance of features in explanations. Pre-processing modifies the training data through re-balancing or other transformations (Caton & Haas, 2024) which can obscure the true relationships between features and outcomes, making it challenging to interpret model behavior accurately. For instance, if minority groups are oversampled to increase their representation, an explainability method may correspondingly overemphasize the importance of features associated to that group. The same issue can occur from in-processing (Wan et al., 2023) as the influence of various features is altered by, for example, fairness constraints added to the loss. Meanwhile, post-processing methods that modify predictions without altering the underlying model can create a disconnect between the model's internal decision-making process and the actual predictions that are used (Di Gennaro et al., 2024).

E $\rightharpoonup$ F: Explanations introduce another potential source of bias in modeling. Even if a model's predictions are considered fair, the fidelity of explanations may be inconsistent across groups – that is, for some groups the features identified as important in the explanation may not truly reflect the features driving the model's predictions. Fidelity disparity can lead to the model's predictions being trusted more for some groups than others, such that the benefits of the model are not evenly experienced across groups (Balagopalan et al., 2022). Dai et al. (2022) found that *post hoc* explanation methods used on neural networks quite commonly have disparate fidelity across groups.

Alternatively, explanations may hide biases in an unfair model. For instance, explanations may fail to accurately represent that a model is relying on sensitive attributes, covering up active discrimination (Lakkaraju & Bastani, 2020; Slack et al., 2020).

3.4 Fairness and Uncertainty Quantification

Background: In conformal prediction the coverage guarantee in Equation (7) holds marginally over the entire distribution $\mathbb{P}$. Hence, it is possible that some groups within the distribution have lower coverage than others, leading to tension between fairness and UQ. A stronger guarantee is *group-wise conditional coverage* with respect to pre-defined groups $\mathcal{A}$,

$$\mathbb{P}[y \in \mathcal{C}(x) \mid A = a] \geq 1 - \alpha, \quad \forall\, a \in \mathcal{A}. \tag{8}$$

Group-wise conditional coverage can easily be obtained by partitioning $\mathcal{D}_{\text{cal}}$ by groups, and performing CP on each $\mathcal{D}_a$, giving distinct thresholds q_a. At test time, sets are generated using Equation (6) with the appropriate q_a (Vovk et al., 2003).

F $\rightharpoonup$ UQ: The idea of providing equal levels of coverage across groups $\mathcal{A}$ for the sake of fairness was discussed by Romano et al. (2020a), who argued that *Equalized Coverage* should be the standard of fairness for CP. Using notation similar to Equation (1), we can express Equalized Coverage as

$$\Delta_{\text{Cov}} = \max_{a,b \in \mathcal{A}} \left(\mathbb{P}[y \in \mathcal{C}(x) \mid A = a] - \mathbb{P}[y \in \mathcal{C}(x) \mid A = b] \right) \approx 0. \tag{9}$$

Marginal coverage gives no guarantee that Equation (9) will hold, but group-wise conditional coverage does. However, Equalized Coverage negatively impacts the usefulness of prediction sets for uncertainty quantification by increasing their average size, meaning the model expresses a greater level of uncertainty than it would using marginal CP (Romano et al., 2020b; Gibbs et al., 2025; Ding et al., 2024). Additionally, partitioning $\mathcal{D}_{\text{cal}}$ means each individual calibration is done with fewer datapoints n_{cal}. This increases variance, and the probability that the desired coverage level $1 - \alpha$ is breached in practice (Angelopoulos & Bates, 2021).

UQ $\rightharpoonup$ F: Classes $\mathcal{Y}$ in a decision problem often represent mutually exclusive actions, hence a prediction set cannot be acted on by itself. Instead, prediction sets can be given to a human decision maker as a form of model assistance (Straitouri & Gomez Rodriguez, 2024; Cresswell et al., 2024). The usefulness of prediction sets is correlated to set size – humans have higher accuracy on tasks when given smaller prediction sets (Cresswell et al., 2024). Average set sizes $\mathbb{E}|\mathcal{C}_q|$ typically vary across groups when the underlying model f_θ has some accuracy disparity $\Delta_{\text{acc}} > 0$ (Equation (1)). As a result, human accuracy will improve more for groups which have smaller sets on average, causing Disparate Impact (Cresswell et al., 2025).

Equalized Coverage makes this unfairness worse. If a group in the data is under-covered using marginal CP, equalizing its coverage requires increasing set sizes, harming downstream accuracy even more. Substantive fairness – achieving comparable accuracy across groups – would be better served by equalizing set sizes (Cresswell et al., 2025).

3.5 Privacy and Robustness

P $\rightharpoonup$ R: Models trained with DPSGD (Equation (3)) tend to be less adversarially robust than the same models trained without DP guarantees. The clipping and noise addition steps in DPSGD slow the convergence of models (Tramèr & Boneh, 2021) giving decision boundaries that are less smooth (Hayes et al., 2022) which has a strong impact on adversarial robustness (Fawzi et al., 2018). Empirical tests confirm this intuition (Boenisch et al., 2021; Tursynbek et al., 2021).

R $\rightharpoonup$ P: Adversarial training (Equation (5)), designed to improve adversarial robustness, can increase the influence of individual datapoints on the model. This in turn makes the model more susceptible to MIAs (Yeom et al., 2020). Song et al. (2019) tested six common adversarial defence methods and found all six increased the success rates of MIAs compared to the same model trained without any specific defence.

Incorporating DP alongside adversarial defences to protect against MIAs is also non-trivial, as the methodologies conflict in practice. Adversarial training creates several augmentations $x^\dagger$ of data points x, and backpropagates gradients over them in a batch. By comparison, DPSGD computes per-sample gradients, which is on its own computationally inefficient (Yousefpour et al., 2021). Incorporating augmented data points would drastically increase the time and memory costs of training, and require careful accounting of how much privacy budget is consumed by the use of augmented versions of x (Wu et al., 2024).

3.6 Privacy and Explainability

P $\rightharpoonup$ E: DPSGD is designed to obscure the details of any single element of $\mathcal{D}_{\text{train}}$, but its addition of noise to gradient updates can degrade the fidelity of *post hoc* explanations by clouding the true relationships between input and output variables (Patel et al., 2022). Saifullah et al. (2024) found severe deterioration of explanation fidelity across many model architectures and data domains when DPSGD was used.

Applying DPSGD during training protects elements of $\mathcal{D}_{\text{train}}$, but not inference data x_{test} whose predictions need to be explained. DP can be applied to the explanation mechanism to protect x_{test}

(Patel et al., 2022), but since DP requires randomization, explanations will necessarily differ each time they are generated for the same x_{test}. Unstable and potentially inconsistent explanations undermine the notion that we can understand the reasons behind model predictions.

E $\rightharpoonup$ P: Local explanations $E_\theta(x)$ are supplemental to the model's outputs $F_\theta(x)$, and as such pose an additional avenue for private information about x or $\mathcal{D}_{\text{train}}$ to leak from the model. Explanations are designed to reveal details about how specific inputs influence model predictions, so it is unsurprising that they can be exploited to make MIAs more effective (Shokri et al., 2021). For instance, the explanations generated by LIME (Section 2.4) consist of a simple model that locally approximates $F_\theta(x)$ around x. The behaviour of these local models will vary depending on whether x was included in $\mathcal{D}_{\text{train}}$, and attackers can exploit these differences in their MIAs (Quan et al., 2022; Huang et al., 2024).

3.7 Privacy and Uncertainty Quantification

P $\rightharpoonup$ UQ: Conformal Prediction could be applied to any model trained with DPSGD without affecting the privacy of $\mathcal{D}_{\text{train}}$, because CP is merely post-processing on the privatized model $f_\theta(x)$. However, CP requires an additional calibration set $\mathcal{D}_{\text{cal}}$ which would not inherit any DP guarantee and hence could be vulnerable to privacy attacks. To mitigate the risk to $\mathcal{D}_{\text{cal}}$, one can generate *private prediction sets* via a DP quantile routine (Angelopoulos et al., 2022), such that prediction sets $\mathcal{C}_q(x)$ would satisfy Equation (2) for $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}'_{\text{cal}}$ differing by one element. However, the noise added for DP degrades the empirical coverage of $\mathcal{C}_q(x)$, requiring larger prediction sets to retain the same coverage $1-\alpha$, hence overestimating the true model uncertainty.

Even without protecting the privacy of $\mathcal{D}_{\text{cal}}$, the quality of UQ with a differentially private model will suffer compared to a non-private model. The per-example clipping used in DPSGD causes miscalibration (Bu et al., 2023; Zhang et al., 2022), which affects the utility of CP, again by increasing the size of prediction sets (Xi et al., 2024).

UQ $\rightharpoonup$ P: Uncertainty quantification techniques by design provide additional information to supplement the model's prediction, which broadens the attack surface. As proof of concept, Zhu et al. (2024) developed and tested MIAs targeting prediction sets, showing empirically that an attacker who receives prediction sets has a higher rate of successfully identifying datapoints x that were used in $\mathcal{D}_{\text{train}}$.

3.8 Robustness and Explainability

R $\rightharpoonup$ E: Adversarial training fundamentally alters the representations that are learned by $F_\theta(x)$ (Tsipras et al., 2019; Zhang & Zhu, 2019). In image classification, features learned with adversarial training are often more interpretable to humans (Ilyas et al., 2019), but other data modalities do not share the same alignment between robust features and human-perceptible patterns (Jia & Liang, 2017; Carlini & Wagner, 2018). In such domains, adversarial training can lead to unexplainable behaviours and reduced explanation fidelity (Zhou et al., 2025).

E $\rightharpoonup$ R: *Post hoc* explanations are susceptible to adversarial perturbations which do not change the model's prediction $F_\theta(x)$, but greatly change the explanation $E_\theta(x)$ (Ghorbani et al., 2019). Users may expect there to be a single, interpretable explanation for any given prediction, and hence the possibility of non-robust explanations casts doubt on the veracity of all explanations. Alternatively, adversarial examples can be used to generate explanations from methods like LIME (Section 2.4) which are not faithful to the model's actual behaviour (Slack et al., 2020).

3.9 Robustness and Uncertainty Quantification

R $\rightharpoonup$ UQ: Conformal prediction sets may fail to be robust if the underlying model is non-robust. Standard CP methods use softmax scores $f_\theta(x)$ to compute the conformal score $s(x, y)$ (Romano et al., 2020b), so if the elements of $f_\theta(x^\dagger_{\text{test}})$ vary wildly, so will $\mathcal{C}_q(x^\dagger_{\text{test}})$. Hence, adversarial training on the underlying model may appear to be a natural defence for CP. However, Liu et al. (2024) demonstrated that adversarial training increases the overall uncertainty of models, leading to larger set sizes even for clean datapoints x_{test}.

UQ $\rightharpoonup$ R: CP techniques are highly susceptible to adversarial attacks because they introduce additional assumptions on the test data which can easily be violated by an attacker. The coverage guarantee (Equation (7)) relies on exchangeability of x_{test} with $\mathcal{D}_{\text{cal}}$, but adversarial perturbations

can imperceptibly force x_{test} out-of-distribution (Gendler et al., 2022). Prediction sets under attack will grossly overestimate the certainty of predictions and often fail to cover the true label. Alternatively, x_{test} can be perturbed such that coverage is maintained, but prediction set sizes are greatly increased, which reduces the utility of those sets (Ghosh et al., 2023).

3.10 Explainability and Uncertainty Quantification

E $\rightharpoonup$ UQ: When generating explanations of a model's predictions, not only should we quantify the uncertainty in the predictions, but also in the explanations themselves. Explanations help users determine when to rely on predictions, investigate potential issues, and confirm that predictions are not influenced by biases. However, high uncertainty about the validity of explanations can erode trust (Kindermans et al., 2019; Bykov et al., 2020; Ahn et al., 2023; Löfström et al., 2024). For example, Slack et al. (2021) highlighted that the feature importances $E_\theta(x)$ generated by LIME strongly depend on the random noise introduced when constructing the sparse linear model around x, and on the number of perturbed samples used. These factors can lead to significant variations in the rank order of important features, indicating a considerable degree of uncertainty in LIME's explanations that is often overlooked.

UQ $\rightharpoonup$ E: To help practitioners understand the limitations of a model, explanations should be given as to why it is more uncertain on some inputs than others (Antoran et al., 2021). For CP, explanations must extend to prediction sets $\mathcal{C}_q(x)$. However, the task of explaining why the model has predicted the entire set is inherently more difficult than explaining the top prediction $F_\theta(x)$, especially when some elements of the set may be contradictory or incompatible with others. Yapicioglu et al. (2024) recognize this as an issue and develop techniques to explain the relative impact different features have on the coverage and size of $\mathcal{C}_q(x)$.

4 Trustworthy AI Must Account for Intersectionality

The overall goal of Trustworthy AI research is to enable not one or two aspects of trust, but *many* simultaneously. Current research in TAI very commonly follows the same formula: one or two TAI aspects are selected and genuine issues with typical AI models are used to motivate improving these aspects. Then, technical solutions are developed and evaluated to show improvement on the selected aspects; possible interactions with aspects outside the ones selected are rarely considered. The most straightforward attempt to achieve the overall goal of trust would be to overlay several technical solutions. However, the examples from Section 3 demonstrate that negative interactions between TAI aspects are not rare, are sometimes unexpected, and may only be documented years after a method is first deployed. Based on these observations we take the position that combining solutions to individual TAI aspects will not resolve the trust and alignment problems facing AI. Trustworthiness is not achieved by overlaying isolated technical solutions, but emerges from integrating TAI aspects within a holistic framework that accounts for intersectionality.

Intersectional TAI considers all relevant aspects simultaneously, not in a sequential or siloed manner. It relies on interdisciplinary expertise from ethicists, legal experts, and of course computer scientists, to bring together knowledge from disparate fields. It is context-aware and adapted to its deployment domain, taking account of specific requirements - like the primacy of patient safety in healthcare. Trust in AI systems will be achieved not solely through technical solutions, but by aligning AI with societal needs, and recognizing real-world constraints.

To advance intersectional TAI, we provide guidance to researchers and practitioners on how to achieve it in practice, acknowledging the likelihood of negative interactions, trade-offs, and challenges from combining many objectives.

- Prior to model development, enumerate all relevant TAI aspects and prioritize them by importance in the application at hand. Involve stakeholders including developers, users, and subjects of the model.
- Establish clear metrics and develop automated tests for each relevant aspect when possible, but also recognize soft goals and constraints within the deployment context.
- Deliberately analyze how TAI aspects could interact, positively and negatively, before implementing technical solutions or optimizing for metrics.
- Evaluate the potential risks of negative interactions by quantifying their likelihood and severity.

- When applying technical solutions to improve any single aspect, perform ablations to measure impact on all other aspects, not just accuracy.
- When negative interactions or trade-offs are observed, assess the impacts to each aspect and manage compromises according to the pre-established priorities.

These steps will help to develop a risk-based prioritization of TAI aspects and balance competing constraints, enabling users to proactively anticipate, measure, and mitigate negative interactions.

4.1 Case Study: Financial Industry

We now provide a case study to demonstrate how a siloed approach to TAI might fail to establish trust. SiloBank is a very typical (but fictional) regulated financial institution (FI) that uses AI to automate decisions on credit card applications. Like many FIs, SiloBank manages model risk using "three lines of defence" (Bantleon et al., 2021), where the 1st line are the developers who build and operate models, the 2nd line provides oversight and independent challenge to the 1st line, while the 3rd line is an internal audit function that assesses the effectiveness of the 1st and 2nd lines.

Due to regulations, SiloBank must ensure that its models preserve data privacy, are fair across groups protected by law, can provide actionable explanations to customers, and are robust enough to withstand sudden shifts in the lending environment. To achieve these goals, SiloBank established specialized teams within the 2nd line to provide expert oversight on their subject matter areas. The Privacy team, of course, certifies that data privacy is protected at all times, Regulatory Compliance evaluates models for fairness, while Model Validation tests models for their robustness and ability to generate meaningful explanations. By hiring experts in each area, SiloBank's leadership feels confident that each aspect of TAI will be accounted for.

Excited by recent developments in ML, the 1st line has created a new credit card decision model that greatly outperforms what SiloBank has in place. Eager to put the model to use, the 1st line shows their work to each 2nd line team. Model Validation approves the model's robustness and explainability aspects, while Compliance confirms the model is unbiased. However, Privacy requests better protection against information leaks. Upon revision, the developers decide that retraining their model with DPSGD (Equation (3)) would provably protect customer's data, and Privacy is satisfied with the mathematically rigorous approach. With all 2nd line teams on board, the model is deployed.

Several months later, SiloBank finds itself in the headlines. A married couple who share finances both applied for the same credit card, but became frustrated when only one of them was approved. More stories begin to surface of denied applications from financially secure women, and customers leave the bank. Internally, SiloBank's 3rd line audit team begins investigating and finds the new credit card model heavily disadvantages women, despite Compliance's earlier findings to the contrary.

While each 2nd line team was composed of experts in their field, no team effectively managed risks across jurisdictions. The 1st line failed to account for the negative interactions that DPSGD can cause, and did not build in automated tests that would run after each model change. In the end, Audit recommends that leadership break down siloed divisions and instead establish intersectional teams that account for the interactions between TAI aspects.

4.2 Beyond Pairwise Interactions

In Section 3 we surveyed a wide range of literature that has considered specific pairwise interactions. We briefly mention work that has gone beyond pairwise interactions and considered the intersection of multiple TAI aspects in the direction we are advocating. Ferry et al. (2023) systematically review three aspects, fairness, privacy, and explanability, pointing out the isolated nature of prior research. They survey the literature on pairwise interactions considering all three combinations, but stop short of considering novel challenges of integrating all three aspects at once. Sharma et al. (2020) integrate fairness and robustness into their explainability method by design in the spirit of intersectional TAI, but do not propose a holistic framework that goes beyond these three aspects. Meanwhile, Li et al. (2023) discuss many TAI aspects including fairness, privacy, robustness, and explainability, building a framework of when to consider each aspect throughout the model lifecycle. While advocating for the combination of many TAI aspects, they do not give detailed insights on the negative interactions that can occur.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016. doi: 10.1145/2976749.2978318.

Surin Ahn, Justin Grana, Yafet Tamene, and Kristian Holsheimer. Uncertainty quantification for local model explanations without model access. *arXiv:2301.05761*, 2023.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv:1606.06565*, 2016.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv:2107.07511*, 2021.

Anastasios Nikolas Angelopoulos, Stephen Bates, Tijana Zrnic, and Michael I. Jordan. Private Prediction Sets. *Harvard Data Science Review*, 4(2), 2022.

Javier Antoran, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato. Getting a CLUE: A Method for Explaining Uncertainty Estimates. In *International Conference on Learning Representations*, 2021.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, volume 32, pp. 15479–15488, 2019.

Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. The road to explainability is paved with bias: Measuring the fairness of explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1194–1206. Association for Computing Machinery, 2022. ISBN 9781450393522. doi: 10.1145/3531146.3533179.

Ulrich Bantleon, Anne d'Arcy, Marc Eulerich, Anja Hucke, Burkhard Pedell, and Nicole V.S. Ratzinger-Sakel. Coordination challenges in implementing the three lines of defense model. *International Journal of Auditing*, 25(1):59–74, 2021.

Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In *NeurIPS 2020 Workshop on Preregistration in Machine Learning*, volume 148, pp. 325–342, 2021.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases: European Conference*, pp. 387–402, 2013.

Franziska Boenisch, Philip Sperl, and Konstantin Böttinger. Gradient masking and the underestimated robustness threats of differential privacy in deep learning. *arXiv:2105.07985*, 2021.

Zhiqi Bu, Hua Wang, Zongyu Dai, and Qi Long. On the convergence and calibration of deep learning with differential privacy. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018.

Kirill Bykov, Marina M-C Höhne, Klaus-Robert Müller, Shinichi Nakajima, and Marius Kloft. How Much Can I Trust You?–Quantifying Uncertainties in Explaining Neural Networks. *arXiv:2006.09000*, 2020.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops*, pp. 1–7, 2018. doi: 10.1109/SPW.2018.00009.

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The privacy onion effect: Memorization is relative. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13263–13276, 2022.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56 (7), 2024. ISSN 0360-0300. doi: 10.1145/3616865.

Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy*, pp. 292–303, 2021. doi: 10.1109/EuroSP51992.2021.00028.

Hongyan Chang, Ta Duy Nguyen, Sasi Kumar Murakonda, Ehsan Kazemi, and Reza Shokri. On adversarial bias and the robustness of fair machine learning. *arXiv:2006.08669*, 2020.

A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 864–876, 2022. ISBN 9781450393522. doi: 10.1145/3531146.3533150.

Kimberle Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(8), 1989.

Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal Prediction Sets Improve Human Decision Making. In *Proceedings of the 41th International Conference on Machine Learning*, 2024.

Jesse C. Cresswell, Bhargava Kumar, Yi Sui, and Mouloud Belbahri. Conformal Prediction Sets Can Cause Disparate Impact. In *International Conference on Learning Representations*, 2025.

Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H. Bach, and Himabindu Lakkaraju. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 203–214, 2022. ISBN 9781450392471. doi: 10.1145/3514094.3534159.

Federico Di Gennaro, Thibault Laugel, Vincent Grari, Xavier Renard, and Marcin Detyniecki. Post-processing fairness with minimal changes. *arXiv:2408.15096*, 2024.

Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. *Advances in Neural Information Processing Systems*, 36, 2024.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, pp. 265–284. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-32732-5.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012. ISBN 9781450311151. doi: 10.1145/2090236.2090255.

Maria S. Esipova, Atiyeh Ashari Ghomi, Yaqiao Luo, and Jesse C. Cresswell. Disparate impact in differential privacy from gradient misalignment. In *The Eleventh International Conference on Learning Representations*, 2023.

Rosanna Fanni, Valerie Eveline Steinkogler, Giulia Zampedri, and Jo Pierson. Enhancing human agency through redress in artificial intelligence systems. *AI & Society*, 38(2):537–547, 2023.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

Sina Fazelpour and Maria De-Arteaga. Diversity in sociotechnical machine learning systems. *Big Data & Society*, 9(1):20539517221082027, 2022.

Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. SoK: Taming the Triangle–On the Interplays between Fairness, Interpretability and Privacy in Machine Learning. *arXiv:2312.16191*, 2023.

Iason Gabriel. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3):411–437, 2020.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2022.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688, Jul. 2019. doi: 10.1609/aaai.v33i01.33013681.

Subhankar Ghosh, Yuanjie Shi, Taha Belkhouja, Yan Yan, Jana Doppa, and Brian Jones. Probabilistically robust conformal prediction. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216, pp. 681–690, 2023.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkaf008.

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

Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems*, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Jamie Hayes, Borja Balle, and M Pawan Kumar. Learning to be adversarially robust and differentially private. *arXiv:2201.02265*, 2022.

Dan Hendrycks. *Introduction to AI Safety, Ethics and Society*. Taylor & Francis, 2024. ISBN 9781032798028.

Catherine Huang, Martin Pawelczyk, and Himabindu Lakkaraju. Explaining the model, protecting your data: Revealing and mitigating the data privacy risks of post-hoc model explanations via membership inference. *arXiv:2407.17663*, 2024.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable artificial intelligence approaches: A survey. *arXiv:2101.09429*, 2021.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017. doi: 10.18653/v1/D17-1215.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. *The (Un)reliability of Saliency Methods*, pp. 267–280. Springer International Publishing, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_14.

Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. Disparate vulnerability to membership inference attacks. *Proceedings on Privacy Enhancing Technologies Symposium*, 2022. doi: 10.2478/popets-2022-0023.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Himabindu Lakkaraju and Osbert Bastani. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 79–85, 2020. ISBN 9781450371100. doi: 10.1145/3375627.3375833.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.*, 55(9), 2023. ISSN 0360-0300. doi: 10.1145/3555803.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1), 2021. ISSN 1099-4300. doi: 10.3390/e23010018.

Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surv.*, 54(2), 2021. ISSN 0360-0300. doi: 10.1145/3436755.

Ziquan Liu, Yufei Cui, Yan Yan, Yi Xu, Xiangyang Ji, Xue Liu, and Antoni B. Chan. The pitfalls and promise of conformal inference under adversarial attacks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 30908–30928, 2024.

Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen. A pragmatic approach to membership inferences on machine learning models. In *2020 IEEE European Symposium on Security and Privacy*, pp. 521–534, 2020. doi: 10.1109/EuroSP48549.2020.00040.

Helena Löfström, Tuwe Löfström, Ulf Johansson, and Cecilia Sönströd. Calibrated explanations: With uncertainty information and counterfactuals. *Expert Systems with Applications*, 246:123154, 2024. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.123154.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi Differential Privacy of the Sampled Gaussian Mechanism. *arXiv:1908.10530*, 2019.

Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 466–477, 2021. ISBN 9781450383097.

Office of the Comptroller of the Currency. Fair lending, 2025. URL `https://www.occ.treas.gov/topics/consumers-and-communities/consumer-protection/fair-lending/index-fair-lending.html`. Accessed: 2025-01-01.

Neel Patel, Reza Shokri, and Yair Zick. Model explanations with differential privacy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1895–1904. Association for Computing Machinery, 2022. ISBN 9781450393522. doi: 10.1145/3531146.3533235.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

Pengrui Quan, Supriyo Chakraborty, Jeya Vikranth Jeyakumar, and Mani Srivastava. On the amplification of security and privacy risks by post-hoc explanations in machine learning models. *arXiv:2206.14004*, 2022.

Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang. Membership inference attack against differentially private deep learning model. *Transactions on Data Privacy*, 11(1), 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016. ISBN 9781450342322. doi: 10.1145/2939672.2939778.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2 (2), 2020a.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, 2020b.

Stuart Russell. *Human Compatible: AI and the Problem of Control*. Penguin UK, 2019.

Saifullah Saifullah, Dominique Mercier, Adriano Lucieri, Andreas Dengel, and Sheraz Ahmed. The privacy-explainability trade-off: Unraveling the impacts of differential privacy and federated learning on attribution methods. *Frontiers in Artificial Intelligence*, 7:1236947, 2024.

Shubham Sharma, Jette Henderson, and Joydeep Ghosh. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 166–172, 2020. ISBN 9781450371100. doi: 10.1145/3375627.3375812.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy*, pp. 3–18, 2017. doi: 10.1109/SP.2017.41.

Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 231–241, 2021. ISBN 9781450384735. doi: 10.1145/3461702.3462533.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020. ISBN 9781450371100. doi: 10.1145/3375627.3375830.

Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9391–9404, 2021.

Christian Soize. *Uncertainty Quantification*. Springer, 2017.

Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 241–257, 2019. ISBN 9781450367479. doi: 10.1145/3319535.3354211.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, 2024.

Eleni Straitouri and Manuel Gomez Rodriguez. Designing decision support systems using counterfactual prediction sets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, 2024.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv:1312.6199*, 2013.

Huan Tian, Guangsheng Zhang, Bo Liu, Tianqing Zhu, Ming Ding, and Wanlei Zhou. When fairness meets privacy: Exploring privacy threats in fair binary classifiers via membership inference attacks. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024. doi: 10.24963/ijcai.2024/57.

Shakila Mahjabin Tonni, Dinusha Vatsalan, Farhad Farokhi, Dali Kaafar, Zhigang Lu, and Gioacchino Tangari. Data and model dependencies of membership inference attack. *arXiv:2002.06856*, 2020.

Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations*, 2021.

Cuong Tran, Ferdinando Fioretto, Pascal Van Hentenryck, and Zhiyan Yao. Decision making with differential privacy under a fairness lens. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021. doi: 10.24963/ijcai.2021/78.

Cuong Tran, Keyu Zhu, Pascal Van Hentenryck, and Ferdinando Fioretto. On the effects of fairness to adversarial vulnerability. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024. doi: 10.24963/ijcai.2024/58.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Nurislam Tursynbek, Aleksandr Petiushko, and Ivan Oseledets. Robustness threats of differential privacy. *arXiv:2012.07828*, 2021.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. *Technical Report*, 2003.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Trans. Knowl. Discov. Data*, 17(3), 2023. ISSN 1556-4681. doi: 10.1145/3551390.

Jiapeng Wu, Atiyeh Ashari Ghomi, David Glukhov, Jesse C. Cresswell, Franziska Boenisch, and Nicolas Papernot. Augment then smooth: Reconciling differential privacy with certified robustness. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Huajun Xi, Jianguo Huang, Lei Feng, and Hongxin Wei. Does confidence calibration help conformal prediction? *arXiv:2402.04344*, 2024.

Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021.

Fatima Rabia Yapicioglu, Alessandra Stramiglio, and Fabio Vitali. ConformaSight: Conformal Prediction-Based Global and Model-Agnostic Explainability Framework. In *Explainable Artificial Intelligence*, pp. 270–293, 2024. ISBN 978-3-031-63800-8.

Dayong Ye, Sheng Shen, Tianqing Zhu, Bo Liu, and Wanlei Zhou. One parameter defense—defending against data inference attacks via differential privacy. *IEEE Transactions on Information Forensics and Security*, 17:1466–1480, 2022a. doi: 10.1109/TIFS.2022.3163591.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3093–3106, 2022b. ISBN 9781450394505. doi: 10.1145/3548606.3560675.

Samuel Yeom, Irene Giacomelli, Alan Menaged, Matt Fredrikson, and Somesh Jha. Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning. *Journal of Computer Security*, 28(1):35–70, 2020.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv:2109.12298*, 2021.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, 2013.

Hanlin Zhang, Xuechen Li, Prithviraj Sen, Salim Roukos, and Tatsunori Hashimoto. A closer look at the calibration of differentially private learners. *arXiv:2210.08248*, 2022.

Tianyuan Zhang and Zhanxing Zhu. Interpreting adversarially trained convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, 2019.

Dehua Zhou, Ziyu Song, Zicong Chen, Xianting Huang, Congming Ji, Saru Kumari, Chien-Ming Chen, and Sachin Kumar. Advancing explainability of adversarial trained convolutional neural networks for robust engineering applications. *Engineering Applications of Artificial Intelligence*, 140:109681, 2025. ISSN 0952-1976. doi: 10.1016/j.engappai.2024.109681.

Meiyi Zhu, Caili Guo, Chunyan Feng, and Osvaldo Simeone. Uncertainty, calibration, and membership inference attacks: An information-theoretic perspective. *arXiv:2402.10686*, 2024.