The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

I Prefer Not To Say: Protecting User Consent in Models with Optional Personal Data

Tobias Leemann[1, 2], **Martin Pawelczyk**[3], **Christian Thomas Eberle**[1], **Gjergji Kasneci**[2]

[1]University of Tübingen, Tübingen, Germany
[2]Technical University of Munich, Munich, Germany
[3]Harvard University, Cambridge, MA, USA
tobias.leemann@uni-tuebingen.de, martin.pawelczyk.1@gmail.com, ct.eberle@protonmail.ch, gjergji.kasneci@tum.de

Abstract

We examine machine learning models in a setup where individuals have the choice to share optional personal information with a decision-making system, as seen in modern insurance pricing models. Some users consent to their data being used whereas others object and keep their data undisclosed. In this work, we show that the decision not to share data can be considered as information in itself that should be protected to respect users' privacy. This observation raises the overlooked problem of how to ensure that users who protect their personal data do not suffer any disadvantages as a result. To address this problem, we formalize protection requirements for models which only use the information for which active user consent was obtained. This excludes implicit information contained in the decision to share data or not. We offer the first solution to this problem by proposing the notion of Protected User Consent (PUC), which we prove to be loss-optimal under our protection requirement. We observe that privacy and performance are not fundamentally at odds with each other and that it is possible for a decision maker to benefit from additional data while respecting users' consent. To learn PUC-compliant models, we devise a model-agnostic data augmentation strategy with finite sample convergence guarantees. Finally, we analyze the implications of PUC on challenging real datasets, tasks, and models.

Introduction

While the day-to-day impact of automated data processing is steadily growing, modern regulations such as the European Union's General Data Protection Regulation (GDPR) (GDPR 2016) or the California Consumer Privacy Act (CCPA) (OAG 2021) strive to give individuals more control over their personal data. In light of these regulations, we consider machine-learned classifiers in which individuals have the freedom to decide themselves on which data they would like to provide to an automated decision system.

Such systems are increasingly being deployed (Henning 2022): As a running example, we consider a realistic use-case of health insurance pricing: Suppose in an automated pricing model all potential customers are asked to fill out an application form where they enter certain *base features*, for instance information such as their state of residence and age. To improve the pricing model, the insurance offers an additional service, a "companion fitness app", through which additional health data about the customer's physical condition are collected. The customers decide whether to use the app or not; alternatively, customers can sign up for a policy without consenting to use the app. The health data that customers share may however influence the premium of the insurance policy they receive. We refer to data that provide additional, non-mandatory information beyond the base features as *optional features*. With fitness trackers and smartwatches rapidly gaining popularity (Reeder and David 2016; Zimmer et al. 2020; Statista 2023), such systems are already being deployed in practice, e.g., by major health insurance firms in Australia (Henning 2022).

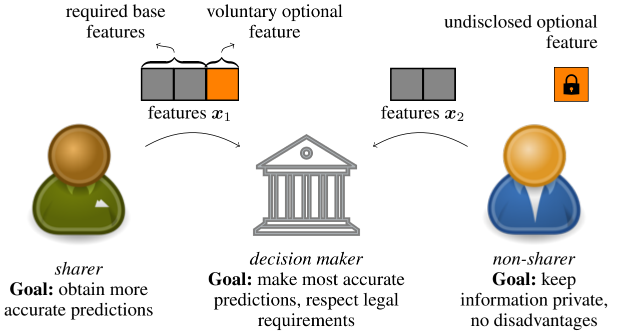
required base features
voluntary optional feature
undisclosed optional feature
features $\boldsymbol{x}_1$
features $\boldsymbol{x}_2$
sharer
Goal: obtain more accurate predictions
decision maker
Goal: make most accurate predictions, respect legal requirements
non-sharer
Goal: keep information private, no disadvantages

Figure 1: Overview of the relevant stakeholders. We consider a case where users can voluntarily provide information on optional features or choose to leave them undisclosed. The goals of sharers, non-sharers, and the decision maker have to be reconciled.

The outlined scenario is challenging as there are three groups of stakeholders whose interests need to be reconciled: (1) The group of non-sharing individuals who do not want to provide additional information, for instance due to privacy concerns. We refer to them as *non-sharers*. For this group, the decision maker does not want to or cannot force them to provide the additional information for legal reasons. Consequently, the non-sharers do not want the additional information to be considered in the decision making process; in return, they are willing to sacrifice some accuracy, but they do not want to face other systematic disadvantages. (2)

On the other hand, individuals who voluntarily share data (*sharers*) explicitly want the additional information to be considered and want to obtain more accurate predictions. (3) Finally, the decision makers themselves desire the most accurate predictions with the lowest overall costs while respecting the users' privacy and legal requirements.

Among these requirements, it is crucial to the non-sharers to explicitly exclude the information contained in the decision to share or not to share. To see this, we note that smartwatch users are more likely to exercise in general than non-wearers (DeMarco 2023) which usually create lower costs for the insurance company as fitter customers take less sick days on average. Thus, only through observing the decision to share data, the insurance firm could make inferences about a person's fitness. This is problematic for two reasons: First, the company would unethically infer private data, that the non-sharers explicitly did not give consent to. Prior work (Wachter and Mittelstadt 2019) has argued for a "right to reasonable inferences". This rules out inferences from unrelated factors that are purely predictive and may infringe privacy, as they open the door for discriminatory and invasive decision-making (Mittelstadt et al. 2016). Second, this would lead to non-sharers being assigned a higher insurance premium than the estimate of the legacy model which only considered their base features. Many countries have laws that prohibit insurers from raising the base premium for users who do not share their data, as this is seen as a coercive and unfair practice. For example, the US only permits five factors to affect the premium, which are location, age, tobacco use, plan category, and dependent coverage (US Government. U.S. Centers for Medicare & Medicaid Services. 2023). It is however possible for insurers – and desired by many users – to award bonuses which reduce the premium based on participation in optional reward and incentive programs (Madison, Schmidt, and Volpp 2013; Henning 2022).

To summarize, we study machine learning models that can handle optional features and meet legal requirements and desiderata of three groups of stakeholders: the sharers, the non-sharers, and the decision makers. We consider it essential for these models to not make inferences based on the unavailability of a feature value for the non-sharers, a constraint that we term *Availability Inference Restriction (AIR)*. Finally, we are interested in obtaining models with optimal performance under this requirement.

Contribution. We address the problem of how to fairly and privately predict outcomes for users who share optional data and those who do not. We tackle this overlooked issue by making the following contributions:

- **Definition.** We introduce models with Protected User Consent (PUC), which are optimal under our protection requirement AIR. We derive performance guarantees, which formally show that it is possible to reconcile the decision maker's interest in improved predictions and the non-sharer's privacy preferences.
- **Algorithm.** We propose a PUC-inducing data augmentation (`PUCIDA`) technique that can be applied to any type of predictive architecture (e.g., tree or neural network) and any convex loss function (e.g., mean squared error or cross-entropy loss) to obtain such models
- **Analysis.** We prove that predictive models trained with `PUCIDA` satisfy PUC asymptotically, and provide finite sample convergence results that demonstrate that `PUCIDA` produces PUC-compliant models in practice.
- **Empirical evaluation.** We empirically show that without enforcing PUC, the average absolute prediction outcome (e.g., insurance quote) of users who do not share data can be almost 20 % worse than justified by their base data. We then evaluate our data augmentation technique on various ML models and show that PUC is achieved regardless of the model.

Related Work

In this Section, we review the most relevant streams of related work (see Appendix A.1 for additional references).

Classification with Missing Values. Classification models that can handle missing data have been studied previously with the goal of minimizing costs or increasing performance (Zhang et al. 2005; Aleryani, Wang, and De La Iglesia 2020), obtaining uncertainty estimates (Kachuee et al. 2020), or fulfilling classical fairness notions (Zhang and Long 2021; Jeong, Wang, and Calmon 2022; Wang and Singh 2021; Fernando et al. 2021). However, the mechanisms underlying missingness is different in this work, as missing values indicate explicit non-consent by the user, leading to different implications. In a related line of work, classification with noisy (Fogliato, Chouldechova, and G'Sell 2020) or missing labels (Kilbertus et al. 2020; Rateike et al. 2022) has been investigated, where the missingness is often a result of *selection bias*. The setting considered in this work is different in the sense that we are not concerned with fulfilling a fairness notion with respect to a sensitive attribute, but consider the interests of subjects that have and have not provided optional information.

Data Minimization. The principle of Data Minimization is anchored in the GDPR (GDPR 2016). Data Minimization demands minimal data collection. Several works are concerned with implementing (Goldsteen et al. 2021) or auditing compliance with this principle (Rastegarpanah, Gummadi, and Crovella 2021). Rastegarpanah et al. (Rastegarpanah, Crovella, and Gummadi 2020) consider decision systems that can handle optional features from a data minimization perspective where the decision maker decides which features are collected for each individual. This principle is distinct from the "right to be forgotten" (Biega et al. 2020), which enables individuals to submit requests to have their data deleted. In response to these regulations, several works consider the problem of updating an ML model without the need of retraining the entire model (Wu, Dobriban, and Davidson 2020; Ginart et al. 2019; Izzo et al. 2021; Golatkar, Achille, and Soatto 2020) or the effect of removals on model explanations (Rong et al. 2022; Pawelczyk et al. 2023). Our work differs from these works as our goal is to train a model where users decide themselves which data they deem relevant through sharing one or many optional features.

Algorithmic Fairness. A multitude of formal fairness definitions have been put forward in the literature (Verma and Rubin 2018). Examples include statistical parity (Dwork et al. 2012), predictive parity (Chouldechova 2017), equalized odds, equality of opportunity (Hardt, Price, and Srebro 2016), and individual fairness (Dwork et al. 2012). However, they are still a topic of discussion, for instance, because these definitions are known to be incompatible (Kleinberg, Mullainathan, and Raghavan 2016; Lipton, McAuley, and Chouldechova 2018). Additionally, there are a several definitions that rely on causal mechanisms to assess fairness, e.g., counterfactual fairness (Kusner et al. 2017), and the notion of unresolved discrimination (Kilbertus et al. 2017). While causal approaches to fairness might be preferable, they require information about the causal structure of the data generating process. Moreover, it has recently been shown that causal definitions may lead to adverse consequences, such as lower diversity (Nilforoshan et al. 2022). We discuss how existing fairness definitions could possibly be applied to the setting with optional features, but we find that none of the fairness definitions aligns with our desiderata theoretically and experimentally (see Appendix A.2).

Strategic Classification. In an even broader context, this work also relates to the field of strategic classification (Hardt et al. 2016). However, it is worth noting that in strategic classification research, the focus primarily revolves around users strategically manipulating their features for optimal outcomes, which may also involve information withholding (Krishnaswamy et al. 2021). In contrast to our work, privacy concerns are neglected in this research stream. As far as we are aware, there are no prior works on the specific problem of balancing the interests of *all three* groups of stakeholders (the non-sharers, sharers, and the decision makers).

Problem Formulation

Formalization and Notation

In this work, each data instance contains a realization of a number of base features $\mathbf{b} \in \mathcal{X}^b$, where $\mathcal{X}^b \subseteq \mathbb{R}^n$ is the space of the base features. Furthermore, let there be some optional information $z \in \mathcal{X}^z$, where $\mathcal{X}^z \subseteq \mathbb{R}$ is the value space of the optional feature.[1] It is the users' choice to decide if they want to disclose z to the system, which results in an availability variable $a \in \{0, 1\}$. Accordingly, only imputed samples $z^* = \{z \text{ if } a{=}1, \text{ else } \texttt{N/A}\}$ are observed, where a value of N/A indicates that a user did not reveal the optional information, e.g., did not use the companion app. In summary, the data observations are tuples $\mathbf{x} = (\mathbf{b}, a, z^*)$ that reside in $\mathcal{X} = \mathcal{X}^b \times \{0, 1\} \times (\mathcal{X}^z \cup \{\texttt{N/A}\})$. Each training sample comes with a label $y \in \mathcal{Y}$. Further, there is a data generating distribution $\mathbf{p}$ with support $\mathcal{X} \times \mathcal{Y}$ and we have access to an i.i.d. training sample $(\mathbf{x}, y) \sim \mathbf{p}$. Figure 2 shows such a data sample. We denote the random variables for the respective quantities by $\boldsymbol{B}, A, Z, Z^*, Y$. The label is probabilistically determined through the base features $\mathbf{B}$ and the hidden feature Z but the sharing decision does not influence the true label for a given $\mathbf{B}, Z$, such that $Y \perp\!\!\!\perp A | \mathbf{B}, Z$.

base features $\mathbf{b}$		opt. feat. z^*	a	label y
state	plan	fitness score	avail.	treatment costs
New South Wales	basic	87 %	1	3k$
Queensland	gold	N/A	0	17k$
New South Wales	basic	92 %	1	5k$
New South Wales	basic	N/A	0	64k$
Victoria	premium	56 %	1	22k$

Figure 2: Samples for the insurance use-case. We have two base features $\mathbf{b}$ and one optional feature z^*, which either takes an observed value z, or it takes a value of N/A if unobserved. The variable $a \in \{0, 1\}$ indicates the availability of the feature. The goal is to predict the label y.

In many applications, the goal is to find a function $f : \mathcal{X} \to \mathcal{Y}$ that models the observed data. In particular, $f : \mathcal{X} \to [0, 1]$ may predict a probability of a positive outcome or $f : \mathcal{X} \to \mathbb{R}$ may return a numerical score. The test data for which the model will be used come from the same distribution $\mathbf{p}$, though with the label y unobserved, and we suppose that the information provided is always correct. We consider a convex loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, e.g., mean-squared-error (MSE) or binary cross entropy (BCE), for which we minimize the expected loss for a sample from the data distribution. For instance, using the common MSE loss $\mathcal{L}(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$, an optimal predictor is given by $f^*_{\mathcal{L}}(\mathbf{x}) = \arg\min_{f(\mathbf{x})} \mathbb{E}_{\mathbf{p}(Y|\mathbf{x})} \left[(f(\mathbf{x}) - Y)^2\right] = \mathbb{E}[Y|\mathbf{x}]$, the conditional expectation. However, this notion can be generalized to other loss functions: An optimal predictor $f^*_{\mathcal{L}}(\mathbf{x})$ for the loss function $\mathcal{L}$ fulfills $\forall \mathbf{x}$:

$$f^*_{\mathcal{L}}(\mathbf{x}) = \mathbb{F}^{\mathcal{L}}_{\mathbf{p}}[Y|\mathbf{x}] := \arg\min_{f(\mathbf{x})} \mathbb{E}_{\mathbf{p}(Y|\mathbf{x})}\left[\mathcal{L}(f(\mathbf{x}), Y)\right]. \tag{1}$$

We use $\mathbb{F}^{\mathcal{L}}[Y|\mathbf{x}]$ to denote a generalized expected value that minimizes the expected loss conditioned on $\mathbf{x}$. To ease our derivations, we suppose this minimum to be unique and finite. Intuitively, it represents the best guess of Y given $\mathbf{x}$. For the MSE-Loss, $\mathbb{F}^{\mathcal{L}}$ is equivalent to the expectation operator $\mathbb{E}$. In the following statements, the reader may thus mentally replace $\mathbb{F}^{\mathcal{L}}$ with an expectation $\mathbb{E}$ without further ramifications in order to get the high level intuition. Finally, we introduce two key terms, namely, *base feature model* and *full feature model*. The former refers to a model trained on the base features only, while the latter refers to a model trained on all features where some strategy is used to replace unavailable feature values. Typically these strategies are called *imputation* and replace unavailable values by zeros, a feature's mean or median (Emmanuel et al. 2021).

Desiderata

Our goal is to learn models $f : \mathcal{X} \to \mathcal{Y}$ that comply with the desideratum of *Availability Inference Restriction*, which we briefly introduced in Section , to protect the interests of the non-sharers. Under this constraint, the model should provide the best predictive performance to reflect the need of

[1] We extend our definitions to integrate multiple optional features a later section.

the sharers and the decision maker for most accurate predictions.

Desideratum 1: Availability Inference Restriction. We start by considering the intricate case of individuals who *do not want to share optional information*. In this case, the model should compute the prediction based on the information the user gave their consent to. In particular, (a) the model should only use the base features *and* (b) should not use information that could be derived from the unavailability of the optional features to compute the prediction to avoid violating the user's consent.

For (a), this requires that the predictor does not use the information as an explicit input, i.e., the predictor should behave as if it only used base features $\mathbf{b}$ via some function $g : \mathcal{X}^b \to \mathcal{Y} : f_{|a=0}(\mathbf{b}, a, z^*) = g(\mathbf{b})$. For (b), although a=0 is not an explicit input to g, a sufficiently complex function may still be implicitly adapting to the group $a = 0$ and thus incorporate information that the user did not give their consent to. We would like to make sure that the predictions of g cannot use more information than contained in the overall conditional distribution, given the base features $\mathbf{b}$. This overadaption can be prevented by constraining the model's loss on the population of non-sharers to match the loss of the optimal base model $f^*_{\mathcal{L}}$ on this population. The reasoning behind this rationale is that all models that would beat the performance of this model must implicitly use some additional side knowledge about this group that was not provided by the users.

Definition 1 (Availability Inference Restriction). *For individuals that choose not to provide the optional feature (a=0), only the provided data* $\mathbf{b}$ *is used to compute the outcome in the decision process, i.e.,* $f_{|a=0}(\mathbf{b}, a, z^*)$=$g(\mathbf{b})$, *where* $g : \mathcal{X}^b \to \mathcal{Y}$ *is a base feature model. Further, we require*

$$\mathbb{E}\left[\mathcal{L}\left(g(\mathbf{B}), Y\right)|A=0\right] \geq \mathbb{E}\left[\mathcal{L}\left(f^*_{\mathcal{L}}(\mathbf{B}), Y\right)|A=0\right]. \quad (2)$$

This definition summarizes our intuition that the information encoded through the unavailability of feature information should neither be used explicitly (a) nor implicitly (b). We show how this constraint can analogously be derived from information-theoretic considerations in Appendix B.3.

Desideratum 2: Optimality. Our Definition 1 restricts the information that the predictor can use when the optional information is unavailable. To meet the interests of the decision maker and the sharers, we also want to find models with optimal performance, i.e., lowest loss, under this constraint.

Protecting User Consent

We are therefore looking for an *optimal* model within the class of predictors that comply with Availability Inference Restriction. In this Section, we derive a novel notion called Protected User Consent (PUC) that fulfills this purpose.

One-Dimensional PUC

The next result encodes an intuitive notion of protection for the users that do not want to share data on the optional features (a=0): Their prediction under f is then constrained to the best estimate for a user with the same base characteristics, no matter if additional data was provided. Contrarily, when additional information through the optional feature is provided, the predictor returns the best estimate using the available optional information:

Theorem 1 (1D-PUC). *Let* $f : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}$ *be a full feature model (i.e., including optional features). Among all predictors compatible with the Availability Inference Restriction, a model* f *with minimal loss is given by:*

$$f^*_{PUC}(\mathbf{b}, a, z^*) = \begin{cases} \mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}], & \text{if } a=0 \\ \mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}, A=1, Z^*=z^*] & \text{if } a=1. \end{cases}$$

We defer all proofs in this work to Appendix D. PUC is different from existing notions of group fairness, that do not fulfill the two desiderata in general (see Appendix A.2 for a discussion). Under the mentioned requirements, there is no model that can outperform f^*_{PUC}. We stress that 1D-PUC-compliant models have performance guarantees. These models match or improve upon an optimal base feature model $f^*_{\mathcal{L}}(\mathbf{B}) = \mathbb{F}_{\mathcal{L}}\left[Y|\mathbf{b}\right]$. This model can be seen as an upper bound for practical models obtained after model selection. Therefore, models that can beat its performance may offer improvements even after extensive hyper-parameter tuning and model selection, a property which we refer to as Predictive Non-Degradation (PND): a model f fulfills PND if its loss is smaller than that of the base feature model:

$$\mathbb{E}[\mathcal{L}\left(Y, f^*_{\mathcal{L}}(\mathbf{B})\right)] \geq \mathbb{E}[\mathcal{L}\left(Y, f(\mathbf{B}, A, Z^*)\right)]. \quad (3)$$

We prove the following result:

Corollary 1 (Predictive Non-Degradation of f^*_{PUC}). *For any density* $\mathbf{p}$, *a PUC-compliant model* f^*_{PUC} *fulfills Predictive Non-Degradation, i.e., it has a loss upper-bounded by the optimal base feature model* $f^*_{\mathcal{L}}$.

This is a remarkable result as it testifies that the decision maker can benefit from additional information in terms of loss, while protecting the privacy of users. This highlights that the interests of the different stakeholders are not contradictory and models that benefit all stakeholders do exist.

PUC under Strategic Considerations and Monotonicity Constraints

We have initially considered the case where the users desire the highest possible accuracy under data usage restrictions. However, in some cases such as our initial insurance example, the motivation to receive a lower premium might be a more important concern to some users than receiving an accurate prediction or their privacy concerns. If all users have full information (i.e., they see premiums with and without their optional data) and act strategically by sharing the value of z only if it would decrease their premiums, we obtain the following result.

Theorem 2 (Optimality of f^*_{PUC} under strategic actions). *Let* $\mathbf{p}'(\mathbf{B}, Z, Y)$ *be any prior density on base features, true optional features and labels and let* $f(\mathbf{b}, a=0, z) = \mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}]$, *i.e., the decision maker uses the base feature model when no optional data is available. Further suppose that users* *strategically choose to share the optional feature z only if $f(\mathbf{b}, a = 1, z) \leq f(\mathbf{b}, a = 0, \texttt{N/A})$. Under these conditions, the model f^*_{PUC} (Theorem 1) has minimal loss among all predictors.*

This result underlines that PUC models remain optimal if the decision maker cannot increase the premiums beyond the predictions of the current base model for the non-sharers. This is reasonable in many cases, where legal constraints mandate that the decision maker cannot implicitly force users to share data by inflating the base premium, as outlined in the introduction. The sharing decision can also be automated for the users by simply dropping the optional feature if it does not lead to a decrease in premiums. This would result in the aforementioned bonus systems, where sharing more data cannot increase the premium. We show that among the class of models with such a monotonicity constraint, the outlined PUC-model with automatic sharing decisions is still optimal under the same conditions as in Theorem 2 in Appendix D.5.

r-dimensional PUC

Next, we generalize our notion such that r features can be provided optionally. For example, the insurance firm might also accept voluntary results from prior medical examinations or diagnostic tests. Therefore, let there now be r optional features such that $\boldsymbol{z} \in \mathcal{X}^z_1 \times \cdots \times \mathcal{X}^z_r$ and $\boldsymbol{a} \in \{0, 1\}^r$, where $\mathcal{X}^z_i$ are the respective supports of each optional feature. By $\mathcal{I} \subseteq [r] = \{1, \ldots, r\}$, we denote an index set that contains all feature indices present, i.e., $\mathcal{I}(\boldsymbol{a}) = \{i \mid \boldsymbol{a}_i = 1, i = 1, \ldots, r\}$. When we index vectors with this set, e.g., $\boldsymbol{Z}_{\mathcal{I}}$, we refer to the subvector that only contains the indices in $\mathcal{I}$.

Definition 2 (Protected User Consent, PUC)**.** *Let $f : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}$ be a full feature model. The model f^*_{PUC} that fulfills Protected User Consent is given by*

$$f^*_{PUC}(\mathbf{b}, \boldsymbol{a}, \boldsymbol{z}^*) = \mathbb{F}^{\mathcal{L}}_{(\mathbf{B}, \boldsymbol{A}, \boldsymbol{Z}^*) \sim \mathbf{p}} \left[Y \middle| \mathbf{B} = \mathbf{b}, \boldsymbol{A}_{\mathcal{I}(\boldsymbol{a})} = \mathbf{1}, \boldsymbol{Z}_{\mathcal{I}(\boldsymbol{a})} = \boldsymbol{z}^*_{\mathcal{I}(\boldsymbol{a})} \right],$$

where $\boldsymbol{A}_{\mathcal{I}(\boldsymbol{a})} = \mathbf{1}$ means that each element that is set to 1 in $\boldsymbol{a}$ needs to be one in $\boldsymbol{A}$ as well.

For a single feature (r=1), the index set can either be $\mathcal{I} = \emptyset$ or $\mathcal{I} = \{1\}$ and the definition corresponds to 1D-PUC. The conditional expectation with $\boldsymbol{A}_{\mathcal{I}(\boldsymbol{a})} = \mathbf{1}$ effectively constrains the features in $\mathcal{I}$ to be available, but marginalizes over samples with or without further information.

Implementing Protected User Consent

In this section, we derive a model-agnostic approach called *PUC-inducing data augmentation* (PUCIDA) to achieve protected user consent. By using theoretical analysis, we establish that PUCIDA will result in exact protected user consent. Furthermore, we establish performance guarantees that provide an upper bound on the deviation between practical, finite sample-based PUC-compliant models and their theoretical infinite sample limits.

	state	plan	score	costs
	NSW	basic	87 %	3k$
⊕	**NSW**	**basic**	**N/A**	**3k$**
	NSW	basic	92 %	5k$
⊕	**NSW**	**basic**	**N/A**	**5k$**
	NSW	basic	N/A	64k$

Figure 3: Explaining PUCIDA. Our data augmentation procedure expands each instance with optional information into two samples: The original instance and a synthetic sample (⊕). The synthetic samples retain the base features and the labels, but the information on the optional features is dropped (fitness score → N/A). The model sees samples with the same base features with a missing value and will thus base its decision only on the base features. In this example, given the base features ("NSW", basic) and no optional statements, the model would estimate the costs to be 24k$, which is the dataset average conditioned on these values.

PUCIDA: PUC-inducing Data Augmentation

Intuitively, we want to prevent the model from making inference from a feature's missingness patterns. The core insight is to leverage synthetic samples that make the *distribution of the labels given missingness equal to the overall label distribution*. Thereby, we prevent the derivation of predictive information from the missingness itself (see Table 3).

For a single optional feature, extensively enumerating all samples as in the table is possible while for multiple features this may be intractable. Therefore, we do not list all samples but propose a stochastic, multifeature variant of the algorithm: **(1)** Instead of drawing samples with uniform probability from the distribution $\mathbf{p}$, we use non-normalized weights w:

$$w(\mathbf{x}) = w(\mathbf{b}, \boldsymbol{a}, \boldsymbol{z}^*) = 2^{|\mathcal{I}(\boldsymbol{a})|}. \tag{4}$$

This step corresponds to the expansion of an instance into $2^{|\mathcal{I}(\boldsymbol{a})|}$ synthetic ones; e.g., a sample with a single optional feature is assigned a weight of two (cf. Figure 3). Training instances are drawn with a probability proportional to these weights. This results in data instances with optional information being more frequently sampled. **(2)** We require a sample modification where optional features are randomly dropped from the samples. For each sampled item, we drop each available optional feature with probability p=0.5:

$$\mathbf{q}_i \sim \text{Bern}(0.5), i = 1, \ldots, r; \quad \overline{\boldsymbol{a}} = \mathbf{q} \odot \boldsymbol{a}; \tag{5}$$

$$\overline{\boldsymbol{z}}^*_i = \{\boldsymbol{z}^*_i \text{ if } \overline{\boldsymbol{a}}_i{=}1, \text{ else } \texttt{N/A}\}, i = 1, \ldots, r. \tag{6}$$

(3) We train the predictive model on the modified samples $(\overline{\mathbf{x}}, y) = ((\mathbf{b}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{z}}^*), y) \sim \overline{\mathbf{p}}$ derived through this procedure.

Theoretical Analysis

We summarize PUCIDA in pseudo-code in Appendix D.8 and provide the following theorem to demonstrate that PUCIDA leads to PUC-compliant models.

Theorem 3. *The loss-minimal model $f(\mathbf{b}, \boldsymbol{a}, \boldsymbol{z}^*) = \mathbb{F}^{\mathcal{L}}_{\overline{\mathbf{p}}}[Y | \mathbf{b}, \boldsymbol{A} = \boldsymbol{a}, \boldsymbol{Z}^* = \boldsymbol{z}^*]$ on the modified distribution $\overline{\mathbf{p}}$*

fulfills Protected User Consent with respect to $\mathbf{p}$*, i.e.,*

$$\mathbb{F}^{\mathcal{L}}_{\overline{\mathbf{p}}}\left[Y|\mathbf{B}=\mathbf{b},\boldsymbol{A}=\boldsymbol{a},\boldsymbol{Z}^*=\boldsymbol{z}^*\right]=$$

$$\mathbb{F}^{\mathcal{L}}_{\mathbf{p}}\left[Y\middle|\mathbf{B}{=}\mathbf{b},\boldsymbol{A}_{\mathcal{I}(\boldsymbol{a})}{=}\mathbf{1},\boldsymbol{Z}_{\mathcal{I}(\boldsymbol{a})}{=}\boldsymbol{z}^*_{\mathcal{I}(\boldsymbol{a})}\right]=f^*_{PUC}\left(\mathbf{b},\boldsymbol{a},\boldsymbol{z}^*\right).$$

This result is remarkable in its generality as it enables PUC-compliant models using standard optimization procedures by modifying the distribution of the data; i.e., *PUCIDA can be combined with any existing model and training pipeline*. Next, 'we' study the theoretical convergence behavior for PUCIDA on finite samples. To this end, we define the PUC-Gap as the expected squared deviation from PUC:

$$\text{PUC-Gap}^2(f,\mathbf{p})= \qquad (7)$$
$$\mathbb{E}_{(\mathbf{B},\boldsymbol{A},\boldsymbol{Z}^*)\sim\mathbf{p}}\left[\left(f(\mathbf{B},\boldsymbol{A},\boldsymbol{Z}^*)-f^*_{\text{PUC}}(\mathbf{B},\boldsymbol{A},\boldsymbol{Z}^*)\right)^2\right].$$

We will restrict ourselves to $\mathcal{L}\equiv$ MSE and thus $\mathbb{F}^{\mathcal{L}}\equiv\mathbb{E}$, and study a *baseline conditional expectation estimator* $\hat{\mu}$ which averages the labels conditional on all observations with the same features $\mathbf{x}$. For brevity, we refer to Appendix D.7 (Eqn. 51) for a formal definition of this estimator. Since we usually cannot compute the exact expectation from Theorem 3, we are interested in the number of samples required from $\overline{\mathbf{p}}$ to obtain a fixed average estimation error for which we establish the following result.

Theorem 4 (Finite Sample Convergence)**.** *Let* $\mathcal{X}=\mathcal{X}^b\times(\mathcal{X}^z\cup\{\texttt{N/A}\})$ *be finite feature space and let* $\mathcal{Y}\subseteq\mathbb{R}$ *be the label space. All conditional expectations* $\mu(\mathbf{x}){:=}\mathbb{E}_{\overline{\mathbf{p}}}\left[y|\mathbf{x}\right]$ *and the conditional variances* $\sigma^2(\mathbf{x}){:=}\operatorname{Var}_{\overline{\mathbf{p}}}\left[y|\mathbf{x}\right]$ *exist and are finite. Then there exists a baseline non-parametric regressor* $\hat{\mu}:\mathcal{X}\mapsto\mathbb{R}$ *from a finite number of* N *independent, identically distributed observations* $\left(\overline{\mathbf{x}}_i,y_i\right)_{i=1\ldots N}$ *from* $\overline{\mathbf{p}}$ *with a convergence rate of* $\mathcal{O}(N^{-1})$*; more specifically*

$$\begin{aligned}\text{PUC-Gap}^2(\hat{\mu},\mathbf{p})&=\mathbb{E}_{\mathbf{X}\sim\mathbf{p}}\left[\left(\hat{\mu}(\mathbf{X})-\mu(\mathbf{X})\right)^2\right]\\&\leq\frac{2^r|\mathcal{X}|^2(\sigma^2_{\max}+\mu^2_{\max})}{N}+\mathcal{O}\left(\frac{1}{N^2}\right),\end{aligned}$$

with $\sigma^2_{\max}{:=}\max_{\mathbf{x}\in\mathcal{X}}\sigma^2(\mathbf{x})$ *and* $\mu^2_{\max}{:=}\max_{\mathbf{x}\in\mathcal{X}}\mu^2(\mathbf{x})$*.*

In conjunction with Theorem 3, this result provides a bound on the expected gap to perfect protected user consent that is dependent of the sample size, which decreases with a rate of $\mathcal{O}\left(N^{-1}\right)$. Several remarks are in place: We obtain a multiplicative constant which depends on the number of optional features r and the size of the feature space $|\mathcal{X}|$. The square of this quantity enters the result because the number of samples available to estimate each conditional mean is not independent, as they need to sum up to N. For large feature spaces, however, they are almost independent and we expect the constant to scale almost linearly in $|\mathcal{X}|$. The growth of 2^r is attributed to the re-sampling strategy which might assign a very low probability to certain inputs, which may only be well approximated with a high number of samples. As the number of optional features is typically limited in realistic use-cases it will be well outgrown by N. Note that more powerful model (e.g., Tree based model + PUCIDA) usually outperform this baseline.

data	base model	Full feature model	PUCIDA
diab.(C)	33.84% ±2.47	31.44% ±2.19	34.01% ±1.71
compas (C)	44.47% ±0.37	41.47% ±1.09	44.54% ±0.54
adult (C)	13.37% ±0.07	12.84% ±0.28	13.41% ±0.12
water (C*)	10.65% ±1.64	10.00% ±1.58	10.97% ±1.21
colic (C*)	13.81% ±0.82	11.34% ±0.46	15.05% ±0.68
income (R)	109.56 ±1.00	109.11 ±1.29	110.73 ±1.29
calif. (R)	15.79 ±0.10	15.16 ±0.28	16.18 ±0.06
insurance (R)	283.47 ±0.53	279.78 ±0.42	285.31 ±0.39

Table 1: Availability Inference Restriction is violated by full feature models (Random Forests). As expected, the full feature models always have lower losses than the base-models, indicating that Availability Inference Restriction is violated while PUCIDA fullfils Availability Inference Restriction. We report misclassification error rates for classification models and MSE loss (× 100) for regression models.

Practical considerations. For smaller datasets, an alternative approach to random sampling is to use all possible samples to approximate the distribution $\overline{\mathbf{p}}$ by a method we call "exhaustive augmentation". This involves enumerating all possible variations of the original samples, including any optional features, to form a larger dataset $\mathcal{D}'$. The model is then trained on this expanded dataset.

Experimental Evaluation

Here, we empirically validate the effectiveness of our methods using eight real-world datasets and one synthetic dataset. In particular, we highlight that (a) full feature models violate the Availability Inference Restriction and make it harder for non-sharers to obtain the positive outcome, (b) PUCIDA results in PUC-compliant models as suggested by our theory, and that (c) the reduction in terms of model performance due to using PUC are moderate relative to deploying a full feature model.

Common datasets. We use eight real-world datasets commonly found in the related literature. For classification (C), the Diabetes (diab) and the horse colic dataset (colic) study the prediction of diseases, the COMPAS dataset is concerned with estimating likelihood of recidivism and UCI Adult income dataset requires to predict whether individuals have an income of over 50k$. The water treatment dataset (water) predicts the operational state of a facility. We also study the regression tasks (R) of house price estimation in California (calif), income prediction (income), and inferring information from insurance claims (insurance) to link to our initial example. Details about preprocessing, dataset sources and model hyperparameters are provided in Appendix F.2.

Availability. The colic and the water dataset come with inherent missing values that we use (indicated through $*$). For six more datasets we introduce availability dependent on a feature's value. We compute the probability of feature unavailability $\mathbf{p}(A_i=0|z_i)$ by applying a sigmoid function centered at the feature mean and sample the availability a from the respective conditional distribution. We additionally

				Full feature model		PUCIDA	
task	data	optional	Base feature model	pred.	change	pred.	change
C	diab.	Glucose	60.27%	45.19%	-15.08% ±2.01	61.20%	0.93% ±0.93
C	compas	#priors	51.19%	32.86%	-18.33% ±0.89	51.34%	0.15% ±0.59
C	adult	edu-num	13.86%	11.44%	-2.42% ±0.07	13.92%	0.06% ±0.05
C*	water	oxygen. dem.	87.10%	84.52%	-2.58% ±2.81	87.42%	0.32% ±1.58
C*	colic	abdom. app.	6.39%	1.24%	-5.15% ±0.92	7.01%	0.62% ±1.64
R	income	WKHP	100.0%	81.2%	-18.8% ±0.61	101.2%	1.2% ±0.19
R	calif.	m_income	100.0%	94.4%	-5.6% ±0.67	103.8%	3.8% ±0.42
R	insurance	experience	100.0%	94.8%	-5.2% ±0.09	100.1%	0.1% ±0.05

Table 2: Measuring the average predictions for non-sharers. For classification tasks we report the positive outcomes (in %), and for regression tasks, we report relative predictions to the base feature model (set to 100 %). The non-sharers face disadvantages for not providing the voluntary information and are assigned less favorable prediction outcomes by the full feature models. This discrepancy vanishes when PUCIDA is applied.

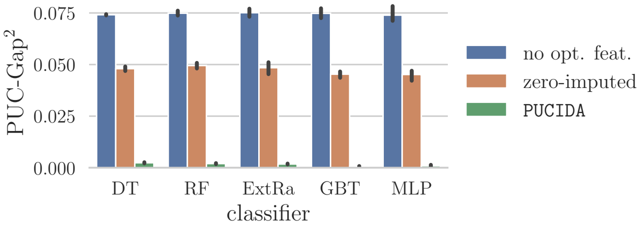

Figure 4: PUCIDA is model-agnostic. The PUC-gaps are close to zero when applying our technique across a variety of common models on the simulated dataset.

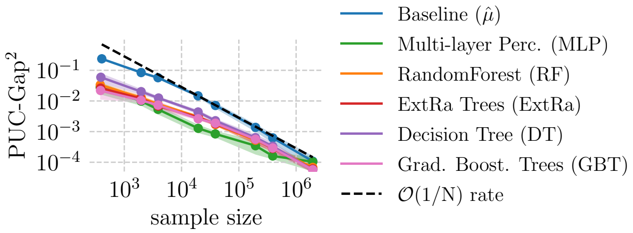

Figure 5: Convergence rate of models under PUCIDA. The estimate of PUC converges to the true value at a rate of $\mathcal{O}(\frac{1}{N})$ for the baseline estimator $\hat{\mu}$ and other commonly used models.

study these datasets in the setting of strategic withholding.

Evaluating PUCIDA

Availability Inference Restriction is violated by full feature models. First, we demonstrate the effect that full feature models have on Availability Inference Restriction. We follow common practices and use zero-imputation to deal with unavailable feature values (Emmanuel et al. 2021). Then, we train a Random Forest model on all features of the dataset where we have introduced stochastic availability into one feature (see previous paragraph). We also train a base feature model that fully drops the optional feature from the dataset. We consider the subset of individuals with unavailable feature values (i.e., a=0) and report the average loss and absolute prediction of the positive class for both models in Table 2. We observe that the full feature models use the information contained in the missingness to obtain a lower loss. This can reduce the chance of obtaining the positive outcome from the full feature model compared to the base feature model by significant margin of up to 18 % for non-sharers. Hence, these results impressively show how the full feature model implicitly infers information from missingness and thereby violates protection requirements. This stays the same when applying established fairness constraints on the models (see Appendix F.1). In contrast, when applying PUC using PUCIDA this gap vanishes or is significantly reduced. We show that the same effect can be observed independently of the imputation techniques, the model class, and the model hyperparameters in Appendix F.3.

Evaluating the Theoretical Bounds

PUCIDA guarantees Predictive Non-Degradation. Usually model performance degrades when training models with additional constraints (e.g., see Corbett-Davies et al. (2017)). To measure model performance, we use the misclassification rate for classification tasks (ROC-AUC scores lead to qualitatively similar results, see Appendix F.4) and the MSE for regression tasks. The results in Table 3a confirm that PUCIDA (using exhaustive augmentation) improves over the base feature model, suggesting that PUCIDA models benefits from using optional information. This is the case even under under strategic actions where users only provide data if it improves their outcome, and aligns with our theoretical result in Corollary 1. Under non-strategic actions, the performance figures show the same characteristics (Appendix F.4). As expected, PUC-compliant models fare moderately worse than full feature models which have no protection requirements.

We now compare two different PUCIDA variants on multiple optional features: the first strategy ensures a fixed dataset size, i.e., the number of samples is equivalent to the original dataset size. The second strategy, which uses ex-

task	data	opt. feature	base model	PUCIDA	Full feature model
C	diab.	Glucose	29.30% ±0.62	**26.61%** ±0.56	23.41% ±0.69
C	compas	#priors	42.89% ±0.10	**40.85%** ±0.15	36.67% ±0.36
C	adult	edu-num	16.05% ±0.03	**15.94%** ±0.05	14.86% ±0.06
R	income	WKHP	85.07 ±0.17	**80.22** ±0.15	73.25 ±0.16
R	calif.	m_income	15.62 ±0.14	**14.79** ±0.08	13.40 ±0.03
R	insurance	experience	262.43 ±0.21	**254.35** ±0.39	236.92 ±0.42

(a) One dimensional case, strategic withholding. Metrics: C: (1-Acc)×100, R: MSE×100

		Fair models				Full feature model
task	data (# opt.)	Base feature model	PUCIDA (f)	PUCIDA (e)	(×)	zero-imputed
C	diab. (2)	29.74 ±2.92	26.23 ±4.42	**25.58** ±3.69	2.2	24.16 ±4.18
C	compas (5)	40.83 ±0.56	37.65 ±0.23	**37.21** ±0.71	7.6	36.86 ±1.20
C	adult (5)	17.98 ±0.37	15.35 ±0.36	**15.27** ±0.25	7.9	15.15 ±0.33
R	income (3)	52.40 ±0.92	**49.47** ±1.71	51.21 ±0.86	3.4	46.15 ±1.60
R	calif. (4)	6.64 ±0.79	6.83 ±0.32	**6.36** ±0.08	5.1	5.69 ±0.22
R	insurance (3)	271.72 ±4.14	**242.99** ±4.47	260.77 ±2.74	3.2	232.59 ±2.39

(b) r-dimensional case. Metrics: C: (1-Acc)×100, R: MSE×100

Table 3: PUC-compliant models leverage optional information to improve predictive performance relative to base feature models This is in line with Corollary 1. In the bottom table, two strategies are considered to achieve PUC: *fixed-size (f)* and *exhaustive (e) PUCIDA*. When using exhaustive PUCIDA, the predictive performance is always better than the performance of the base feature model, and often similar to the performance of the full feature models.

haustive data augmentation, leads to an increased dataset size. The factor by which the dataset size is increased is indicated by (×) along with the results in Table 3b. We observe that competitive results can often be obtained without any dataset increase; fixed-size PUCIDA even outperforms the exhaustive variant on the larger income and the insurance dataset, whereas the exhaustive augmentation leads to a more reliable performance increase. We study the performance for sharers in Table 6 (Appendix) and find that it remains on par with the full feature model. Overall, our results demonstrate that optional information can be leveraged in a conscious way through PUC-inducing data augmentation without suffering from prohibitive performance decrease for the decision maker and the sharers.

Convergence of PUCIDA. Finally, we study the convergence behavior of PUCIDA. As a measure of approximation quality, we use the PUC-Gap[2] defined in Equation (7), which measures the squared deviation from perfect PUC. As this notion requires the knowledge of the ground truth distribution, we use a synthetic dataset for this experiment. The dataset consists of eight binary features (five base, three optional). All features in this dataset are sampled independently. Labels are induced via a logistic distribution, and availability of the optional information depends on the label. For experiments on a second synthetic dataset with five continuous features (two base, three optional) and more details, see Appendix F.5.

First, we observe that PUCIDA is model agnostic, i.e., it works with a variety of state-of-the-art models leading to negligible PUC-gaps (see Figure 4). Second, we verify that the PUC-Gaps converge to zero at the rate of $\mathcal{O}(\frac{1}{N})$ as the sample size increases (Figure 5), confirming what we derived in Theorem 4. While common models (e.g., RandomForest, MLP) have a lower error than the baseline estimator $\hat{\mu}$ the models approach the baseline estimator with larger datasets and the gap closes at the suggested rate.

Conclusion and Future Work

In this work, we studied machine learning predictions where users have the option to disclose optional information. To comply with legal regulations and respect user consent, we introduced the notion of Protected User Consent (PUC) that strikes a balance between the interests of sharers, nonsharers, and decision-makers. We demonstrated that leveraging optional information from consenting users through PUC results in superior performance compared to models that disregard the optional information entirely.

Our work gives raise to several follow-up questions. It would be interesting to study possible long-term effects of PUC and how PUC incentivizes improvements. Furthermore, we have only considered users that act entirely strategic or on privacy grounds. Modeling heterogeneous users, who might be willing to accept a certain increase in costs in return for their privacy could be a meaningful extension.

Additional Material

An extended version of this work including technical appendices is available online[2]. We also publish our code as an open-source project[3].

[2] https://arxiv.org/abs/2210.13954

[3] https://github.com/tleemann/protectedconsent

References

Aleryani, A.; Wang, W.; and De La Iglesia, B. 2020. Multiple imputation ensembles (MIE) for dealing with missing data. *SN Computer Science*, 1(3): 1–20.

Biega, A. J.; Potash, P.; Daumé, H.; Diaz, F.; and Finck, M. 2020. Operationalizing the legal principle of data minimization for personalization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 399–408.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

DeMarco, J. 2023. Nearly 70% of Americans Would Wear a Fitness Tracker/Smartwatch for Discounted Health Insurance.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Emmanuel, T.; Maupong, T.; Mpoeleng, D.; Semong, T.; Mphago, B.; and Tabona, O. 2021. A survey on missing data in machine learning. *Journal of Big Data*, 8(1): 1–37.

Fernando, M.-P.; Cèsar, F.; David, N.; and José, H.-O. 2021. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 36(7): 3217–3258.

Fogliato, R.; Chouldechova, A.; and G'Sell, M. 2020. Fairness evaluation in presence of biased noisy labels. In *International Conference on Artificial Intelligence and Statistics*, 2325–2336. PMLR.

GDPR. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Official Journal of the European Union*.

Ginart, A.; Guan, M. Y.; Valiant, G.; and Zou, J. 2019. Making AI Forget You: Data Deletion in Machine Learning. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*.

Golatkar, A.; Achille, A.; and Soatto, S. 2020. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Goldsteen, A.; Ezov, G.; Shmelkin, R.; Moffie, M.; and Farkash, A. 2021. Data minimization for GDPR compliance in machine learning models. *AI and Ethics*, 1–15.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 111–122.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Henning, L. 2022. Wellness apps and fitness trackers: Why insurers love your smartwatch. *Sydney Morning Herald*.

Izzo, Z.; Anne Smart, M.; Chaudhuri, K.; and Zou, J. 2021. Approximate Data Deletion from Machine Learning Models. In Banerjee, A.; and Fukumizu, K., eds., *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130. PMLR.

Jeong, H.; Wang, H.; and Calmon, F. P. 2022. Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9558–9566.

Kachuee, M.; Karkkainen, K.; Goldstein, O.; Darabi, S.; and Sarrafzadeh, M. 2020. Generative imputation and stochastic prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kilbertus, N.; Rodriguez, M. G.; Schölkopf, B.; Muandet, K.; and Valera, I. 2020. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, 277–287. PMLR.

Kilbertus, N.; Rojas Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Krishnaswamy, A. K.; Li, H.; Rein, D.; Zhang, H.; and Conitzer, V. 2021. Classification with strategically withheld data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 5514–5522.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Lipton, Z.; McAuley, J.; and Chouldechova, A. 2018. Does mitigating ML's impact disparity require treatment disparity? *Advances in neural information processing systems*, 31.

Madison, K.; Schmidt, H.; and Volpp, K. G. 2013. Smoking, obesity, health insurance, and health incentives in the Affordable Care Act. *Jama*, 310(2): 143–144.

Mittelstadt, B. D.; Allo, P.; Taddeo, M.; Wachter, S.; and Floridi, L. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2): 2053951716679679.

Nilforoshan, H.; Gaebler, J. D.; Shroff, R.; and Goel, S. 2022. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, 16848–16887. PMLR.

OAG, C. 2021. CCPA regulations: Final regulation text. *Office of the Attorney General, California Department of Justice*.

Pawelczyk, M.; Leemann, T.; Biega, A.; and Kasneci, G. 2023. On the Trade-Off between Actionable Explanations and the Right to be Forgotten. In *International Conference on Learning Representations (ICLR)*.

Rastegarpanah, B.; Crovella, M.; and Gummadi, K. P. 2020. Fair inputs and fair outputs: The incompatibility of fairness in privacy and accuracy. In *Adjunct Publication of the 28th*

ACM Conference on User Modeling, Adaptation and Personalization, 260–267.

Rastegarpanah, B.; Gummadi, K.; and Crovella, M. 2021. Auditing black-box prediction models for data minimization compliance. *Advances in Neural Information Processing Systems*, 34: 20621–20632.

Rateike, M.; Majumdar, A.; Mineeva, O.; Gummadi, K. P.; and Valera, I. 2022. Don't Throw it Away! The Utility of Unlabeled Data in Fair Decision Making. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1421–1433.

Reeder, B.; and David, A. 2016. Health at hand: A systematic review of smart watch uses for health and wellness. *Journal of biomedical informatics*, 63: 269–276.

Rong, Y.; Leemann, T.; Borisov, V.; Kasneci, G.; and Kasneci, E. 2022. Evaluating feature attribution: An information-theoretic perspective. In *International Conference on Machine Learning*, 18770 – 18795.

Statista. 2023. Wearable Shipments Worldwide.

US Government. U.S. Centers for Medicare & Medicaid Services. 2023. How insurance companies set health premiums.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (Fairware)*, 1–7. IEEE.

Wachter, S.; and Mittelstadt, B. 2019. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Colum. Bus. L. Rev.*, 494.

Wang, Y.; and Singh, L. 2021. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics*, 12(2): 101–119.

Wu, Y.; Dobriban, E.; and Davidson, S. 2020. DeltaGrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning (ICML)*, 10355–10366. PMLR.

Zhang, S.; Qin, Z.; Ling, C. X.; and Sheng, S. 2005. " Missing is useful": Missing values in cost-sensitive decision trees. *IEEE transactions on knowledge and data engineering*, 17(12): 1689–1693.

Zhang, Y.; and Long, Q. 2021. Assessing Fairness in the Presence of Missing Data. *Advances in neural information processing systems*, 34: 16007–16019.

Zimmer, M.; Kumar, P.; Vitak, J.; Liao, Y.; and Chamberlain Kritikos, K. 2020. 'There's nothing really they can do with this information': unpacking how users manage privacy boundaries for personal fitness information. *Information, Communication & Society*, 23(7): 1020–1037.