Detecting and Mitigating Algorithmic Bias in Treatment Effect Estimation: Theory, Methods, and Empirical Evidence

Besides ethical and moral reasons, regulatory and legal frameworks increasingly mandate that machine learning (ML) algorithms should not make errors that systematically disadvantage individuals from different groups. Such errors may arise in ML algorithms for predicting treatment effects, leading to fairness problems around treatment targeting. In this paper, we develop a novel framework for detecting and mitigating bias in ML prediction of treatment effects. We define algorithmic bias as systematic group-wise disparities in the prediction errors of the heterogeneous treatment effects (HTEs). Detecting algorithmic bias can then be cast as a statistical testing problem. For this, we provide an estimator of the algorithmic bias, and we show that our estimator is asymptotically normal. For mitigation, we propose a data-driven procedure that minimizes algorithmic bias in new predictions of HTEs. Our framework is general; it makes minimal assumptions about the predicted HTEs and the ML algorithm. We empirically demonstrate our framework using data from a randomized experiment on the leading travel platform *Booking.com* with 36 million website visitors. Our methods detect and subsequently mitigate algorithmic bias while providing evidence of a fairness-accuracy trade-off in predicted treatment effects. Altogether, our work provides a path forward for ensuring fairness in ML prediction of treatment effects.

Key words: heterogeneous treatment effect; machine learning; algorithmic fairness; field experiment; digital platforms

1. Introduction

Algorithmic bias in machine learning (ML) has led to fundamental concerns in real-world applications (Barocas and Selbst 2016, Campolo et al. 2017, Chouldechova and Roth 2020, De-Arteaga et al. 2022). At a high level, algorithmic bias refers to systematic prediction errors by algorithms, which lead to unfair outcomes for individuals from different groups (Chouldechova and Roth 2020). There are many reasons why algorithmic bias in ML should be addressed. Besides ethical and moral reasons, algorithmic bias may lead to reputational damage for businesses and organizations. Examples are recent media scandals around Amazon's use of an ML-based hiring tool that was biased against women (Reuters 2018), gender bias in Google's targeting of job ads (Lambrecht and Tucker 2019), and Staple's coupon targeting that was biased against low-income customers (Journal 2012). To this end, regulatory and legal frameworks across the world increasingly mandate that

ML algorithms should not make systematic errors that disadvantage individuals of certain groups (Barocas and Selbst 2016, Campolo et al. 2017, Kleinberg et al. 2018, Rambachan et al. 2020). As such, it is imperative for businesses and organizations to address algorithmic bias in their use of ML.

Algorithmic bias can also be a concern when ML is used to predict heterogeneous treatment effects (HTE). In marketing, ML for predicting HTEs is widely used to evaluate the effects of treatments and optimize the allocation of treatments across customers, for instance, in the context of customer retention management (Ascarza 2018, Lemmens and Gupta 2020, Yang et al. 2023), pricing (Smith et al. 2022), promotions (Simester et al. 2020a, Ellickson et al. 2022, Daljord et al. 2023), and personalizing the length of software trials (Yoganarasimhan et al. 2022).[1] In these examples, the predicted HTEs quantify the relative benefit for a company when assigning a customer to a treatment. Hence, if the predicted HTEs suffer from algorithmic bias, a company will make incorrect inferences about which groups of customers are beneficial to treat. This, in turn, can lead to unfair allocations of treatments (e. g., coupons, free shipping, discounts) across different customer groups.

There are many reasons why prediction errors in HTEs can systematically vary across groups (and thus lead to algorithmic bias). Potential reasons include bias in the training data and bias in the ML model (Alaiz-Rodríguez and Japkowicz 2008, Kane et al. 2014, Campolo et al. 2017, Chouldechova and Roth 2020, Simester et al. 2020b, Mehrabi et al. 2021). As an example of the former, the training data may be imbalanced (i.e., one group is more frequent than other groups). This naturally occurs in practice, as many businesses and services are used more frequently by certain customer groups. For example, travel platforms are primarily used by high-income customers who can afford to travel either for business or leisure. If such imbalances are present in the data, then the algorithm will – unless explicitly accounted for – predominantly focus on the prediction performance for the larger groups, leading to larger prediction errors in HTEs for prospective customers from smaller groups. As a result, imbalances in the training data can lead to systematic disparities in prediction errors across groups. As an example of the latter, customers from low-income groups have larger HTEs in many marketing contexts due to a greater price elasticity of demand given a treatment. For instance, a "buy one, get one free" promotion is often more effective for customers who have less disposable income. Hence, if the heterogeneity in the

[1] In marketing, the use of ML models for predicting HTEs is also known as *uplift modeling* (see, e.g., Guelman et al. 2012, Jaskowski and Jaroszewicz 2012, Rzepakowski and Jaroszewicz 2012, Nassif et al. 2013, Kane et al. 2014, Sołtys et al. 2015, Michel et al. 2019, Goldenberg et al. 2020, for specific uplift models). While the causal inference literature typically aims at the prediction of HTEs from non-randomized data under sufficient identification conditions, the uplift modeling literature typically assumes that the training data come from a randomized experiment (Kane et al. 2014).

treatment effect is not explicitly accounted for in the ML model, then companies may infer that the low-income groups should not be targeted, making the company miss out on profits and low-income customers not receiving a beneficial offer.

In this paper, we propose a novel framework with theoretical guarantees for detecting and mitigating algorithmic bias in ML for predicting HTEs. Our approach is motivated by the ethical, regulatory, and operational challenges that businesses and organizations using ML for predicting HTEs at scale (e. g., digital platforms, online marketplaces, and streaming services) face in addressing algorithmic bias. To this end, we define the notion of algorithmic bias in HTEs as group-wise disparities in the prediction errors of the heterogeneous treatment effects (HTEs). Different from much of the research in algorithmic fairness (Chouldechova and Roth 2020), our notion does not focus on systematic disparities in prediction errors for outcomes (which are observable quantities) but in treatment effects (which are counterfactual quantities). Importantly, our framework is general; it makes minimal assumptions about the predicted HTEs and is applicable to various ML models for predicting HTEs. Our contribution is three-fold.

First, we show that detecting algorithmic bias can be cast as a statistical testing problem and then present a tailored estimator. A challenge for the detection of algorithmic bias is that the prediction errors in the HTEs depend on the true but unobservable HTEs. To address this, we use the concept of collapsibility of HTEs (see, e. g., Huitfeldt et al. 2019, Didelez and Stensrud 2022, Colnet et al. 2023) and show that estimating the algorithmic bias per group only requires estimating how well a (weighted) average of the HTE predictions per group recover the associated average treatment effect (ATE). This greatly facilitates the applicability of our framework, as ATEs are considerably easier in terms of identification than HTEs. We further show that our estimator is asymptotically normal and that it recovers the true algorithmic bias in expectation, allowing us to construct statistical tests with large-sample guarantees to detect the true algorithmic bias per group.

Second, we propose a statistical procedure to mitigate algorithmic bias in HTE predictions. Here, it is tempting to simply de-bias the HTE predictions by subtracting the prediction error per group such that the disparities are eliminated. However, we show that such an approach is generally suboptimal when predicting HTEs of new customers. The reason is that it neglects the uncertainty in the detection. That is, if the estimated prediction errors of the ML model relative to the true HTEs per group are incorrect, then a simple subtraction may increase algorithmic bias in unknown ways. As a solution, we derive a data-driven procedure for mitigation that optimally scales the amount of de-biasing so as to minimize the amount of algorithmic bias in the HTE predictions for new customers. We finally show theoretically that, the better we detect the true algorithmic bias,

the better we can eliminate disparities in the prediction errors of HTEs across groups through our data-driven procedure.

Third, we empirically demonstrate our framework using data from a randomized field experiment at the travel platform *Booking.com* with 36 million website visitors from across the globe. Here, we aim to mitigate algorithmic bias in an ML model predicting HTEs that inform the allocation of a free travel benefit across users from different countries. We highlight three findings. (i) Without mitigation, there is substantial heterogeneity in prediction errors and algorithmic bias. In particular, even though the average prediction error in HTEs is zero, there are a few countries where the predictor error in HTEs is more than two standard deviations away from zero. (ii) Our framework mitigates the algorithmic bias but at the cost of shifting the distribution of prediction errors off-center. (iii) Our proposed mitigation strategy that accounts for the variance in the prediction errors per group leads to the largest reduction in algorithmic bias, thus confirming the effectiveness of our framework.

Our work has several implications for research and practice. First, algorithmic biases in prediction algorithms that inform decision-making should be assessed in terms of predicted treatment effects, not predicted outcomes. However, this makes the detection of algorithmic bias challenging as treatment effects are counterfactual estimands, and, thus, any notion of algorithmic bias or fairness defined in terms of prediction errors is also a counterfactual estimand. As a remedy, we provide a framework with theoretical guarantees for the correct detection and mitigation of algorithmic bias in HTE predictions. Second, our framework is highly general: it is applicable to HTEs measured in magnitude or in relative terms, to binary or continuous outcomes, and to various prediction models of HTEs, as long as they provide estimates for the chosen measure of HTEs. Examples of the latter for which our framework is applicable are causal forests (Wager and Athey 2018), ensembles (Sołtys et al. 2015), meta learners (Künzel et al. 2019), double ML (Chernozhukov et al. 2018), and doubly robust ML (Kennedy 2020). Third, our results point to an accuracy-fairness trade-off in that ML models for predicting HTEs may also have a loss in the overall prediction performance when the disparities in the prediction errors are equalized across groups. Thus, as an implication for practice, decision-makers may need to take a stance on whether the ML models predicting HTEs should be optimized for prediction performance or fairness.

The rest of our paper is structured as follows. In Section 2, we position our paper to related work and thereby show how our setting of algorithmic bias in HTEs is different from earlier research. In Section 3, we discuss potential reasons for algorithmic bias in the use of ML for predicting HTEs. In Section 4, we formalize the problem setup of our framework. Sections 5 and 6 then describe our framework for detecting and mitigating algorithmic bias, respectively. Section 7 demonstrates the effectiveness of our framework at *Booking.com*. Finally, Section 8 discusses the implications of our work and concludes.

2. Related Work

Our work relates to two literature streams of previous research, namely, treatment effect estimation and algorithmic fairness. In the following, we briefly review both streams and discuss how our paper contributes.

2.1. Estimating Heterogeneous Treatment Effects

Our work relates to methodological and empirical research on ML models for predicting heterogeneous treatment effects (HTEs). One stream of work known as uplift modeling has proposed adaptions of ML models to predict HTEs (see, e.g., Guelman et al. 2012, Jaskowski and Jaroszewicz 2012, Rzepakowski and Jaroszewicz 2012, Nassif et al. 2013, Kane et al. 2014, Sołtys et al. 2015, Michel et al. 2019, Goldenberg et al. 2020). Uplift modeling is typically used by companies and organizations with the aim to optimize the targeting of treatments to customers. As such, uplift modeling has found widespread use in marketing applications from e-commerce (Gubela et al. 2019), direct marketing (Rzepakowski and Jaroszewicz 2012), and digital platforms (Goldenberg et al. 2020). Methodologically, the literature on uplift modeling commonly assumes that the ML model was trained on data from a randomized experiment. Then, HTEs can be reliably predicted for each individual, as standard causal inference assumptions such as no hidden confounders and overlap are satisfied by design.

Another stream of literature builds upon the theory in causal inference, statistics, and ML to develop models for estimating HTEs (e.g., Athey and Imbens 2016, Wager and Athey 2018, Oprescu et al. 2019). Compared to uplift modeling, this stream focuses more on establishing statistical properties and theoretical guarantees for the methods. Recently, research in marketing has adopted and further tailored ML models from the above streams to different marketing contexts, such as customer retention management (Ascarza 2018, Lemmens and Gupta 2020, Yang et al. 2023), pricing and promotions (Simester et al. 2020a, Ellickson et al. 2022, Smith et al. 2022, Daljord et al. 2023), subscription services (Simester et al. 2020b, Yoganarasimhan et al. 2022), and direct marketing (Hitsch and Misra 2018).

2.2. Algorithmic Bias

2.2.1. Notions of Algorithmic Bias Our work relates to notions of fairness and discrimination in ML, economics, and law. ML research has focused on the properties and implications of criteria that define bias (or fairness) in terms of statistical disparities with respect to sensitive attributes such as race, gender, or nationality (Hardt et al. 2016, Chouldechova 2017, Kleinberg et al. 2017, Carey and Wu 2022). Defining fairness based on a sensitive attribute is in part motivated by legislative frameworks, where such an approach is common (Feldman et al. 2015, Barocas and Selbst 2016, Carey and Wu 2022).

Barocas and Selbst (2016) showed that most criteria proposed in the ML literature are encompassed by three representative criteria formalized as statistical independence relations between predictions, binary classifications, and covariates. One of the criteria is independence. It requires the share of positive and negative labels to be equal across groups. Another criterion is separation. It requires the false-positive or false-negative rate of correct labels to be equal across groups. The final representative criterion is sufficiency. It requires the share of positive or negative predictions to be equal across groups. The representative criteria are developed for the standard ML setting where the objective is to predict outcomes. A canonical example from the algorithmic fairness literature is the decision of a judge to release a defendant based on their predicted risk of recidivism. The prediction model is fitted to data on re-offenders of previously released defendants. Because only released defendants can re-offend, outcome prediction is sufficient. Our setting is different, as it concerns the prediction of treatment effects, which are functions of counterfactual outcomes given a change in the decision. Motivated by this, we define a novel notion of algorithmic bias based on the accuracy of predicted HTEs.

2.2.2. Addressing Algorithmic Bias Having defined a notion of algorithmic bias, the question is how to mitigate it using data. Previous research has classified mitigation into three types of methods: (1) pre-processing, which seeks to alter the training data such that predictive covariates are uncorrelated with the sensitive attribute. (2) in-processing methods, which add a criterion as a constraint to the loss function of the prediction model. (3) post-processing, which adjusts the predictions of the model to be uncorrelated with the sensitive attribute (Pleiss et al. 2017, Mehrabi et al. 2021, Wan et al. 2023).

Previous pre-processing methods include data processing (Kamiran and Calders 2012, Calmon et al. 2017, Johndrow and Lum 2019), learning "fair" predictors (Woodworth et al. 2017), and learning fair representations of the data (Zemel et al. 2013). Many of these methods adjust the data such that the predictive covariates lose explainability. This is a weakness for settings such as ours, where explaining treatment effect heterogeneity is important. In-processing methods include adding a regularization penalty to the loss function such that solutions that do not satisfy a fairness criterion are penalized Bechavod and Ligett (2017), Berk et al. (2017), Ascarza and Israeli (2022), or using constrained optimization techniques to directly optimize the loss function given that a fairness criterion is met (Zafar et al. 2017a,b, Agarwal et al. 2018, Donini et al. 2018, Jiang et al. 2020, Cohen et al. 2021). A weakness with in-processing for mitigating bias is it requires re-training and re-deploying the prediction model, which, as noted earlier, can be highly expensive in large-scale applications such as those on digital platforms. Post-processing includes methods based on graphs (Petersen et al. 2021) and doubly robust estimators from the causal inference literature

(Mishler et al. 2021). Post-processing methods are model-agnostic. Hence, they have the benefit that they work for any prediction model regardless of how it arrives at its predictions (Barocas et al. 2019, Chapter 3). In addition, they do not require altering the covariates or re-training the model, but only its outputs. Hence, post-processing methods such as ours are also suitable for settings where the auditor is an external stakeholder whose objective is not to fix the algorithm but to evaluate it for bias only using data on its inputs and outputs.

2.2.3. Evidence of Algorithmic Bias Our work further contributes to research in marketing, operations, and economics documenting the presence and impact of bias in algorithms. An empirical line of work has studied bias in empirical contexts such as cable news (Goli and Mummalaneni 2023), advertising (Ali et al. 2019, Lambrecht and Tucker 2019), and hiring (Raghavan et al. 2020). A more theoretical stream of papers has proposed notions and the impact of algorithmic fairness in specific decision-making problems. For instance, Kallus and Zhou (2021) study metrics for fairness, equality, and welfare in personalized pricing, Speicher et al. (2018) propose metrics for measuring discrimination in targeted advertising on online platforms, Kozodoi et al. (2022) study the assessment, implementation, and profit implications of fairness in credit scoring, and (Bertsimas et al. 2011, 2012, Nicosia et al. 2017) study the price of fairness in terms of efficient resource allocation, (Celis et al. 2019) study methods for ensuring fairness in users' exposure to auction-based targeting of search ads, and (Goli et al. 2023) propose methods for removing interference bias in product ranking experiments on online marketplaces. We add to these two streams of work by (a) proposing an empirical framework with theoretical guarantees for assessing algorithmic bias in a different context of relevance to marketing, operations, and economics; (b) empirically documenting the presence and mitigation of, as well as the tension between, statistical and algorithmic bias in an ML model for HTE prediction deployed at Booking.com.

2.3. Research Gap

Our work aims to address gaps in the literature on HTE estimation and algorithmic fairness. As for the former, the referenced works on HTE estimation focus on developing or applying estimation methods for optimal learning of HTEs that may potentially be used for downstream targeting decisions. In contrast, we consider the problem of detecting whether prediction errors in HTEs are systematically biased across groups and, if so, how to best mitigate it. Most similar to our work are Ascarza and Israeli (2022) and Huang and Ascarza (2023). The former proposes a tailored splitting criterion for decision trees that ensures HTE predictions are personalized while satisfying group and individual fairness, whereas the latter proposes how to correct a prediction model of HTEs for bias introduced by applying privacy-preserving methods. However, both have salient differences. The former work by Ascarza and Israeli (2022) considers how to address biases during

model training, whereas we consider the problem of detecting whether a model is biased and, if so, how to correct its predictions. The latter work by Huang and Ascarza (2023) does use posthoc correction, but via model training and for bias arising from privacy-preserving noise injected in the HTEs. In contrast, we do not assume a specific source of the bias and address it by adjusting the HTE predictions to recover experimental treatment effect estimates.

As for the algorithmic fairness literature, much of the existing works do not explicitly motivate their approach to address algorithmic bias by the constraints that the user of the method faces in practice. In contrast, we motivate our framework by two stylized facts. First, prediction models may be biased in deployment due to a variety of factors related to the training data and the estimation procedure or combinations thereof that may be highly challenging, if not impossible, to fully characterize. Even if one source of bias is pinpointed and eliminated, other sources of bias may remain. Thus, we intentionally do not seek to attribute sources to the bias, but simply mitigate it overall. Second, training and deploying prediction models of HTEs can be extremely costly and time-consuming in modern application areas such as online platforms, marketplaces, and streaming services. Specifically, companies must first design and run A/B tests to get good training data, then put in significant engineering efforts to train and deploy the model, and finally continuously monitor the model to ensure that it runs smoothly in deployment. Companies may thus value methods that can detect whether a deployed prediction model is biased and, if so, mitigate it. We contribute with an experimental causal inference approach that achieves this aim.

3. Reasons for Bias

In the following, we discuss potential reasons why an uplift model might exhibit algorithmic bias with respect to users' country of origin. This is by no means an exhaustive list of algorithmic bias in general but is instead reasons that are likely to apply to our particular application on Booking.com (see e.g. Chouldechova and Roth 2020, Mehrabi et al. 2021, De-Arteaga et al. 2022, for comprehensive surveys).

- *Bias in the data*: First, users from different groups are not necessarily equally represented in the training data. This has been referred to as sample bias and representation bias (Mehrabi et al. 2021) and is a challenge both for standard outcome prediction (Campolo et al. 2017, Chouldechova and Roth 2020) as treatment effect prediction (Simester et al. 2020b). If observations of a group are not weighted according to their relative sample size, then minimizing the model's loss function is expected to cause a greater prediction accuracy for the majority groups (Chouldechova and Roth 2020). Another source of bias in the data pertains to the distribution of HTEs across groups. Specifically, groups with greater variation in HTEs will tend to have greater variance in predicted HTEs, leading to greater prediction errors given all else equal.

- *Bias in the model:* Another reason for systematic disparities in prediction errors is omitted group-wise treatment effect heterogeneity. This has been referred to as omitted variable bias (Mehrabi et al. 2021), and can arise from the indicator variable of group membership is not included in the model or because it was dropped by the regularization of the ML model. On average, we expect that individuals from certain groups have a higher HTE on average. If the model does not capture this heterogeneity, it will estimate the pooled-average HTE across all individuals, leading to aggregation bias (Mehrabi et al. 2021).
- *Bias in the environment:* The environment that groups are subject to may change from the training of the prediction model to its deployment. In particular, groups may be subject to different market conditions, competitors' actions, and external events that affect the ability of the prediction model to generalize equally well to groups of new individuals (Simester et al. 2020b). The challenge of changes in the environment for predictive accuracy is widely recognized in ML research (Alaiz-Rodríguez and Japkowicz 2008), and has received attention in the literature on uplift modeling (Kane et al. 2014, Simester et al. 2020b).

Potential sources of algorithmic bias as those mentioned above may interact in non-linear and unknown ways, leading to complex forms of algorithmic bias that may be highly difficult to characterize in practice. In light of this, we do not aim to characterize the sources of algorithmic bias. Instead, we seek to detect the algorithmic bias whatever it may be, quantify its magnitude, and then mitigate it.

4. Setup

This section introduces the main objects of interest in our framework. Our presentation is deliberately high-level, as our approach is meant to cover a broad range of cases.

Our approach is motivated by the challenges that companies using ML models of HTEs at scale (e. g., digital platforms, online marketplaces, and streaming services) face in addressing algorithmic bias. One approach to address algorithmic bias is to pre-process the training data. However, this may hinder the explainability of the predicted HTEs, which is of crucial importance for informing decision-making. Another approach is to address the algorithmic bias in the training of the ML model. This necessitates re-estimating and re-deploying the ML model, which is often prohibitively costly in terms of running new A/B tests, expenses for computational power for estimation, and man-hours for putting the ML model into production. Motivated by this, our framework instead takes a novel post-processing approach that only requires the ability to calculate sample moments on data from an existing A/B test containing pre-treatment covariates, a treatment assignment, outcomes, and the HTE predictions of the ML model. Notably, our framework applies to different measures of HTEs, ML models of the HTEs, and forms of algorithmic bias, whether it stems from the data or the training.

4.1. Preliminaries

We first introduce notation. We use capital letters to refer to random variables, where normal font letters are scalars and boldface letters are random vectors. Lower-case letters are sample realizations of the random variables. Greek letters symbolize parameters. Sets are written in calligraphic font.

Let $\mathcal{I} \coloneqq \{1, \ldots, N\}$, $N \in \mathbb{N}$, be the set of individuals of interest. The set of groups is defined as

$$\mathcal{G} \coloneqq \left\{ g_i \subseteq \mathcal{I} \colon i \in \{1, \ldots, N_\mathcal{G}\},\ N_\mathcal{G} \in \mathbb{N}, g_i \neq \emptyset,\ g_i \cap g_j = \emptyset \text{ for } i \neq j,\ \bigcup_i g_i = \mathcal{I} \right\}, \tag{1}$$

i. e., a finite set of non-empty, pairwise disjunct subsets of $\mathcal{I}$. Thus, each individual in $\mathcal{I}$ belongs to exactly one group $g \in \mathcal{G}$. The groups are assumed to be independent. For each group $g \in \mathcal{G}$, we observe data $d_g = \{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^{n_g}$ that are realizations of i.i.d. random variables $(\mathbf{X}_i, T_i, Y_i)$ where, for each individual $i \in g$, $\mathbf{X} \in \mathcal{X}_i \subseteq \mathbb{R}^p$ is a vector of pre-treatment covariates, $T_i \in \{0, 1\}$ is a binary treatment assignment, and $Y_i \in \mathcal{Y} \subset \mathbb{R}$ is an outcome of interest that may be continuous or binary. We assume without loss of generality that higher values of Y are preferred. Following the potential outcomes framework (Rubin 1974), let $Y_i(t)$ denote the potential outcome under treatment assignment $t = 0, 1$. By the fundamental problem of causal inference (Holland 1986), we only observe the potential outcome corresponding to the assigned treatment,

$$Y_i = Y_i(1)T_i + Y_i(0(1 - T_i). \tag{2}$$

In the following, we will drop the subscript i indexing individuals, unless necessary. All notations then refer to an individual drawn at random. This is without loss of generality as the data are i.i.d.

4.2. Heterogeneous Treatment Effect

A heterogeneous treatment effect (HTE) measures heterogeneity in an average treatment effect (ATE) as a function of the pre-treatment covariates. As a function, the HTE is defined as $\tau \colon \mathcal{X} \to \text{Supp}(\tau(\mathbf{X}))$, where $\tau(\mathbf{X})$ is a random variable whenever the pre-treatment covariates $\mathbf{X}$ are random, and $\tau(\mathbf{x})$ is a constant for given covariate profile $\mathbf{X} = \mathbf{x}$. With slight abuse of notation, we let τ denote the ATE corresponding to a HTE $\tau(\mathbf{X})$.

As an estimand, the HTE can be measured in several ways. We provide two common examples below that our framework covers.

Example 1 (Magnitude HTE). The HTE is often defined as the difference

$$\tau(\mathbf{x}) \coloneqq \mathbb{E}[Y(1) \mid \mathbf{X} = \mathbf{x}] - \mathbb{E}[Y(0) \mid \mathbf{X} = \mathbf{x}], \tag{3}$$

which measures heterogeneity with respect to pre-treatment covariates in the population-level magnitude average treatment effect measure (ATE) $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$.

Example 2 (Relative HTE). In many areas, the HTE is commonly defined in unit-less terms as the ratio

$$\tau(\mathbf{x}) :== \frac{\mathbb{E}[Y(1) \mid \mathbf{X} = \mathbf{x}]}{\mathbb{E}[Y(0) \mid \mathbf{X} = \mathbf{x}]}, \tag{4}$$

which measures heterogeneity with respect to pre-treatment covariates in the population-level relative ATE $\tau = \mathbb{E}[Y(1)]/\mathbb{E}[Y(0)]$.

The above measures apply both when the outcome is continuous ($\mathcal{Y} = \mathbb{R}$) and when it is binary ($\mathcal{Y} = \{0, 1\}$). Economics and quantitative social sciences tend to study HTEs for continuous outcomes. Marketing and the health sciences, which are common application areas of HTE estimation, frequently also study HTEs for binary outcomes (e. g., disease or not, conversion or not). With a binary outcome, the conditional potential outcome expectation $\mathbb{E}[Y(t) \mid \mathbf{X}]$ can be expressed as $\mathbb{P}[Y(t) = 1 \mid \mathbf{X}]$, i. e., the conditional probability that the potential outcome is the positive event ($Y = 1$). Then, the magnitude HTE becomes $\tau(\mathbf{x}) = \mathbb{P}[Y(1) = 1 \mid \mathbf{X} = \mathbf{x}] - \mathbb{P}[(0) = 1 \mid \mathbf{X} = \mathbf{x}]$ and the relative HTE becomes $\tau(\mathbf{x}) = \mathbb{P}[Y(1) = 1 \mid \mathbf{X} = \mathbf{x}]/\mathbb{P}[Y(0) = 1 \mid \mathbf{X} = \mathbf{x}]$, where the latter requires the assumption $\mathbb{P}[Y(0) = 1 \mid \mathbf{X} = \mathbf{x}] > 0$ to be well-defined. The magnitude HTE for binary outcomes corresponds to a conditional causal estimand of risk difference, which is a common treatment effect measure in the health sciences. Similarly, the relative HTE for binary outcomes is a conditional estimand of causal relative risk (Hernán and Robins 2006), also known as uplift, or lift factor, in advertising and the tech sector.

By Eq. (2), we only observe one of the two potential outcomes. Because of this, the HTE is a counterfactual quantity that cannot be observed. Thus, the decision-maker relies on estimates from a prediction function f trained on previous data, which maps a user covariate profile $\mathbf{x}$ to a prediction $f(\mathbf{x}) = \widehat{\tau}^f(\mathbf{x})$ of either the magnitude HTE or the relative HTE. In principle, the prediction function is arbitrary as long as it provides estimates for the HTE measure of interest. Nowadays in marketing, it is typically an ML model trained by an algorithm tailored for HTEs. For instance, it could be an ensemble of uplift models (Sołtys et al. 2015), an uplift response transformation model (Jaskowski and Jaroszewicz 2012, Gubela et al. 2020), double ML (Chernozhukov et al. 2018), or a meta-learning algorithm for HTE estimation (e.g., T-learner, S-learner, X-learner, or DR-learner (Künzel et al. 2019, Kennedy 2020)) that provides estimates for the chosen HTE measure. In this context, the benefit of ML models is their data-driven and non-parametric form, aiding in prediction accuracy and avoiding the need to impose inherently unverifiable assumptions on the unobservable potential outcomes. In principle, however, the prediction function could also be a parametric statistical model or even a rule-based procedure. For purposes of generality, we will thus use the terms prediction model and ML model interchangeably to refer to the function f that predicts HTEs.

A benefit of our framework is that it makes minimal assumptions about the prediction model. We only assume that it was estimated on data with randomized treatment assignment, which is considered best practice for estimating and predicting HTEs and commonly used in practice, also by Booking.com in our empirical application. Besides this assumption, we only require access to its predictions and that it satisfies standard regularity conditions. Without loss of generality, we thus leave the functional form, specification, and estimation technique of the prediction model unspecified. In the following, we define our notions of algorithmic bias for ML algorithms of HTEs.

4.3. From Prediction Error to Algorithmic Bias

We now show how prediction errors in HTEs relate to our notion of algorithmic bias. We first define a notion of overall prediction errors.

DEFINITION 1. Fix a prediction function f of a HTE measure. Let τ_g^f be the ATE implied by the predictions for group $g \in \mathcal{G}$ in the population, and let τ_g be the corresponding true ATE for the same group. The population-mean *prediction error* in f for group g is defined as

$$b_g \coloneqq \tau_g^f - \tau_g. \tag{5}$$

We make two remarks on our notation: First, we denote the true bias with a lowercase letter to signify that it is a (unknown) constant, not a random variable. Second, the mean prediction error b_g clearly depends on the choice of prediction model f as well as the data that it was trained on. However, to simplify notation we will omit the dependence of the bias on f and the training data view the prediction model as a fixed object to audit. This mimics the typical setting in practice, where the aim is commonly to detect biases in a specific prediction model.

We use our definition of prediction error to define our notion of algorithmic bias.

DEFINITION 2. Let $\varepsilon \in \mathbb{R}^+$ be a slack parameter and b_{-g} be analogous to Def. 1 but across all observations except those for group g. A prediction function f of an HTE measure has an algorithmic bias against group $g \in \mathcal{G}$ if

$$\left| b_g - b_{-g} \right| \geq \varepsilon. \tag{6}$$

We note three aspects of our notion of algorithmic bias. First, testing for algorithmic bias only requires estimating how "off" the HTE predictions are from the true ATE for a group compared to the rest. Second, algorithmic bias is agnostic to the source or the form of the bias. This allows us to detect and mitigate algorithmic bias without necessarily identifying its causes. This is useful as identifying the causes of effects is an inverse problem, which is either highly challenging or impossible to solve empirically (Maclaren and Nicholson 2019). Nonetheless, possible reasons for algorithmic bias are discussed in Sec. 3. Third, if the prediction model is used to inform treatment

decisions (as in, e. g., customer targeting), then detecting and mitigating our notion of algorithmic bias will address the mechanism (i.e., the ML model) producing the disparities in decisions across groups, not the symptoms thereof (i.e., that different groups may have different outcomes). This is important, as heterogeneity in outcomes across groups is not necessarily evidence of bias, whereas heterogeneity in decisions across groups given identical input to the decisions often is.

Our definition of algorithmic bias motivates a plug-in estimator by replacing the population-level predicted ATE τ_g^f and true ATE τ_g of a group g with sample estimates thereof. That is,

$$\widehat{B}_g = \widehat{\tau}_g^f - \widehat{\tau}_g. \tag{7}$$

This shows that estimating the prediction error b_g simply comes down to estimating how well the predicted HTEs of the ML model recover the ATE per group. An estimate of true ATE $\widehat{\tau}_g$ is straightforward to obtain, as we can simply estimate it directly from the data using an appropriate estimator for the population-level ATE implied by the HTE measure. The challenge lies in obtaining the sample predicted ATE $\widehat{\tau}_g^f$. Here, the problem is that, depending on the HTE measure, the average HTE does not necessarily collapse to the corresponding ATE. The following section details this problem and our solution.

4.4. Collapsibility

To fix ideas, we first define the collapsibility of HTE measures.

Definition 3 (Collapsibility). Let $P\{\mathbf{X}, Y(0)\}$ be the joint distribution of pre-treatment covariates and baseline potential outcome. A treatment effect measure τ is said to be *collapsible* if there exists weights $W = w(\mathbf{X}, P(\mathbf{X}, Y(0)))$ such that for all joint distributions $P(\mathbf{X}, Y(0), Y(1))$ over $\tau(\mathbf{X})$ we have

$$\mathbb{E}\big[W\tau(\mathbf{X})\big] = \tau, \quad \text{with } w \geq 0, \text{ and } \quad \mathbb{E}\big[W\big] = 1. \tag{8}$$

The definition states that an HTE measure is collapsible if we can recover its corresponding ATE via a weighted average. Both the magnitude HTE and the relative HTE that we consider in this work are collapsible measures. Note that other common measures for treatment effects such as the odds ratio are log-odds ratio are inherently non-collapsible (Colnet et al. 2023).

Proposition 1. *Let b_g be given by Def. 1. Then for the magnitude HTE measure, we have*

$$b_g = \mathbb{E}_{P_g}\big[\widehat{\tau}^f(\mathbf{X}) - \tau(\mathbf{X})\big] \tag{9}$$

whereas for the relative HTE measure, we have

$$b_g = \mathbb{E}_{P_g}\left[\frac{\mathbb{E}[Y(0) \mid \mathbf{X}]}{\mathbb{E}[Y(0)]}\big\{\widehat{\tau}^f(\mathbf{X}) - \tau(\mathbf{X})\big\}\right]. \tag{10}$$

Proof. See Appendix A.1 □

The proposition shows that, if we view the ML model as an estimator and its HTE prediction as an estimate, then our definition of prediction error is equivalent to the canonical definition of (weighted) statistical bias. In the remainder of this paper, we will thus use the terms prediction error and statistical bias interchangeably, where convenient.

Proposition 1 clarifies how our notion of algorithmic bias in the prediction model of HTEs actually depends on its HTE predictions rather than their implied ATEs. This insight leads to another result motivating our approach to detect and mitigate algorithmic bias.

PROPOSITION 2. *Let f be a prediction model of HTEs that is optimal in the sense that $\widehat{\tau}^f(\mathbf{x}) = \tau(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. Then f satisfies has no algorithmic bias for any group.*

Proof. The result follows from that algorithmic bias is defined as disparities in prediction error, which, by the proposition, f does not have. □

The proposition states that an ML model with perfect performance in predicting HTEs has no algorithmic bias. The result coincides with that of Corbett-Davies et al. (2017) but, instead of predicting outcomes, the objective is to predict HTEs. The proposition implies that our notion of algorithmic bias can be satisfied in two ways: by estimating the model to not be biased in the first place, or by correcting its bias afterward. The problem with the former approach is that, in practice, we do not actually know whether the model is biased. This motivates the second approach by first detecting whether the makes prediction errors per group and, if so, by how much. The next section presents our approach to this.

5. Detection

5.1. Identification and Estimation

Our methods require prediction error estimates $\widehat{B}_g$ per group g. Reliable estimation of this quantity requires both of its components – the predicted HTE and the true HTE – to collapse to ATE estimates that are unbiased of what they are *supposed to* to identify. To clarify, note that

$$\mathbb{E}[\widehat{B}_g] = \mathbb{E}[\widehat{\tau}_g^f - \widehat{\tau}_g] = \mathbb{E}[\widehat{\tau}_g^f] - \mathbb{E}[\widehat{\tau}_g]. \tag{11}$$

Thus, $\mathbb{E}[\widehat{B}_g] = b_g$ will only reliably hold in practice if (a) $\mathbb{E}[\widehat{\tau}_g^f] = \tau_g^f$, and (b) $\mathbb{E}[\widehat{\tau}_g] = \tau_g$. Here, (a) means that if the HTE predictions $\widehat{\tau}^f(\mathbf{X})$ for group g are in truth off by an amount b_g on average, then the predicted ATE $\widehat{\tau}_g^f$ obtained by aggregating the HTE predictions should, on average across many samples, be off by b_g. But since b_g depends on the unobservable true ATE τ_g, we can only evaluate this if our estimate $\widehat{\tau}_g$ equals the true ATE on average across the same samples. Thus (b) must also hold.

In the following, we discuss unbiased estimators for collapsing the predicted and the true HTE to respective ATE under either the absolute difference measure or the relative measure of the HTE. Following the setup of our framework, We assume access to data from a randomized experiment containing a binary treatment assignment, outcomes, HTE predictions, and potentially pre-treatment covariates for all individuals.[2]

The ATE is a counterfactual estimand unobservable in data. We thus make the following identification assumptions that are standard in the potential outcomes framework for causal inference (Imbens and Rubin 2015, Hernan and Robins 2023).

ASSUMPTION 1. *For all* $t \in \{0,1\}$ *and* $i = 1, \ldots, N$ *we have:*

Consistency*:* $Y_i = Y_i(T)$.

No interference*:* $T_i \perp\!\!\!\perp Y_j(T_j)$ *for all* $j \neq i$.

Strong exchangeability*:* $\{Y_i(t)\}_{t\in\{0,1\}} \perp\!\!\!\perp T_i \mid \mathbf{X}_i$.

Positivity*:* $\mathbb{P}(t_i \mid \mathbf{X}_i = \mathbf{x}) > 0$ *for all* $\mathbf{x}$ *such that* $p_{\mathbf{X}}(\mathbf{x}) > 0$.

Consistency means that the potential outcomes under any treatment assignment equals the observed outcomes given that assignment. It connects the potential outcomes to the observed outcomes and is necessary to identify causal effects. In our case, it follows from that $Y(T) = Y(1)T + Y(0)(1-T)$. No interference means that the potential outcomes of a user do not depend on the treatment assignment to other units. In our application, there is no network dependence between users and, hence, the assumption plausibly holds.[3] Strong exchangeability states that the set of potential outcomes of a user is independent of her treatment assignment and potentially conditional on covariates. This assumption holds by the randomized assignment of the treatment. The positivity assumption states that all users had a positive probability of being treated. This also holds by the randomization. The combination of strong exchangeability and positivity implies strong ignorability, meaning that the treatment assignment mechanism can be ignored.

In the following, we show how to identify and estimate the predicted and the true ATE when the HTE.

[2] Our framework extends to non-random treatment assignments (e. g., observational data) via the use of an appropriate estimator.

[3] We note that in many other applications on digital platforms, marketplaces, and social media networks there may be spillover effects between users that violate the assumption. For instance, imagine that a social media platform A/B tests a feature meant to increase the number of comments on content. If the feature causes an exposed user to post a comment on some content, those who were not exposed but also saw the same content may also post more simply because they want to engage in the conversation.

5.1.1. True Average Treatment Effect. The challenge in estimating the true ATE is that we do not know the true HTE. Hence, we cannot simply take an average to collapse the measure. Instead, we must rely on the identification assumptions to estimate the true ATE only from observed outcomes that either belong to the treatment group or the control groups. It is well known in the causal inference literature that, when the HTE is measured as a magnitude, then the true ATE is identified by the difference in conditional mean outcome among the treated and the controls, i. e.,

$$\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \tag{12}$$

$$= \mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0] \tag{13}$$

$$= \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]. \tag{14}$$

Hence, an unbiased estimator of the true ATE for a group is given by group-wise difference-in-means estimator

$$\widehat{\tau}_g = \widehat{\mathbb{E}}_{n_g}[Y \mid T = 1] - \widehat{\mathbb{E}}_{n_g}[Y \mid T = 0] = \frac{1}{\sum_{i \in g} T_i} \sum_{i \in g} Y_i T_i - \frac{1}{\sum_{i \in g}(1 - T_i)} \sum_{i \in g} Y_i (1 - T_i), \tag{15}$$

where $\widehat{\mathbb{E}}_{n_g}(\cdot)$ is the sample average over group g. The identification for the relative HTE measure follows by the same arguments, i. e.

$$\tau = \frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]} = \frac{\mathbb{E}[Y \mid T = 1}{\mathbb{E}[Y \mid T = 0]}, \tag{16}$$

leading to the ratio-of-means estimator

$$\widehat{\tau}_g = \frac{\widehat{\mathbb{P}}_{n_g}[Y = 1 \mid T = 1]}{\widehat{\mathbb{P}}_{n_g}[Y = 1 \mid T = 0]} = \frac{\left(\sum_{i \in g} T_i\right)^{-1} \sum_{i \in g} T_i Y_y}{\left(\sum_{i \in g}(1 - T_i)\right)^{-1} \sum_{i \in g}(1 - T_i) Y_i}, \tag{17}$$

This estimator is consistent with finite-sample bias vanishing asymptotically (Jewell 1986). Just like the HTE measures themselves, both of these estimators of the true ATE apply irrespective of whether the outcome is continuous or binary. We have thus shown that $\mathbb{E}[\widehat{\tau}_g] = \tau_g$ for both considered HTE measures. Thus, our requirement on one of the components in the statistical bias estimate in our framework is fulfilled.

We make two remarks on the choice of estimator for the true ATE. First, the estimator shall be guided by the definition of HTE estimand in the particular application. Second, for smaller sample sizes than in our application, one may wish to use more efficient estimators, such as parametric or non-parametric regression adjustment, inverse probability weighting, or augmented inverse probability weighting (see e. g., Imbens and Rubin 2015, Kennedy 2020, for details). At a high level, these estimators increase efficiency by controlling for variation in the outcome not explained by

treatment assignment.[4] Again, things become slightly more complicated when the ATE is defined as a ratio. Regression adjustment for ratio-ATEs based on maximum likelihood estimation has been studied in biostatistics, epidemiology, and public health (see e. g., Jewell 1986, Zhang and Kai 1998, Marschner and Gillett 2012, Richardson et al. 2017).

5.1.2. Predicted Average Treatment Effect. We now consider identification and unbiased estimation of the population-level predicted ATE τ_g^f. We first consider the case that the HTE is measured in magnitude. In this scenario, the identification of τ_g^f follows from that, by the definition of collapsibility, its result for the magnitude HTE in proposition 1, and linearity of expectations, we have $\mathbb{E}_{P_g}[\widehat{\tau}_g^f(\mathbf{X})] = \tau_g^f(\mathbf{X})$. Hence, an unbiased estimator is given by simply averaging the HTE predictions, i. e.,

$$\widehat{\tau}_g^f = \widehat{\mathbb{E}}_{n_g}[\widehat{\tau}^f(\mathbf{X})] = \frac{1}{n_g}\sum_{i\in g} f(\mathbf{X}_i). \tag{18}$$

We now consider the alternative case that the HTE is defined as a ratio. Eq. (10) in Proposition 1 shows that the collapsibility of the ratio-HTE to a ratio-ATE involves weighting the ratio-HTEs with $W = \mathbb{E}[Y(0) \mid \mathbf{X}]/\mathbb{E}[Y(0)]$. This suggests two approaches to estimating the predicted ratio-ATE from the predicted ratio-HTEs. The first approach is to assume that baseline potential outcomes $Y(0)$ are independent of the pre-treatment covariates $\mathbf{X}$, formalized in the following.

Assumption 2. $Y(0) \perp\!\!\!\perp \mathbf{X}$

Under the assumption, $\mathbb{E}[Y(0) \mid \mathbf{X}] = E[Y(0)]$ and so the weight $W = \mathbb{E}[Y(0) \mid \mathbf{X}]/\mathbb{E}[Y(0)]$ equals 1. As a result, Eq. (10) reduces to Eq. (9), and identification follows from the same arguments as for the magnitude-ATE. Hence, if the assumption holds, the sample average of the predicted ratio-HTEs is an unbiased estimator of the predicted ratio-ATE. However, the assumption is unlikely to hold in practice, since it implies that all pre-treatment covariates only induce heterogeneity in the treatment effects via their effect on the potential outcome under treatment, $Y(1)$. Furthermore, because the potential outcomes are unobservable this cannot be tested. This leads to a more credible approach, which collapses the predicted ratio-HTEs to the predicted ratio-ATE via weighting. Then, the identification is immediate by Eq. (10) in Proposition 1 and an unbiased estimator is the weighted sample average

$$\widehat{\tau}_g^f = \widehat{\mathbb{E}}_{n_g}[\widehat{W}\widehat{\tau}^f(\mathbf{X})] = \frac{1}{n_g}\sum_{i\in g} \widehat{w}_i f(\mathbf{X}_i). \tag{19}$$

[4] For instance, regression adjustment estimators use that $t = 0, 1$, we have $\mathbb{E}[Y \mid T = t] = \mathbb{E}(\mathbb{E}[Y \mid T = t, \mathbf{X}])$ where the outer expectation is over the distribution of the pre-treatment covariates $\mathbf{X}$.

with

$$\widehat{w}_i = \frac{\widehat{\mathbb{E}}_{n_g}[Y(0) \mid \mathbf{X} = \mathbf{x}]}{\widehat{\mathbb{E}}_{n_g}[Y(0)]} = \widehat{Y}_i(0) \times \left(\frac{1}{\sum_{i \in g}(1 - T_i)} \sum_{i \in g} (1 - T_i) Y_i \right)^{-1}. \tag{20}$$

Here, $\widehat{Y}_i(0)$ is the predicted potential outcome of individual i as a function of the pre-treatment covariates from a suitable model providing asymptotically Gaussian estimates (e. g., parametric models fitted with least squares or maximum likelihood or non-parametric models fitted with double machine learning, doubly robust learning, causal forests, etc.).

The above estimators apply both when the outcome is continuous and when it is binary. We now consider the special case that the outcomes are binary and that the ratio-HTE measure is used. In this case, it is possible to improve upon the weighted average approach suggested above. In particular, a problem with the weighting average estimator is that it involves estimating three expectations. If just one of those is misspecified the estimate of the predicted ATE will be biased, leading to incorrect inferences for detecting algorithmic bias. Hence, as an alternative for settings with binary outcomes, we propose an estimator that only uses observations with a positive outcome ($Y = 1$). That way, violations of Assumption 2 do not apply. However, this comes at the cost of imposing another assumption, stated below.

Assumption 3.

$$\frac{\mathbb{P}[Y(1) = 1 \mid \mathbf{X} = \mathbf{x}]}{\mathbb{P}[Y(0) = 1 \mid \mathbf{X} = \mathbf{x}]} \perp\!\!\!\perp Y \mid g \quad \text{for all} \quad \mathbf{x} \in \mathcal{X}, g \in \mathcal{G}, \tag{21}$$

The assumption states that, within each group, the predicted ratio-HTEs are the same for those with the positive outcome ($Y = 1$) as the those with the negative outcome ($Y = 0$). The assumption is required because the estimator only uses predicted ratio-HTEs of individuals with positive outcomes to estimate the predicted ratio-ATE among all individuals in a group. If the assumption holds, then estimating the predicted ratio-ATE only requires estimating two expectations – the predicted ATE among the treated with positive outcomes, and the predicted ATE among the controls with positive outcomes – and then weighting them for correct collapsibility of the HTEs to the ATE. The first required average is the sample average HTE prediction among the treated. For a given group g, it is given by

$$\widehat{\psi}_g = \frac{1}{\sum_{i \in g} T_i Y_i} \sum_{i \in g} T_i Y_i f(\mathbf{X}_i). \tag{22}$$

The second required average is the analog among the controls, given by

$$\widehat{\lambda}_g = \frac{1}{\sum_{i \in g}(1 - T_i) Y_i} \sum_{i \in g} (1 - T_i) Y_i f(\mathbf{X}_i). \tag{23}$$

The predicted relative ATE for group g by the uplift model is then given by

$$\widehat{\tau}_g^f = \frac{\sum_{i \in g}(1-T_i)Y_i\widehat{\lambda}_g^2 + \sum_{i \in g} T_i Y_i \widehat{\psi}_g}{\sum_{i \in g}(1-T_i)Y_i\widehat{\lambda}_g + \sum_{i \in g} T_i Y_i}. \tag{24}$$

See Appendix C for a derivation.

We have thus presented two three estimators of the predicted ATE when the predicted HTEs are defined as a ratio. The choice between them boils down to ease of use and a bias-variance trade-off. On the one hand, the sample average estimator is very simple to calculate but only unbiased if the potential outcomes under no treatment are the same for everybody irrespective of their pre-treatment covariates. If it does not hold, the estimated predicted ATE will be biased. On the other hand, the weighted sample average estimators are unbiased irrespective of whether the assumption holds, but will have higher variance. For the estimator in Eq. (20), the increase in variance increase stems from the additional estimation of weights, whereas for the estimator in Eq. (66), the increase in variance stems from the aggregation of two averages estimated on fewer observations. To choose an estimator, one should first judge the plausibility of Assumption 2. If reasonable, then the non-weighted sample average estimator in Eq. (18) may be used. If not and outcomes are binary, one should judge the plausibility of Assumption 3. If reasonable, our proposed estimator in Eq. (66) may be used. Otherwise, the estimator in Eq. (66) should be used. Note that for HTEs defined as a difference, the non-weighted sample average estimator can always be used.

Given that we choose an estimator that is correct for the data, and the data are randomly sampled from the population, then all three of the above estimators are unbiased of the population-level predicted ATE, i. e., $\mathbb{E}[\widehat{\tau}_g^f] = \tau_g^f$. Note that this holds irrespective of the size of the unknown true bias. In particular, for any true statistical bias $b \in \mathbb{R}$ such that $f(\mathbf{X}) = \tau(\mathbf{X}) + b$, we have $\widehat{\tau}_g^f$ estimated with either Eq. (20) or Eq. (66) will equal $\mathbb{E}[\widehat{\tau}_g] + b_g$. Then, unbiased estimation of the true statistical bias b_g only requires $\mathbb{E}[\widehat{\tau}_g] = \tau_g$, i. e., unbiased estimation of the true ATE. The next section details our estimators for this.

5.2. Theoretical Results

Our main theoretical result for detection is that, if τ_g^f and τ_g are estimated with the appropriate sample estimators for the definition of the HTE estimand provided in the previous section, then the plug-in estimator of the prediction error in Eq. (7) is asymptotically normal with mean equal to the true prediction error given by Def. 1.

THEOREM 1. *Let n be the size of g, $Z_t \coloneqq \widehat{\mathbb{P}}_n[Y = 1 \mid T = t]$ for $t \in \{0,1\}$ so that $\widehat{\tau}_g \coloneqq Z_0/Z_1$ is the sample means estimator of τ_g, $\widehat{\tau}_g^f = n^{-1}\sum_{i \in g}\widehat{\tau}_g^f(\mathbf{X}_i)$ be the average prediction of f for group g, and let $\mathrm{Var}[\widehat{\tau}_g^f], \mathrm{Var}[\widehat{\tau}_g] < \infty$ so that $\sigma_g^2 \coloneqq Var(\widehat{B}_g) = \mathbb{E}[(\widehat{B}_g - \mathbb{E}[\widehat{B}_g])^2] \in \mathbb{R}^+$. Then*

$$\sqrt{n}\,\widehat{B}_g \xrightarrow{d} \mathcal{N}\Big(b_g, \sigma_g^2\Big) \quad as \quad n \to \infty. \tag{25}$$

Proof. See Appendix A.2. □

Corollary 1.

$$\sqrt{n}\Big(\widehat{B}_g - b_g\Big) \xrightarrow{d} \mathcal{N}(0, \sigma_g^2) \quad as \quad n \to \infty. \tag{26}$$

Corollary 2.

$$\sqrt{n}\Big(\widehat{\tau}_g^f - \widehat{B}_g\Big) \xrightarrow{d} \mathcal{N}(\tau_g, \sigma_{\widehat{\tau}_g}^2) \quad as \quad n \to \infty. \tag{27}$$

The corollaries can be easily verified using the result of the theorem. The proposition states that the probability of incorrectly estimating the prediction error vanishes in the sample limit. The first corollary provides a theoretical guarantee for the use of hypothesis tests to detect algorithmic bias, whereas the second corollary suggests de-biasing by subtracting the estimated prediction error from the mean HTE prediction of a group.

5.3. Inference

We now present hypothesis tests for detecting algorithmic bias according to our notion. By Def. 2, we wish to test the null hypothesis that the prediction error in the ML model among the individuals in a group deviates from its average prediction error among all other individuals, that is

$$\mathcal{H}_0 \colon b_g = b_{-g} \quad \text{vs.} \quad \mathcal{H}_A \colon b_g \neq b_{-g} \tag{28}$$

By Corollary 1, the difference in the estimates of prediction error for a group compared to the rest is a test statistic that follows a t-distribution with $n - p + 1$ degrees of freedom. Specifically,

$$|t| = \frac{\widehat{B}_g - \widehat{B}_{-g}}{\sigma_{g,-g}} \sim t_{n-(p+1)}. \tag{29}$$

where $\sigma_{g,-g} := \sqrt{Var(\widehat{B}_g - \widehat{B}_{-g})}$. We thus reject the null hypothesis if $|t|$ exceeds the critical value $t_{n-(p+1)}(\alpha/2)$ at a pre-specified significance level α. A rejection of the null implies that we have detected algorithmic bias for the group where the size of the bias is our estimate of the prediction error for the group, whereas a failure to reject means that no bias is detected. In practice, a multiple testing correction should be applied to control for the false discovery rate associated with running one test per group. A simple solution is to use the classical Bonferroni correction, which simply scales the significance level to the number of tests. With G groups, the significance level for rejecting the null hypothesis then becomes α/G.

We now state how the slack parameter ε that determines whether a difference in prediction error constitutes algorithmic bias is set.

Proposition 3. *Let ε be defined as in Def. 2. Fix a significance level $\alpha \in [0, 1]$. Then*

$$\varepsilon = t_{n-(p+1)}(\alpha/2) \times \sigma_{g,-g}. \tag{30}$$

The result follows immediately by Def. 2 and the construction of the hypothesis test in Eq. (28). The proposition shows that our procedure to test for algorithmic bias according to Def. 2 is data-adaptive; the confidence with which we detect algorithmic bias is determined by the uncertainty in our estimate of the prediction error.

To provide insights into the algorithmic bias, we may test for prediction error per group. The hypothesis test to conduct is

$$\mathcal{H}_0 \colon b_g = 0 \quad \text{vs.} \quad \mathcal{H}_A \colon b_g \neq 0 \tag{31}$$

Again, it follows immediately by Corollary 1 that

$$|t| = \frac{\widehat{B}_g}{\sigma_g} \sim t_{n-(p+1)}. \tag{32}$$

where $\sigma_{g,-g} := \sqrt{Var(\widehat{B}_g)}$. Thus, a simple t-test on the estimated prediction error per group allows us to detect whether the HTE predictions of the ML model are biased relative to the true HTE for a country on average. Again, a correction for multiple testing may be used.

6. Mitigation

6.1. Challenges

The methods presented so far enable the detection of algorithmic bias via estimation and inference of the prediction error per group, where the prediction error measures the sign and size of the average prediction error in the HTE. Knowledge of the sign and size is useful for characterizing the algorithmic bias, as they tell us for which groups the predictions are, on average, over or underestimated and by how much. By Theorem 1, the estimated bias $\widehat{B}_g$ will asymptotically converge in distribution to a normal random variable centered at the true bias b_g. Hence, a tempting strategy to mitigate algorithmic bias is to, per group, subtract the estimated prediction error from the predicted HTE. By Corollary 1, we would expect the remaining prediction error to be zero for all groups, implying no algorithmic bias.

However, such a strategy may not be optimal. The reason is that it neglects uncertainty. If the estimated prediction error exceeds its true value or is of the wrong sign, it is possible that we increase algorithmic bias rather than reduce it. And since the true bias is unobservable, we will not know whether we overcorrect or not. Appendix. B illustrates this mathematically. To address these challenges, we propose not to simply subtract the full amount of prediction error per group, but to adjust for it according to a chosen loss function, where the expected loss gives the risk of residual bias.

6.2. Objective

We set up our objective for minimizing algorithmic bias as follows. Let $\gamma_g \in [0,1]$ be a correction factor for group g. Then

$$\widehat{\tau}_g^f(\gamma) = \widehat{\tau}_g^f - \gamma_g \widehat{B}_g \tag{33}$$

is the expected HTE predicted by model f for an individual drawn at random for group g corrected by a factor γ_g for its estimated prediction error $\widehat{B}_g$. The remaining prediction error is

$$\begin{aligned} \widehat{\tau}_g^f(\gamma_g) - \tau_g &= \widehat{\tau}_g^f - \gamma_g \widehat{B}_g - \tau_g && (34) \\ &= b_g - \gamma_g \widehat{B}_g. && (35) \end{aligned}$$

The specific value of the prediction error estimate $\widehat{B}_g$ depends on the data with which it is estimated. We thus write $\widehat{B}_g(d_g)$ to denote the estimate we would obtain from our framework given a realized sample $d_g = \{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^{n_g} \overset{i.i.d}{\sim} P_g$ from group $g \in \mathcal{G}$. Let $L(b_g, \gamma_g \widehat{B}_g(d_g))$ be the loss in terms of remaining prediction error. Following statistical decision theory, we define the risk as the expected loss, i. e.,

$$R(b_g, \gamma_g \widehat{B}_g) := \mathbb{E}_{P_g}\left[L\left(b_g, \gamma_g \widehat{B}_g(d_g)\right)\right] = \int_{\mathcal{D}_g} L\left(b_g, \gamma_g \widehat{B}_g(d_g)\right) \mathrm{d}P_g(d_g), \tag{36}$$

where the expectation is over the loss given repeated realizations of data $d_g = \{(\mathbf{x}_i, t_i, y_i)\}_{i=1}^{n_g}$ from the joint distribution P_g. Thus, the risk measures the long-run average loss in terms of remaining prediction errors for group g after applying correction factor γ_g. Because algorithmic bias is defined as disparities in prediction errors, it follows that minimizing the risk per group minimizes algorithmic bias. Hence, the objective is to choose a correction factor that minimizes the risk, i. e.,

$$\gamma_g^* \in \underset{\gamma_g \in [0,1]}{\arg\min}\, R(b_g, \gamma_g \widehat{B}_g). \tag{37}$$

A benefit of our approach is its generality. Each feasible value of the correction factor corresponds to a different bias mitigation strategy in terms of the degree of correction. Here, the boundary case $\gamma = 0$ corresponds to not making a correction, whereas the boundary case $\gamma = 1$ corresponds to the naïve approach of subtracting the full amount of estimated bias. Thus, our approach nests a broad range of feasible mitigation strategies. Suitable correction factors are obtained by solving for the optimal solution to the objective function in Eq. (36) given a chosen loss function. A decision-maker tasked with mitigating algorithmic bias may either choose the optimal solution corresponding to their notion of risk implied by a loss function or, alternatively, derive several correction factors under different loss functions and compare their risk. Loss functions particularly appropriate for mitigating algorithmic bias are the absolute error loss and squared error loss, detailed below.

EXAMPLE 3 (ABSOLUTE ERROR LOSS). We have $L(b_g, \gamma_g \widehat{B}_g) = \left|b_g - \gamma_g \widehat{B}_g)\right|$, implying that the risk function is simply the mean absolute error (MAE) loss

$$R(b_g, \gamma_g \widehat{B}_g) = \mathbb{E}_{P_g}\left[\left|b_g - \gamma_g \widehat{B}_g\right|\right]. \tag{38}$$

The MAE loss places equal weight on all possible estimates of prediction errors.

EXAMPLE 4 (SQUARED ERROR LOSS). The loss is given by $L(b_g, \gamma_g \widehat{B}_g) = (b_g - \gamma_g \widehat{B}_g)^2$, implying that the risk function is the mean squared error (MSE) loss

$$R(b_g, \gamma_g \widehat{B}_g) = \mathbb{E}_{P_g}\left[(b_g - \gamma_g \widehat{B}_g)^2\right]. \tag{39}$$

The MSE loss penalizes larger estimates of prediction error, and may thus be appropriate when we care about reducing large disparities.

The next section presents solutions to the mitigation optimization problem under the MAE and MSE loss, both in terms of theoretical optimum and feasible estimators and inference procedures for applications.

6.3. Theoretical Results

We now present the optimal correction factors implied by the MAE loss and the MSE loss. We first consider the theoretically best correction that we could obtain if we had access to the true algorithmic bias. We then present how these factors can be feasibly estimated from the data for implementation in practice. We start with the MAE loss and then consider the MSE loss.

PROPOSITION 4. *Let b_g be defined as in Proposition 1 for the respective HTE measure of interest. The* oracle-*estimator of the correction factor for Eq.* (36) *implied the MAE loss is given by*

$$\widehat{\gamma}_g^* = \begin{cases} 1, & \textit{if} \quad b_g = 0 \\ 0, & \textit{else.} \end{cases} \tag{40}$$

Proof. The proof follows immediately from that the mean absolute error loss is minimized by subtracting the error whenever it truly is non-zero. □

The optimal correction factor above assumes access to an oracle with knowledge of the true HTEs, the weights for collapsibility of the HTEs to the ATE, and thus the true prediction error. These are not known in practice, and so we use our detection procedure. The *feasible* estimator of the optimal correction under the MAE loss is given by

$$\widehat{\gamma}_g^*(\alpha) = \begin{cases} 1, & \text{if} \quad \mathcal{H}_0 \colon b_g = 0 \text{ is rejected at sig. level } \alpha \text{ in favor of } \mathcal{H}_A \colon b_g \neq 0, \\ 0, & \text{else.} \end{cases} \tag{41}$$

As detailed in Sec. 5.3, the t-statistic for the test is a function of the standard error $\widehat{\sigma}_2$ of the point estimate of statistical bias $\widehat{B}_g$. Hence, the correction factor $\widehat{\gamma}_g^*(\alpha)$ accounts for uncertainty in the

estimated statistical bias. Here, the significance level $\alpha \in [0,1]$ of the test represents the required confidence in the detection to warrant a correction. A value of α closer to one corresponds to a desire for greater certainty that the detected statistical bias is true in order to correct. However, this has the cost of potentially not detecting true statistical bias. The choice of significance level can either be motivated by an established standard, such as $\alpha = 0.05$, or by aligning it with the potential differential costs of type-I errors vs. type-II errors. We refrain from making any specific recommendations here, as our framework is meant to be general and because both the established standards and the costs of type-I and type-II errors may be context-specific.

The following states a key theoretical result for the optimal correction factor implied by an MSE loss.

PROPOSITION 5. *The* oracle-*estimator of the optimal correction factor for Eq.* (36) *implied by the MSE loss has a closed-form solution that can be expressed in the following equivalent ways:*

$$\widehat{g}amma_g^* = \frac{b_g^2}{\sigma_g^2 + b_g^2} = \frac{b_g^2}{\mathbb{E}[\widehat{B}_g^2]} = \frac{\mathbb{E}[\widehat{B}_g^2] - \sigma_g^2}{\mathbb{E}[\widehat{B}_g^2]}. \tag{42}$$

Proof. See Appendix A.3. □

The three ways of expressing the optimal correction factor under MSE loss provide three distinct insights. The first equality shows that the optimal reduction depends on variance σ_g^2 of our bias estimate. Hence, the above approach accounts for how uncertain we are about the deviation of our estimated bias $\widehat{B}_g$ from its true value b_g. In contrast, the naive approach that implicitly fixes the factor γ at one does not account for uncertainty. The second equality shows that the optimal correction factor is simply the squared ratio between the true bias and the expected value of the estimated bias. This makes intuitive sense, as it scales our correction according to how close our expectation of the estimate is to the truth. The worse our bias estimate, the less correction. Conversely, the better our estimate, the greater the correction. In the best case that our estimate equals the true bias, the optimal correction factor will thus equal 1, leading to perfect bias reductions. The third expression writes the optimal correction factor in terms of only estimable quantities. Here, we expect that $\widehat{B}_g \approx \mathbb{E}[\widehat{B}_g]$ as long as our data d_g is a random sample from the joint distribution P_g. By the plug-in principle, we can thus replace $\mathbb{E}[\widehat{B}_g]$ with $\widehat{B}_g$. Likewise, the population variance σ_g^2 of $\widehat{B}_g$ can be replaced with a sample estimate $\widehat{\sigma}_g^2$. Because $\widehat{B}_g$ depends on two estimated quantities, this variance term has a complicated analytical form. We thus propose to estimate it via the non-parametric bootstrap. We later present our approach to this.

7. Application to Randomized Field Experiment on Booking.com

We apply our framework to large-scale data from a randomized field experiment on Booking.com covering more than 36 million users from across the globe. The experiment sought to evaluate an

ML model predicting the HTE in users' probability of booking a stay at a hotel given a free travel benefit. A strong driver of the heterogeneity in the treatment effect of the offer is the user's country of origin. Reducing systematic disparities in the accuracy of the ML model along this dimension is important for downstream decision-making, not only in terms of fairness but also in terms of efficiency. In particular, it ensures that the offer is targeted towards the groups who otherwise would tend not to travel, thereby leading to the most efficient and fair allocation given that the offer cannot be provided to everyone and that users from different countries tend to have different treatment effects. Hence, the objective is to detect whether the ML model deployed at Booking.com was biased with respect to users' country of origin and, if so, mitigate it.

7.1. Data

The treatment variable is an offer of free benefit that incentivizes users to book a stay at a hotel. Examples of other benefits are free breakfast, late check-out, and room service. Users allocated to the treatment condition were shown the offer next to other benefits on the web pages of the hotels included in the campaign. Here, a hotel web page shows accommodation details such as the price of different rooms, the amenities provided, and benefits. Users arrive at a hotel web page by navigating there after making a search for a stay comprising a destination and date range.

The experiment was designed as follows. Users who arrived at the desktop website of Booking.com and who met the eligibility criteria (explained below) were randomly assigned to receive the offer or not. 18.5 million were assigned to either condition. Thus, the data contain 37 million observations at the user-session level. An observation comprises a user's pre-treatment covariates (i.e., a unique user-identifier, country of origin, and 8 covariates of browsing, search, and purchasing history that have been found to predict HTEs at Booking.com but which we cannot disclose due to confidentiality), treatment assignment, a binary booking outcome, and predicted HTE given the covariate values. The HTEs were predicted using a model estimated on data from an identical experiment that ran in 2019. Thus, the randomization, eligibility criteria, sample sizes of the treatment and control conditions, and the covariates that were collected were the same in the experiment used to estimate the ML model of HTEs as the experiment for which its predictions were evaluated. The ML model was then fitted on the experimental data from 2019 to estimate the HTE function of the offer on the relative increase in the likelihood that a user booked given their covariates. The randomization of users into treatment and control ensured overlap in the pre-treatment covariates such that the HTE could be reliably predicted for users assigned in the second experiment that ran in 2020. Details on the prediction model, experimental design, and data are provided in Goldenberg et al. (2020).

The eligibility criteria were as follows. First, users had to visit the page of a hotel that had the offer. Booking.com selected these hotels prior to the experiments. Second, the stay had to meet a

minimum spend threshold pre-determined by Booking.com so as to offset the cost of the benefit. Third, the user's search had to be for a booking of at most six people. Note that, for both the first and second experiments, the random assignment of users into the experimental conditions continued until each condition had an equal number of eligible users, and observations from non-eligible users were discarded. As a result, the users randomly assigned to treatment vs. control are comparable.

The theoretical guarantees of our framework leverage large-sample theory. To increase the reliability of our empirical results, we thus omit countries with less than 10 thousand observations. We are then left with 60 out of the 200 countries, each with at least 5 thousand treatment observations and 5 thousand control observations per the randomization. Due to the proprietary nature of the data, we cannot disclose summary statistics.

7.2. Mitigation Strategies

We demonstrate our mitigation approach using the optimal correction factors under MAE loss and MSE loss and three baselines also derived from our framework.

1. **No correction:** Setting $\gamma = 0$ yields no bias correction. This strategy is the most conservative in the sense that no correction is at least not expected to increase bias.
2. **Mean error:** Setting $\gamma = 1$ yields the naïve strategy of subtracting the full amount of estimated prediction error irrespective of its variance. This strategy represents a baseline that neglects uncertainty.
3. **Mean error if rejected null:** This is the optimal correction factor under the MAE loss provided in Eq. 41 where we set $\alpha = 0.05$. Hence, it is identical to the above strategy except that it only subtracts estimated bias when it is statistically significant.
4. **Mean squared error minus approach:** This is the optimal correction factor under the MSE loss provided in the last expression of Eq. (42), where we estimate the expected prediction error $\mathbb{E}[\widehat{B}_g^2]$ with the bootstrap sample average $\widehat{\bar{B}}_g^2$ and approximate the population variance σ_g^2 with the bootstrap sample variance $\widehat{\sigma}^2$.
5. **Mean squared error plus approach:** This is a plug-in estimator of the oracle-optimal correction factor under MSE loss given by the first expression in Eq. 42, where we replace b_g^2 with its sample estimate $\widehat{B}_g^2$ and the population variance with the same bootstrap variance estimate $\widehat{\sigma}_g^2$ as in the former correction approach. Hence, this serves as a baseline to the former optima correction factor derived from the MSE loss.

The main difference between the mitigation strategies that do perform a correction lies in whether or not they penalize larger errors and if and how they account for uncertainty.

7.3. Evaluation Procedure

We provide a general cross-validated, sample split, bootstrap procedure for counterfactual evaluation of bias mitigation. The bootstrap allows us to estimate the variance prediction error estimates required by the detection and mitigation of algorithmic bias, the sample split allows us to evaluate the mitigation for new observations not used for detection, and the cross-validation reduces the sensitivity of our results to the random sample split and the random bootstrap samples. The procedure is as follows. For cross-validation fold $k = 1, \ldots, K$:

1. Split the data into a training set $\mathcal{S}$ and a test set $\mathcal{V}$ each of size $N/2$
2. For each group $g \in \mathcal{G}$:
   (a) On $\mathcal{S}$, use the estimators in Sec. 5.1.2 and Sec. 5.1.1 to estimate the predicted and the true ATE for the HTE measure, both for the observations in group g and the pooled rest.
   (b) Calculate the prediction error estimate $\widehat{B}_g = \widehat{\tau}_g^f - \widehat{\tau}_g$ and $\widehat{B}_{-g} = \widehat{\tau}_{-g}^f$ and $\widehat{\tau}_{-g}$and the associated algorithmic bias $\widehat{A}_g = \widehat{B}_g - \widehat{B}_{-g}$. Use the non-parametric bootstrap to get the standard errors of the estimates.
   (c) Test the null hypothesis in Eq. (28) of no algorithmic bias at significance level α
   (d) Calculate the correction factor $\widehat{\gamma}_g^{(c)}$ corresponding to strategy $c = 1, \ldots, C$ in Sec. 7.2
   (e) Repeat steps A–B. on $\mathcal{V}$ to obtain an estimate of the predicted and true ATE $\widehat{\tau}_g^f$ and $\widehat{\tau}_g$ representative for the test data.
   (f) For each mitigation strategy $c = 1, \ldots, C$, get the fold-k test data remaning prediction error $\widehat{B}_{g,l}^{(k)} = \widehat{\tau}_g^f - \widehat{\gamma}_g^{(c)} - \widehat{\tau}_g$ and $\widehat{B}_{-g,c}^{(k)} = \widehat{\tau}_{-g}^f - \widehat{\gamma}_{-g}^{(c)} - \widehat{\tau}_{-g}$ and the associated test data remaning algorithmic bias $\widehat{A}_{g,c}^{(k)} = \widehat{B}_{g,c}^{(k)} - \widehat{B}_{-g,c}^{(k)}$. Use the non-parametric bootstrap to approximate the sampling distribution of the estimates.

Pseudo-code for implementing the procedure is provided in Appendix D.

To implement our procedure on the data, we use $K = 10$ cross-validation folds and set the number of bootstrap runs to $B = 50$. In step 2(a), we apply the ratio-of-means estimator in Eq. (17) to estimate the true ATE and apply our weighted average estimator in Eq. (66) to collapse the ML model's relative HTE predictions to a predicted relative ATE. In step 2(c), we set the significance level α to the 5% standard. In step 2(d), we calculate the correction factors of five mitigation strategies presented in the previous section. Step 2(e) uses the same estimators as in step 2(a) but on the hold-out test data per group.

7.4. Empirical Results

Fig. 1 shows the kernel density estimates of the prediction error and the algorithmic bias across countries prior to mitigation. All densities are estimated with a Gaussian kernel and bandwidth set according to Silverman's "rule of thumb" (p. 48 Silverman 1986), which are the default settings for

ggplot in Python. We then double the selected bandwidth to smooth out wiggles in the densities and infer general patterns more easily. We apply this smoothing to all densities. Finally, the densities are on a standardized scale, meaning that for both measures of bias, we divide the bias for a country by the standard deviation in the bias across countries.

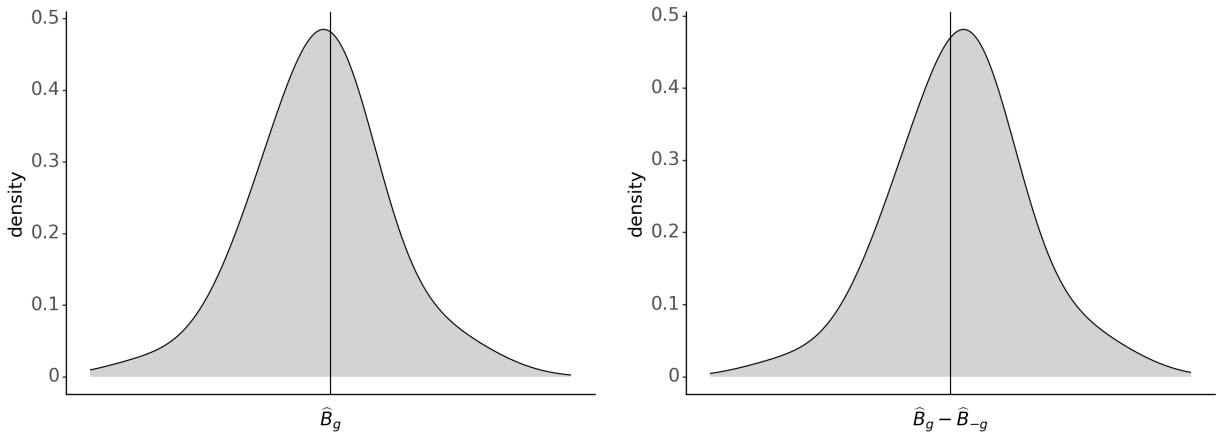
We highlight three findings. First, we find evidence of cross-country variation in the prediction error of the ML model (Fig. 1a). The empirical density is maximized around zero, implying that for most countries there is no prediction error. Moreover, the distribution is roughly symmetric with a center at zero. This suggests that the ML model predicts HTEs accurately for most countries, on average. For a few countries, however, the prediction error is two standard deviations away from zero.

Second, the distribution of the algorithmic bias in the ML model is highly similar to that of the prediction error, except that the mode is positively shifted by about half a standard deviation (Fig. 1b). Thus, for the countries at the mode, the ML model is estimated to overestimate the HTE relative to individuals from the other countries. The discrepancy in the distributions between statistical and algorithmic bias is explained by that our notion of algorithmic bias measures the difference in prediction error of a given country and the pooled rest, where the composition of the rest depends on for which country the algorithmic bias is calculated. Here, the majority groups contribute more to the pooled rest, thereby accounting for disparities due to representation bias in the data.

Third, for most countries, we do not detect algorithmic bias in the ML model when we account for the uncertainty in the bias estimates. This is supported by that the cross-country empirical distribution of the t-statistic from tests aligns with the theoretical distribution of the test statistics under the null hypothesis (Fig. 2). Here, the t-statistic values on the x-axis range from -3 to 3, corresponding to a 99% confidence level. However, only a few countries' t-statistics exceed ± 1.64 and ± 1.96, implying a rejection of a two-sided null of no algorithmic bias at a 90% and 95% significance level, respectively.

Next, we apply our evaluation procedure detailed in Sec. 7.3 to analyze how the different mitigation strategies in Sec. 7.2 affect the bias towards new individuals from different countries of origin. We emphasize that the evaluation uses a train-test split, such that, per country of origin, the individuals that were used to obtain the correction factors are different from those for which they were evaluated in terms of remaining bias.

Fig. 3 shows empirical kernel densities of the statistical and algorithmic bias that remains per country, after a correction according to each of the mitigation strategies. We detail two results. First, the mitigation strategies decreased the modes of the distributions. As for prediction error, the mitigation shifted the mode from zero to slightly below (cf. Fig. 1a and Fig. 3a). As for the



(a) Statistical bias (b) Algorithmic bias

Figure 1 **Empirical density of (a) size of statistical bias, (b) size of algorithmic bias across countries of origin over 10-fold cross-validation each with 50 bootstrap runs for each country.**

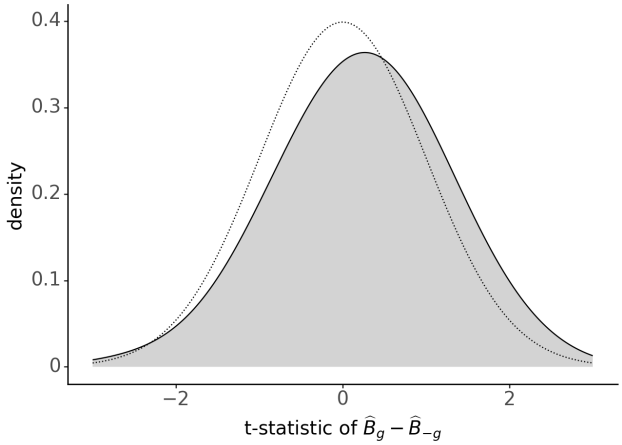
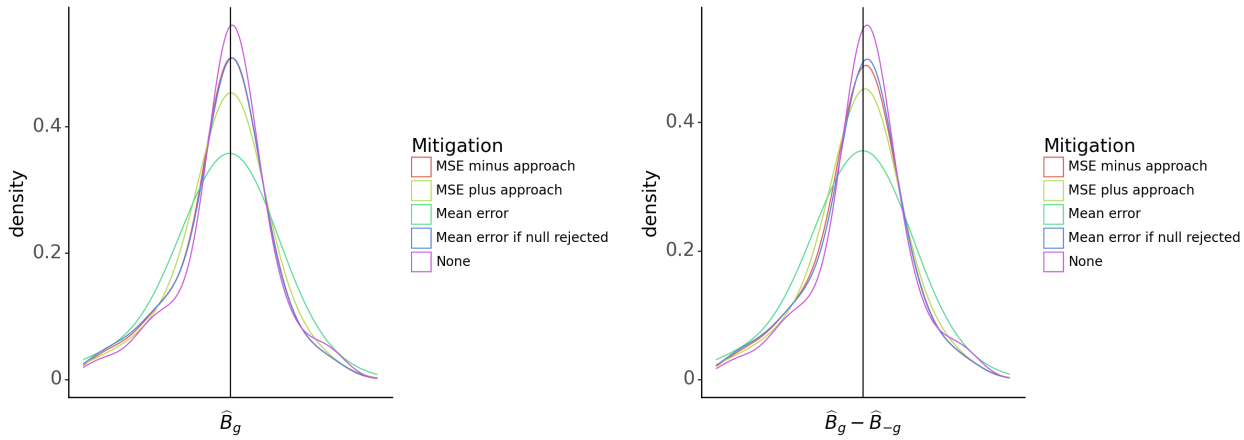


Figure 2 **t-statistic of the test of algorithmic bias across countries of origin with the theoretical density of the test statistic under the null hypothesis overlaid. The t-statistics are computed over 10-fold cross-validation each with 50 bootstrap runs for each country.**

algorithmic bias, the mode was above center before mitigation (Fig. 1b), and so the corrections led to an improvement, albeit with a slight overcorrection (Fig. 3b). Taken together, the mitigation strategies improve the distribution of algorithmic bias at a minor cost in overall prediction performance. Second, the different mitigation strategies were not equally effective. The best mitigation strategies are no correction or a mean error correction conditional on a detected bias, followed by the MSE approach that accounts for sampling variability in the prediction error. This can be seen from the fact that their post-mitigation densities of algorithmic bias are most concentrated at zero with the smallest variance. The worst-performing mitigation strategies are subtracting the

estimated prediction error without accounting for uncertainty and the MSE error approach that does not account for sampling variability.



(a) Remaining statistical bias (b) Remaining algorithmic bias

Figure 3 **Empirical density of remaning (a) statistical bias, and (b) algorithmic bias across all countries of origin over 10-fold cross-validation each with 50 bootstrap runs for each country, after bias mitigation with different correction factors.**

8. Discussion

Our work contributes to previous research along three dimensions. First, previous work in marketing has focused on how to adapt ML for improving the estimation of HTEs or optimizing targeting policies in different empirical contexts (e. g., Ascarza 2018, Hitsch and Misra 2018, Lemmens and Gupta 2020, Simester et al. 2020a,b, Ellickson et al. 2022, Smith et al. 2022, Yoganarasimhan et al. 2022, Daljord et al. 2023, Yang et al. 2023). In contrast, we consider the problem of detecting whether a given prediction model with that aim is biased and, if so, how to mitigate it. Somewhat similar, Ascarza and Israeli (2022) propose an in-processing method for eliminating bias in decision trees, and Huang and Ascarza (2023) propose a model-based post-processing method that removes bias in predicted HTEs caused by privacy-preserving noise in the prediction covariates. We contribute by providing a general, non-parametric causal inference approach to the detection and mitigation of both statistical and algorithmic bias that may arise due to a variety of factors.

Second, research in algorithmic fairness has mostly considered algorithmic bias in binary classification in contexts such as hiring, law, and policing. In contrast, our framework addresses bias in HTE predictions from an ML model in a marketing context. Our work thereby adds to the literature on "counterfactual fairness" (Kusner et al. 2017) by addressing a new problem requiring a tailored approach. For instance, in marketing, treatment effect predictions may systematically vary across

groups without being a case for algorithmic bias. Moreover, treatment effects are unobservable, thereby posing challenges for mitigating bias in their predictions. Our techniques for combating these challenges may be explored for other applications of counterfactual fairness.

Third, recent research in marketing has shown that state-of-the-art causal inference methods for ATE estimation from observational data may not recover ATE estimates from randomized experiments (Gordon et al. 2019). Here, we find that HTE estimates of an ML model may not collapse equally well to experimental ATE estimates for different groups, even though the ML model was trained on experimental data. Such systematic disparities in the ability to predict treatment effects for different groups may arise because of representation bias in the experimental training data, failure of the algorithm's design or training to capture group-wise treatment effect heterogeneity, or because of changes in the environment that shift the distributions of data differently across groups. Irrespective of the source of the bias, we show that our framework can detect the systematic prediction error per group and mitigate it. However, in doing so, our results point to an accuracy-fairness trade-off; ML models for predicting HTEs may lose overall prediction performance when the disparities in the prediction errors are equalized across groups.

Our work provides several implications for future research and practice. Decision-makers seeking to address biases in algorithms should formalize which types of disparities in data constitute bias and which do not. Only with mathematical criteria in hand can the detection and mitigation be rigorously evaluated. We define a notion of algorithmic bias in ML models of HTEs and provide methods that can be used in practice with theoretical guarantees.

Another implication of our work is that any approach to mitigate bias in a prediction algorithm by de-biasing its outputs should account for the uncertainty in the found bias. In particular, the naïve approach of simply subtracting the found bias might lead to under or overcorrections that potentially worsen disparities in unknown ways. Given that any correction may be imperfect, we propose an uncertainty-aware approach that minimizes the risk of bias remaining after a correction.

Finally, our framework is not only applicable to audits by internal stakeholders (e. g., the designer of the prediction algorithm), but also to independent audits by external stakeholders (e. g., organizations interested in algorithmic oversight and policy). A challenge for the latter is the limited insight into the algorithm to audit and resources associated with the new data collection or modeling. Notably, our framework only requires comparing sample averages and prediction averages for different groups on past experimental data. The designer typically already has such data from A/B tests used to train and evaluate the prediction algorithm. Hence, external stakeholders who have been given the right to an independent audit can request such data and then use our framework to detect potential biases and explore the consequences of different mitigation strategies.

8.1. Concluding Remarks

In this paper, we have proposed a framework for detecting and mitigating bias in ML models for HTE prediction. We define notions of statistical and algorithmic bias for HTE prediction and provide estimation and inference methods for addressing the bias in practice. We provide theoretical guarantees for our methods and empirically evaluate them using large-scale data from a randomized experiment on the travel platform Booking.com. We find that the ML model deployed at Booking.com was biased with respect to users' country of origin in predicting HTEs but that our methods can successfully mitigate this, while at the same time illustrating a fairness-accuracy trade-off for treatment effect prediction. Our work provides practitioners with simple-to-implement procedures based on past experimental data to audit and correct algorithms estimating HTEs for bias without the need to re-estimate and re-deploy the prediction model. It thereby offers a cost-effective solution for companies' to mitigate biases against certain groups in downstream treatment decision-making.

Funding and Competing Interests

Authors 1, 4, and 5 have no competing interests. Authors 2 and 3 are employed at the partner company.

References

Agarwal A, Beygelzimer A, Dudík M, Langford J, Wallach H (2018) A Reductions Approach to Fair Classification. *International Conference on Machine Learning*, 60–69 (PMLR).

Alaiz-Rodríguez R, Japkowicz N (2008) Assessing the Impact of Changing Environments on Classifier Performance. *Advances in Artificial Intelligence: 21st Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2008 Windsor, Canada, May 28-30, 2008 Proceedings 21*, 13–24.

Ali M, Sapiezynski P, Bogen M, Korolova A, Mislove A, Rieke A (2019) Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):1–30.

Ascarza E (2018) Retention Futility: Targeting High-Risk Customers Might Be Ineffective. *Journal of Marketing Research* 55(1):80–98.

Ascarza E, Israeli A (2022) Eliminating Unintended Bias in Personalized Policies using Bias-eliminating Adapted Trees (BEAT). *Proceedings of the National Academy of Sciences* 119(11):e2115293119.

Athey S, Imbens G (2016) Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113(27):7353–7360.

Barocas S, Hardt M, Narayanan A (2019) *Fairness and Machine Learning: Limitations and Opportunities* (fairmlbook.org), `http://www.fairmlbook.org`.

Barocas S, Selbst AD (2016) Big Data's Disparate Impact. *California law review* 671–732.

Bechavod Y, Ligett K (2017) Penalizing Unfairness in Binary Classification. *arXiv preprint arXiv:1707.00044* .

Berk R, Heidari H, Jabbari S, Joseph M, Kearns M, Morgenstern J, Neel S, Roth A (2017) A Convex Framework for Fair Regression. *arXiv preprint arXiv:1706.02409* .

Bertsimas D, Farias VF, Trichakis N (2011) The Price of Fairness. *Operations Research* 59(1):17–31.

Bertsimas D, Farias VF, Trichakis N (2012) On the Efficiency-Fairness Trade-off. *Management Science* 58(12):2234–2250.

Calmon F, Wei D, Vinzamuri B, Natesan Ramamurthy K, Varshney KR (2017) Optimized Pre-processing for Discrimination Prevention. *Advances in Neural Information Processing Systems* 30.

Campolo A, Sanfilippo MR, Whittaker M, Crawford K (2017) AI Now 2017 Report .

Carey AN, Wu X (2022) The Causal Fairness Field Guide: Perspectives from Social and Formal Sciences. *Frontiers in Big Data* 5:892837.

Celis E, Mehrotra A, Vishnoi N (2019) Toward Controlling Discrimination in Online Ad Auctions. *International Conference on Machine Learning*, 4456–4465 (PMLR).

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/Debiased Machine Learning for Treatment and Structural Parameters. *Econometrics Journal* 21(1):C1–C68.

Chouldechova A (2017) Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big data* 5(2):153–163.

Chouldechova A, Roth A (2020) A Snapshot of the Frontiers of Fairness in Machine Learning. *Communications of the ACM* 63(5):82–89.

Cohen MC, Miao S, Wang Y (2021) Dynamic Pricing with Fairness Constraints. *Available at SSRN 3930622* .

Colnet B, Josse J, Varoquaux G, Scornet E (2023) Risk Ratio, Odds Ratio, Risk Difference... Which Causal Measure is Easier to Generalize? *arXiv preprint arXiv:2303.16008* .

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

Daljord Ø, Mela CF, Roos JM, Sprigg J, Yao S (2023) The design and targeting of compliance promotions. *Marketing Science* .

De-Arteaga M, Feuerriegel S, Saar-Tsechansky M (2022) Algorithmic Fairness in Business Analytics: Directions for Research and Practice. *Production and Operations Management* 31(10):3749–3770.

Didelez V, Stensrud MJ (2022) On the Logic of Collapsibility for Causal Effect Measures. *Biometrical Journal* 64(2):235–242.

Donini M, Oneto L, Ben-David S, Shawe-Taylor JS, Pontil M (2018) Empirical Risk minimization under Fairness Constraints. *Advances in neural information processing systems* 31.

Doob JL (1935) The Limiting Distributions of Certain Statistics. *The Annals of Mathematical Statistics* 6(3):160–169.

Ellickson PB, Kar W, Reeder III JC (2022) Estimating Marketing Component Effects: Double Machine Learning from Targeted Digital Promotions. *Marketing Science* .

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and Removing Disparate Impact. *International Conference on Knowledge Discovery and Data Mining*, 259–268.

Goldenberg D, Albert J, Bernardi L, Estevez P (2020) Free Lunch! Retrospective Uplift Modeling for Dynamic Promotions Recommendation within ROI Constraints. *Proceedings of the 14th ACM Conference on Recommender Systems*, 486–491.

Goli A, Lambrecht A, Yoganarasimhan H (2023) A Bias Correction Approach for Interference in Ranking Experiments. *Marketing Science* .

Goli A, Mummalaneni S (2023) Gender Diversity on Cable News: An Analysis of On-Screen Talent and Viewership. *Available at SSRN 4462592* .

Gordon BR, Zettelmeyer F, Bhargava N, Chapsky D (2019) A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook. *Marketing Science* 38(2):193–225.

Greenland S, Pearl J, Robins JM (1999) Confounding and Collapsibility in Causal Inference. *Statistical Science* 14(1):29–46.

Gubela R, Bequé A, Lessmann S, Gebert F (2019) Conversion Oplift in E-commerce: A Systematic Benchmark of Modeling Strategies. *International Journal of Information Technology & Decision Making* 18(03):747–791.

Gubela RM, Lessmann S, Jaroszewicz S (2020) Response Transformation and Profit Decomposition for Revenue Uplift Modeling. *European Journal of Operational Research* 283(2):647–661.

Guelman L, Guillén M, Pérez-Marín AM (2012) Random Forests for Uplift Modeling: an Insurance customer Retention Case. *International Conference on Modeling and Simulation in Engineering, Economics and Management*, 123–133.

Hardt M, Price E, Srebro N (2016) Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems* 29.

Hernan M, Robins J (2023) *Causal Inference: What If* (Boca Raton: Chapman & Hall/CRC).

Hernán MA, Robins JM (2006) Estimating Causal Effects from Epidemiological Data. *Journal of Epidemiology & Community Health* 60(7):578–586.

Hitsch GJ, Misra S (2018) Heterogeneous Treatment Effects and Optimal Targeting Policy Evaluation. *Available at SSRN* .

Holland PW (1986) Statistics and Causal Inference. *Journal of the American statistical Association* 81(396):945–960.

Huang TW, Ascarza E (2023) Debiasing Treatment Effect Estimation for Privacy-Protected Data: A Model Audition and Calibration Approach. *Available at SSRN 4575240* .

Huitfeldt A, Stensrud MJ, Suzuki E (2019) On the Collapsibility of Measures of Effect in the Counterfactual Causal Framework. *Emerging Themes in Epidemiology* 16:1–5.

Imbens GW, Rubin DB (2015) *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge University Press).

Jaskowski M, Jaroszewicz S (2012) Uplift Modeling for Clinical Trial Data. *ICML Workshop on Clinical Data Analysis*, volume 46, 79–95.

Jewell NP (1986) On the Bias of Commonly Used Measures of Association for 2 x 2 Tables. *Biometrics* 351–358.

Jiang R, Pacchiano A, Stepleton T, Jiang H, Chiappa S (2020) Wasserstein Fair Classification. *Uncertainty in artificial intelligence*, 862–872 (PMLR).

Johndrow JE, Lum K (2019) An Algorithm for Removing Sensitive Information. *The Annals of Applied Statistics* 13(1):189–220.

Journal TWS (2012) Websites Vary Prices, Deals Based on users' Information. URL `https://www.wsj.com/articles/SB10001424127887323777204578189391813881534`.

Kallus N, Zhou A (2021) Fairness, Welfare, and Equity in Personalized Pricing. *Conference on Fairness, Accountability, and Transparency*, 296–314.

Kamiran F, Calders T (2012) Data Preprocessing Techniques for Classification without Discrimination. *Knowledge and Information Systems* 33(1):1–33.

Kane K, Lo VS, Zheng J (2014) Mining for the Truly Responsive Customers and Prospects using True-Lift Modeling: Comparison of New and Existing Methods. *Journal of Marketing Analytics* 2:218–238.

Kennedy EH (2020) Towards Optimal Doubly Robust Estimation of Heterogeneous Causal Effects. *arXiv preprint arXiv:2004.14497* .

Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR (2018) Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10:113–174.

Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent Trade-Offs in the Fair Determination of Risk Scores. *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*.

Kozodoi N, Jacob J, Lessmann S (2022) Fairness in Credit Scoring: Assessment, Implementation and Profit Implications. *European Journal of Operational Research* 297(3):1083–1094.

Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for Estimating Heterogeneous Treatment Effects using Machine Learning. *Proceedings of the National Academy of Sciences* 116(10):4156–4165.

Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual Fairness. *Advances in Neural Information Processing Systems* 30.

Lambrecht A, Tucker C (2019) Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management science* 65(7):2966–2981.

Lemmens A, Gupta S (2020) Managing Churn to Maximize Profits. *Marketing Science* 39(5):956–973.

Maclaren OJ, Nicholson R (2019) What can be Estimated? Identifiability, Estimability, Causal Inference and Ill-Posed Inverse Problems. *arXiv preprint arXiv:1904.02826* .

Marschner IC, Gillett AC (2012) Relative Risk Regression: Reliable and Flexible Methods for Log-Binomial Models. *Biostatistics* 13(1):179–192.

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR)* 54(6):1–35.

Michel R, Schnakenburg I, Von Martens T (2019) *Targeting Uplift: An Introduction to Net Scores* (Springer Nature).

Mishler A, Kennedy EH, Chouldechova A (2021) Fairness in Risk Assessment Instruments: Post-processing to Achieve Counterfactual Equalized Odds. *Conference on Fairness, Accountability, and Transparency*, 386–400.

Nassif H, Kuusisto F, Burnside ES, Page D, Shavlik J, Santos Costa V (2013) Score as you Lift (SAYL): A Statistical Relational Learning Approach to Uplift Modeling. *Machine Learning and Knowledge Discovery in Databases*, 595–611 (Springer).

Nicosia G, Pacifici A, Pferschy U (2017) Price of Fairness for Allocating a Bounded Resource. *European Journal of Operational Research* 257(3):933–943.

Oprescu M, Syrgkanis V, Wu ZS (2019) Orthogonal Random Forest for Causal Inference. *International Conference on Machine Learning*, 4932–4941 (PMLR).

Petersen F, Mukherjee D, Sun Y, Yurochkin M (2021) Post-processing for Individual Fairness. *Advances in Neural Information Processing Systems* 34:25944–25955.

Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ (2017) On Fairness and Calibration. *Advances in Neural Information Processing Systems* 30.

Raghavan M, Barocas S, Kleinberg J, Levy K (2020) Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. *Conference on Fairness, Accountability, and Transparency*, 469–481.

Rambachan A, Kleinberg J, Mullainathan S, Ludwig J (2020) An Economic Approach to Regulating Algorithms. Technical report, National Bureau of Economic Research.

Reuters (2018) Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women. URL `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`.

Richardson TS, Robins JM, Wang L (2017) On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association* 112(519):1121–1130.

Rubin DB (1974) Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66(5):688.

Rzepakowski P, Jaroszewicz S (2012) Uplift Modeling in Direct Marketing. *Journal of Telecommunications and Information Technology* 43–50.

Silverman BW (1986) *Density Estimation for Statistics and Data Analysis*, volume 26 (CRC press).

Simester D, Timoshenko A, Zoumpoulis SI (2020a) Efficiently Evaluating Targeting Policies: Improving on Champion vs. Challenger Experiments. *Management Science* 66(8):3412–3424.

Simester D, Timoshenko A, Zoumpoulis SI (2020b) Targeting Prospective Customers: Robustness of Machine-Learning Methods to Typical Data Challenges. *Management Science* 66(6):2495–2522.

Slutsky E (1925) Über Stochastische Asymptoten und Grenzwerte. *Metron (in German)* 5(3):3–89.

Smith AN, Seiler S, Aggarwal I (2022) Optimal price targeting. *Marketing Science* 42(3):476–499.

Sołtys M, Jaroszewicz S, Rzepakowski P (2015) Ensemble Methods for Uplift Modeling. *Data Mining and Knowledge Discovery* 29:1531–1559.

Speicher T, Ali M, Venkatadri G, Ribeiro FN, Arvanitakis G, Benevenuto F, Gummadi KP, Loiseau P, Mislove A (2018) Potential for Discrimination in Online Targeted Advertising. *Conference on Fairness, Accountability and Transparency*, 5–19 (PMLR).

Ver Hoef JM (2012) Who Invented the Delta Method? *The American Statistician* 66(2):124–127.

Wager S, Athey S (2018) Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Wan M, Zha D, Liu N, Zou N (2023) In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Transactions on Knowledge Discovery from Data* 17(3):1–27.

Woodworth B, Gunasekar S, Ohannessian MI, Srebro N (2017) Learning Non-discriminatory Predictors. *Conference on Learning Theory*, 1920–1953 (PMLR).

Yang J, Eckles D, Dhillon P, Aral S (2023) Targeting for Long-Term Outcomes. *Management Science* .

Yoganarasimhan H, Barzegary E, Pani A (2022) Design and Evaluation of Optimal Free Trials. *Management Science* .

Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017a) Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *International Conference on World Wide Web*, 1171–1180.

Zafar MB, Valera I, Rogriguez MG, Gummadi KP (2017b) Fairness constraints: Mechanisms for fair classification. *Artificial intelligence and statistics*, 962–970 (PMLR).

Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning Fair Representations. *International Conference on Machine Learning*, 325–333 (PMLR).

Zhang J, Kai FY (1998) What's the Relative Risk?: A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. *JAMA* 280(19):1690–1691.

Online Appendix

Appendix A: Proofs

A.1. Proof of Proposition 1

We have

$$b_g = \tau_g^f - \tau_g \tag{43}$$

$$\text{(Def. 3)} = \mathbb{E}_{P_g}\left[W\widehat{\tau}^f(\mathbf{X})\right] - \mathbb{E}_{P_g}\left[W\tau(\mathbf{X})\right] \tag{44}$$

$$\text{(linearity of expectations)} = \mathbb{E}_{P_g}\left[W\left(\widehat{\tau}^f(\mathbf{X}) - \tau(\mathbf{X})\right)\right]. \tag{45}$$

For the magnitude HTE, the weight $W = 1$ (Greenland et al. 1999), so that

$$b_g = \mathbb{E}_{P_g}\left[\widehat{\tau}^f(\mathbf{X}) - \tau(\mathbf{X})\right]. \tag{46}$$

This proves the first statement of the proposition.

For the relative HTE, we have $W = \mathbb{E}[Y(0) \mid \mathbf{X}]/\mathbb{E}[Y(0)]$ (see e. g., Huitfeldt et al. 2019, Colnet et al. 2023). Hence,

$$b_g = \mathbb{E}_{P_g}\left[\frac{\mathbb{E}[Y(0) \mid \mathbf{X}]}{\mathbb{E}[Y(0)]}\left\{\widehat{\tau}^f(\mathbf{X}) - \tau(\mathbf{X})\right\}\right] \tag{47}$$

as stated in the second part of the proposition. This concludes the proof. □

A.2. Proof of Theorem 1

We provide the proof for the case when the HTE is defined as a ratio. The proof technique is the same when the HTE is defined as a difference.

Let $h\colon \mathbb{R} \times \mathbb{R} \setminus \{0\} \to \mathbb{R}$, $(x, y) \mapsto x/y$ be the standard division. The function h is smooth and it's gradient $\nabla h(x, y) = (1/x, -x/y)$ is continuous on the support of h. Let μ_1 and μ_0 be the true means of random variables Z_1 and Z_0, respectively. We then have

$$\begin{aligned}
\sigma^2 &\coloneqq Var[h(\mu_1, \mu_0)] \\
&\approx \nabla h \begin{pmatrix} \mu_1 \\ \mu_0 \end{pmatrix}^\top \mathbf{\Sigma} \nabla h \begin{pmatrix} \mu_1 \\ \mu_0 \end{pmatrix} \\
&= \nabla h \begin{pmatrix} \mu_1 \\ \mu_0 \end{pmatrix}^\top \begin{pmatrix} \sigma_0^2/n & \sigma_{0,1} \\ \sigma_{0,1} & \sigma_1^2/n \end{pmatrix} \nabla h \begin{pmatrix} \mu_1 \\ \mu_0 \end{pmatrix} \\
&= \frac{\sigma_0^2}{n\mu_1^2} - 2\frac{\mu_0 \sigma_{0,1}}{\mu_1^3} + \frac{\sigma_1^2 \mu_0^2}{n\mu_1^4}
\end{aligned}$$

where $\mathbf{\Sigma}$ is the covariance matrix of τ_g, with $\sigma_{0,1}$ being the covariance of Z_1 and Z_0. It follows by the multivariate version of the delta method (Doob 1935, Ver Hoef 2012) that

$$\sqrt{n}\left(\frac{Z_1}{Z_0} - \frac{\mu_1}{\mu_0}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right) \quad \text{for } n \to \infty. \tag{48}$$

Let $Z_1 = \widehat{\mathbb{P}}_n[Y = 1 \mid T = 1]$ and $Z_0 = \widehat{\mathbb{P}}_n[Y = 1 \mid T = 0]$. Then $\widehat{\tau}_g = Z_1/Z_0$ and we have $\mu_1 \coloneqq \mathbb{P}[Y = 1 \mid T = 1]$, $\mu_0 \coloneqq \mathbb{P}[Y = 1 \mid T = 0]$, and thus $\tau_g = \mu_1/\mu_0$. It follows directly from the above that

$$\sqrt{n}\left(\widehat{\tau}_g - \tau_g\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right) \quad \text{for } n \to \infty. \tag{49}$$

Now, it follows directly from the asymptotic unbiasedness of $\widehat{\tau}_g^f]$ and Def. 1 that

$$\sqrt{n}\left(\widehat{\tau}_g^f - \tau_g\right) \xrightarrow{p} b_g \quad \text{for } n \to \infty. \tag{50}$$

Hence, by Slutsky's Theorem (Slutsky 1925), we get

$$\begin{aligned}
\sqrt{n}\,\widehat{B}_g &= \sqrt{n}\left(\widehat{\tau}_g^f - \widehat{\tau}_g\right) \\
&= \sqrt{n}\left(\left(\widehat{\tau}_g^f - \tau_g\right) - \left(\widehat{\tau}_g - \tau_g\right)\right) \\
&\xrightarrow{d} \mathcal{N}\left(b_g, \sigma^2\right).
\end{aligned}$$

This concludes the proof. □

A.3. Proof of Proposition 5

We first derive how the bias magnitude depends on the correction factor. For ease of exposition we omit the group index g. Let

$$M_\gamma = \tilde{B}_\gamma^2 = (b - \gamma\widehat{B})^2. \tag{51}$$

Using the definition of the expectation of squared random variables, we get

$$\mathbb{E}[M_\gamma] = \gamma^2\sigma^2 + b^2(\gamma - 1)^2. \tag{52}$$

We find the optimal value of γ by solving for the first-order conditions. The first derivative of γ is

$$\frac{\partial \mathbb{E}[M_\gamma]}{\partial \gamma} = \frac{\partial}{\partial \gamma}\left(\gamma^2\sigma^2 + b^2(\gamma - 1)^2\right) = 2\gamma\sigma^2 + 2b^2(\gamma - 1),$$

Setting it to 0 and solving for γ yields

$$\begin{aligned} 2\gamma\sigma^2 + 2b^2(\gamma - 1) &= 0 \\ \iff \gamma\sigma^2 &= b^2 - \gamma b^2 \\ \iff \gamma(\sigma^2 + b^2) &= b^2 \end{aligned}$$

Thus, the optimum reduction of statistical bias in the MSE-sense is

$$\gamma^* = \frac{b^2}{\sigma^2 + b^2} \tag{53}$$

This is the first equality in Proposition 5. Now, simply using that $\mathbb{E}[\widehat{B}^2] = b^2 + \sigma^2$ and swapping terms in the denominator of Eq. (53) directly gives the second sought equality in Proposition 5, i. e.,

$$\gamma^* = \frac{b^2}{\sigma^2 + b^2} = \frac{b^2}{\mathbb{E}[\widehat{B}^2]} \tag{54}$$

Similarly, re-arranging terms to get $b^2 = \mathbb{E}[\widehat{B}^2] - \sigma^2$ and swapping terms in the numerator of Eq.(54) gives the third equality. Thus

$$\gamma^* = \frac{b^2}{\sigma^2 + b^2} = \frac{b^2}{\mathbb{E}[\widehat{B}^2]} = \frac{\mathbb{E}[\widehat{B}^2] - \sigma^2}{\mathbb{E}[\widehat{B}^2]} \tag{55}$$

This concludes the proof. □

Appendix B: The Insufficiency of Simple Bias Subtraction

To simplify notation we omit the subscript g for group. We first analyze what happens if we simply subtract $\widehat{B}$ from $\widehat{\tau}^f$. Let

$$\tilde{\tau}^f = \widehat{\tau}^f - \widehat{B} \tag{56}$$

The remaning bias is given by

$$\begin{aligned}
\tilde{B} &= \mathbb{E}[\tilde{\tau}^f - \tau] && (57)\\
&= \mathbb{E}[\tilde{\tau}^f - \widehat{B} - \tau] && (58)\\
&= \mathbb{E}[b - \widehat{B}] && (59)\\
&= b - \mathbb{E}[\widehat{B}]. && (60)
\end{aligned}$$

If $\widehat{B} \sim \mathcal{N}(b, \sigma^2)$ as established by Theorem 1, then $\tilde{B} \sim \mathcal{N}(0, \sigma^2)$ as stated in Corollary 1. Now,

$$Var[\tilde{B}] = \mathbb{E}[\tilde{B}^2] - (\mathbb{E}[\tilde{B}])^2. \tag{61}$$

Since $\mathbb{E}[\tilde{B}] = 0$, it follows that

$$\mathbb{E}[\tilde{B}^2] = Var(\tilde{B}) = \sigma^2. \tag{62}$$

Hence, simply subtracting the full amount of estimated statistical bias $\widehat{B}$ only leads to a reduction of the squared statistical bias if $b^2 > \sigma^2$, which is not necessarily the case. The most extreme case is $b = 0$, for which the correction introduces bias. Note that even though $\widehat{B}$ is a consistent estimator of b, the squared estimate $\widehat{B}^2$ is not a consistent estimator of b^2. Hence, simply subtracting $\widehat{B}^2$ will not reduce the squared remaning bias $\tilde{B}^2$. This motivates our approach to find a correction fraction to minimize the MSE of remaning squared statistical bias.

Appendix C: Derivation of Estimator

We now show the derivation of the estimator in Sec. 5.1.2. We omit the group subscript g as the results apply to each group separately. We have that

$$\widehat{\tau}_g = \frac{\sum_{i\in g}(1-T_i)Y_i f(\mathbf{X}_i)\widehat{\lambda}_g + \sum_{i\in g} T_i Y_i \widehat{\psi}_g}{\sum_{i\in g}(1-T_i)Y_i\widehat{\lambda}_g + \sum_{i\in g} T_i Y_i \frac{1}{f(\mathbf{X}_i)}\widehat{\psi}_g} \tag{63}$$

$$= \frac{\sum_{i\in g}(1-T_i)Y_i f(\mathbf{X}_i) \times \frac{1}{\sum_{i\in g}(1-T_i)Y_i}\sum_{i\in g}(1-T_i)Y_i f(\mathbf{X}_i) + \sum_{i\in g} T_i Y_i \frac{1}{\sum_{i\in g} T_i Y_i}\sum_{i\in g} T_i Y_i f(\mathbf{X}_i)}{\sum_{i\in g}(1-T_i)Y_i \frac{1}{\sum_{i\in g}(1-T_i)Y_i}\sum_{i\in g}(1-T_i)Y_i f(\mathbf{X}_i) + \sum_{i\in g} T_i Y_i \frac{1}{f(\mathbf{X}_i)} \times \frac{1}{\sum_{i\in g} T_i Y_i}\sum_{i\in g} T_i Y_i f(\mathbf{X}_i)} \tag{64}$$

$$= \frac{\frac{1}{\sum_{i\in g}(1-T_i)Y_i}(\sum_{i\in g}(1-T_i)Y_i f(\mathbf{X}_i))^2 + \sum_{i\in g} T_i Y_i f(\mathbf{X}_i)}{\sum_{i\in g}(1-T_i)Y_i f(\mathbf{X}_i) + \sum_{i\in g} T_i Y_i} \tag{65}$$

$$= \frac{\sum_{i\in g}(1-T_i)Y_i\widehat{\lambda}_g^2 + \sum_{i\in g} T_i Y_i \widehat{\psi}_g}{\sum_{i\in g}(1-T_i)Y_i\widehat{\lambda}_g + \sum_{i\in g} T_i Y_i}. \tag{66}$$

This concludes the derivation of the estimator.

We can check that the estimator correctly identifies the predicted ATE as follows. If the HTE is a constant τ_g for all users in a group g (i.e., the HTE of a user does not depend on her covariates $\mathbf{X}$), then the SATE, SATT, and SATU should all equal τ_g. We thus replace the HTE predictions $f(\mathbf{X}_i)$ with τ_g in the formulas for the ATE for the treated, ATE for the controls, and ATE, and check if they all simplify to τ_g. For the ATE for the treated, we get

$$\widehat{\psi}_g = \frac{\sum_{i\in g} T_i Y_i}{\sum_{i\in g} T_i Y_i \times \frac{1}{\tau_g}} = \frac{1}{\sum_{i\in g} T_i Y_i}\sum_{i\in g} T_i Y_i \tau_g = \tau_g. \tag{67}$$

For the ATE for the controls, we also get that

$$\widehat{\lambda}_g = \frac{1}{\sum_{i\in g}(1-T_i)Y_i}\sum_{i\in g}(1-T_i)Y_i\tau_g = \tau_g. \tag{68}$$

Finally, plugging these into the estimator for the ATE in Eq.. (66) yields

$$\frac{\sum_{i\in g}(1-T_i)Y_i\tau_g^2 + \sum_{i\in g} T_i Y_i \tau_g}{\sum_{i\in g}(1-T_i)Y_i\widehat{\tau}_g + \sum_{i\in g} T_i Y_i} = \tau_g \times \frac{\sum_{i\in g}(1-T_i)Y_i\tau_g + \sum_{i\in g} T_i Y_i}{\sum_{i\in g}(1-T_i)Y_i\tau_g + \sum_{i\in g} T_i Y_i} = \tau_g, \tag{69}$$

as we sought to show. This confirms that the estimator identifies the ATE.

Appendix D: Pseudo-code for Empirical Evaluation of Bias Mitigation

Algorithm 1 details step-by-step pseudo-code for how to implement our cross-validated, sample split, bootstrap procedure to evaluate the mitigation strategies.

Algorithm 1: Evaluation procedure for bias mitigation

Input:
- Number of cross-validation folds K;
- Number of bootstrap runs Z;
- Significance level α;
- Mitigation strategies $c = 1, \ldots, C$;
- Data $d_g = \{(\mathbf{x}_i, t_i, y_i)\}$ per group $g \in \mathcal{G}$

Output: Estimates of remaining statistical and unfairness per group

for $k = 1, \ldots, K$ **do**
  Randomly split the data into training set $\mathcal{S}$ and test set $\mathcal{V}$ of equal size $N/2$;
  for $g \in \mathcal{G}$ **do**
    On $\mathcal{S}$, obtain $\widehat{\tau}_g^f$ and $\widehat{\tau}_{-g}^f$ via Eq. (22), Eq. (23), and then Eq. (66);
    On $\mathcal{S}$, obtain $\widehat{\tau}_g$ and $\widehat{\tau}_{-g}$ via Eq. (17);
    $\widehat{B}_g \leftarrow \widehat{\tau}_g^f - \widehat{\tau}_g$;
    $\widehat{B}_{-g} \leftarrow \widehat{\tau}_{-g}^f - \widehat{\tau}_{-g}$;
    $\widehat{A}_g \leftarrow \widehat{B}_g - \widehat{B}_{-g}$;
    for $z = 1, \ldots, Z$ **do**
      Construct $\mathcal{S}_g^{(z)}$ by randomly sampling with replacement from $\mathcal{S}_g$;
      $\mathcal{S}_{-g}^{(z)} \leftarrow \mathcal{S} \setminus \mathcal{S}_g^{(z)}$;
      Repeat steps 4–7 separately on $\mathcal{S}_g^{(z)}$ and $\mathcal{S}_g^{(z)}$ to get $\widehat{B}_g^{(z)}$ and $\widehat{B}_{-g}^{(z)}$
    $\widehat{\sigma}_g \leftarrow (Var[\widehat{B}_g^{(z)}])^{1/2}$;
    $\widehat{\sigma}_{-g} \leftarrow (Var[\widehat{B}_{-g}^{(z)}])^{1/2}$;
    Test $\mathcal{H}_0 \colon b_g = b_{-g}$ vs. $\mathcal{H}_A \colon b_g \neq b_{-g}$ at significance level α following Sec. 5.3;
    for $c = 1, \ldots, C$ **do**
      Calculate $\widehat{\gamma}_{g,c}$;
    for $z = 1, \ldots, Z$ **do**
      Construct $\mathcal{V}_g^{(z)}$ by randomly sampling with replacement from $\mathcal{V}_g$;
      $\mathcal{V}_{-g}^{(z)} \leftarrow \mathcal{V} \setminus \mathcal{V}_g^{(z)}$;
      Separately repeat steps 1–2 on $\mathcal{V}_g^{(z)}$ and $\mathcal{V}_{-g}^{(z)}$ to get test data estimates $\widehat{\tau}_g^{f,z}$, $\widehat{\tau}_{-g}^{f,z}$, $\widehat{\tau}_g^{(z)}$, and $\widehat{\tau}_{-g}^{(z)}$;
      for $c = 1, \ldots, C$ **do**
        $\widehat{B}_{g,c}^{(k,z)} \leftarrow \widehat{\tau}_g^{f,z} - \widehat{\gamma}_{g,c}\widehat{B}_g - \widehat{\tau}_g^{(z)}$;
        $\widehat{B}_{-g,c}^{(k,z)} \leftarrow \widehat{\tau}_{-g}^{f,z} - \widehat{\gamma}_{-g,c}\widehat{B}_{-g} - \widehat{\tau}_g^{(z)}$;
        $\widehat{A}_{g,c}^{(k,z)} \leftarrow \widehat{B}_{g,c}^{(k,z)} - \widehat{B}_{-g,c}^{(k,z)}$;

return $\widehat{B}_{g,c}^{(k,z)}$ and $\widehat{A}_{g,c}^{(k,z)}$ for $z \in [Z]$, $k \in [K]$, $g \in \mathcal{G}$ and $l \in [L]$;