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

Joel Persson
Spotify, joelpersson@spotify.com

Jurriën Bakker
Booking.com

Dennis Bohle
Booking.com

Stefan Feuerriegel
LMU Munich

Florian von Wangenheim
ETH Zürich

Besides ethical and moral reasons, regulatory and legal frameworks increasingly mandate that machine learning (ML) algorithms should not make errors that systematically disadvantage different groups. Such errors may arise in ML algorithms for predicting treatment effects, leading to group bias in downstream inference and decisions. In this paper, we develop a novel framework for detecting and mitigating group bias in ML prediction of heterogeneous treatment effects (HTEs). We define algorithmic bias as group-wise disparities in expected prediction errors of HTEs. We then show that detection can be cast as estimating and testing how well the predicted HTEs recover the average treatment effects per group. As such, we provide an estimator that is asymptotically normal and present tests with theoretical guarantees. For mitigation, we propose a de-biasing procedure that minimizes bias in new predictions of HTEs. Our framework is general, makes minimal assumption, and is statistically and computationally lightweight. We evaluate its performance in simulation studies and using data from a randomized experiment on the leading travel platform *Booking.com* covering 37 million website sessions. Our methods detect and subsequently mitigate algorithmic bias while providing evidence of a fairness-accuracy trade-off in predicted HTEs. Altogether, our work provides a path forward for addressing bias in ML prediction of treatment effects.

Key words: heterogeneous treatment effect; machine learning; algorithmic fairness; randomized 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 disparities in predictions or 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 Amazon's ML-based hiring tool that was biased against women (Reuters 2018) and Staple's coupon targeting that was biased against low-income customers (The Wall Street Journal 2012). To this end, regulatory and legal frameworks increasingly mandate that ML algorithms should not make systematic errors that disadvantage 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 (HTEs). In tech companies, ML is widely used to predict HTEs for inference and business decision-making across groups, 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 subscription services (Yoganarasimhan et al. 2022).[1] In these examples, the predicted HTEs quantify the relative benefit for a company of providing an individual a treatment (e. g., coupon, free shipping offer, discount). Hence, if the predicted HTEs suffer from algorithmic bias, a company will make incorrect inferences about relative treatment efficacy across 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 contain relative more observations for some groups. This naturally occurs in practice, customer bases are not necessarily equally distributed. For example, travel platforms are primarily used by high-income customers. 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 customers from smaller groups and, thereby, bias across groups. As an example of the latter, group-wise heterogeneity may

[1] The causal inference literature has proposed a vast array of methods aiming at HTE prediction from observational data under sufficient identification conditions (e.g., Athey and Imbens 2016, Wager and Athey 2018, Oprescu et al. 2019). 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), and typically assumes that the training data come from a randomized experiment (Kane et al. 2014).

be unaccounted for by the ML model. In marketing, customers from low-income groups typically have larger HTEs due to a greater elasticity of demand given a treatment. For instance, a "buy one, get one free" promotion is often more effective for customers with less disposable income. If the the ML model is not estimated to capture this heterogeneity, it will estimate the pooled-average treatment effect, leading to biased estimates across groups.

In this paper, we propose a novel framework with theoretical guarantees for detecting and mitigating algorithmic bias in ML for predicting HTEs. We define our notion of algorithmic bias in HTEs as group-wise disparities in the expected prediction errors of the 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 observed outcomes but in treatment effects, which are counterfactual differences in potential outcomes. Our notion does also not require the predicted HTEs themselves to be equal across groups. This is motivated by our considered application areas, where HTEs may systematically vary across groups due to underlying differences between groups without indicating lack of fairness. We thus aim to address the ML model to be equally accurate in predicting systematic heterogeneity in treatment effects across groups, if such heterogeneity truly exists. Our contribution is three-fold.

First, we show that detecting algorithmic bias can be cast as a statistical estimation and inference problem and, for this, present tailored estimators and hypothesis tests with theoretical guarantees. A challenge for the detection is that the prediction errors in the HTEs depend on the true but unobservable HTEs and, therefore, must be estimated. However, estimating the true HTEs is not a viable solution, as the ML model predicting the HTEs is presumed to be inaccurate in the first place. 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 detecting the prediction error per group according to our notion only requires estimating and testing for heterogeneity in how well a (weighted) average of the HTE predictions per group recover the true average treatment effect (ATE) per group. This greatly facilitates the applicability of our framework, as ATEs are considerably easier to identify and estimate than HTEs. Based on our collapsibility result, we provide an estimator that we show is consistent and asymptotically normal. We use this to construct hypothesis tests with theoretical guarantees to detect the true prediction error per group and algorithmic bias across groups.

Second, we propose an optimization procedure to mitigate algorithmic bias when predicting HTEs for new observations. Here, it is tempting to de-bias by simply subtracting the estimated prediction error from the HTE predictions per group. We show that such an approach is generally sub-optimal when predicting HTEs for new observations, as it neglects that the prediction errors are estimates and therefore have variance. As a solution, we propose a procedure that optimally adapts

the amount of de-biasing so as to minimize the expected loss (i. e., statistical risk) of prediction errors in the HTE predictions for new observations from different groups. We draw connections to empirical risk minimization, derive optimal correction factors under common risk functions and provide simple to implement estimators thereof. We show theoretically that, the better we detect the true prediction errors in the HTEs per group, the better our mitigation procedure eliminates bias within and across groups.

Third, we empirically demonstrate our framework using data from a large-scale randomized field experiment on the travel platform *Booking.com*. Here, we aim to mitigate algorithmic bias in an ML model predicting HTEs of a free travel benefit on booking propensity across website sessions from different countries.[2] We highlight three findings. (i) Without mitigation, there is substantial heterogeneity in prediction errors and algorithmic bias across the country of origin of website sessions. In particular, even though the global average prediction error in HTEs is zero, for some countries of origin the prediction 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 the uncertainty-aware optimization of our framework.

Our work has several implications for research and practice. First, detection and mitigation of algorithmic biases in ML models predicting treatment effects should account for that treatment effects are counterfactual estimands and, therefore, the prediction errors are unobservable. Here, we provide a framework for detection and mitigation of algorithmic bias in HTE predictions with theoretical guarantees under minimal assumptions. The latter is important, as causal inference requires assumptions that may be untestable from data. Second, frameworks for addressing algorithmic bias benefit from being general. Our framework is applicable to treatment effects measured in magnitude or in relative terms, to binary or continuous outcomes, and in principle any prediction model 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, detecting and mitigating algorithmic bias in ML models applied at scale requires the methods to be computationally lightweight and easy to implement statistically. Our framework only requires the ability to calculate (weighted) averages, run hypothesis tests, and resample data from historical experiment data. Not only is this a benefit for internal use by, e. g., tech companies, it also facilitates external

[2] *Booking.com* does not collect data on the sensitive attributes race or nationality and, hence, do not use such prohibited information in their ML models. The group variable of interest, country of origin, is not protected grounds under the General Act on Equal Treatment in Netherlands, where *Booking.com* is based.

audits by, e. g., policy organizations, who may be limited in computational power, collecting new data, and modeling efforts. Fourth, and finally, our results point to an accuracy-fairness trade-off; ML models predicting HTEs may lose personalized prediction performance when 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 ML models predicting HTEs should be optimized for prediction performance at the personalized level or be calibrated for fairness at the group level.

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 8 demonstrates the effectiveness of our framework at *Booking.com*. Finally, Section 9 discusses the implications of our work and concludes.

2. Related Work

2.1. Notions of Algorithmic Bias

In the literature, algorithmic bias is often also referred to as "algorithmic discrimination" or "algorithmic fairness" to contextualize the underlying mechanisms and welfare implications, respectively. Several notions of algorithmic bias in ML models have been proposed, which are commonly defined as statistical disparities in predictions or prediction errors across groups that are defined with respect to *sensitive attributes* (Hardt et al. 2016, Chouldechova 2017, Kleinberg et al. 2017, Carey and Wu 2022). In plain words, ML models should then avoid making errors that systematically disadvantage individuals with different sensitive attributes, which are typically indicated by legislative frameworks (Feldman et al. 2015, Barocas and Selbst 2016, Carey and Wu 2022) or business objectives (De-Arteaga et al. 2022).

Most notions proposed in the literature fall into one of three categories defined as independence relations between predictions, outcomes, and groups (see, e. g., Barocas and Selbst 2016, for an overview). At a high level, these so-called representative criteria state that predictions should either be independent of groups, equally well calibrated across groups, or equally accurate across groups. Another and entirely different notion is counterfactual fairness where the prediction should be the same for an individual had (s)he belonged to a different group defined with respect to a sensitive attribute (Kusner et al. 2017). Counterfactual fairness thereby takes a causal inference approach and formalizes fairness as equality in outcomes given a counterfactual change of group membership. Our notion also builds upon causal inference but, in contrast, defines algorithmic bias as a disparity in prediction errors of (counterfactual) treatment effects across groups. Importantly, the

above notions have been developed for supervised ML where the focus is on predicting observable outcomes. Our setting is different as we focus on treatment effects, which are unobservable. Summarizing, the choice of notion should determined by the underlying context (De-Arteaga et al. 2022) and require different approaches for detection and mitigation. We therefore provide a notion of algorithmic bias in HE predictions.

2.2. Mitigating Algorithmic Bias

Several methods have been developed to mitigate algorithmic bias for a specific notions (see Mehrabi et al. (2021) and De-Arteaga et al. (2022) for detailed overviews). In general, existing methods commonly set up a constrained optimization problem of maximizing prediction accuracy subject to fulfilling a fairness notion. Existing methods generally follow three different paradigms: (1) Pre-processing alters the training data so that the notion is ensured after the ML model is trained. For example, the training data can be reweighted so that the predictive covariates are uncorrelated with the sensitive attribute (Kamiran and Calders 2012). Pre-processing methods such as those are generally not amendable to our setting since HTEs may naturally vary across groups due to underlying non-discriminatory group differences and because we intentionally seek to understand how HTEs vary across groups. (2) In-processing incorporates the notion as a constraint to the objective function for training the ML model (Wan et al. 2023). For example, several works incorporate the notion as a regularization term (e.g., Kamishima et al. 2012, Berk et al. 2017, Zafar et al. 2019, Ascarza and Israeli 2022). However, our setting involves counterfactual quantities because of which a simple reformulation through a regularization term is prohibited. (3) Post-processing only consider the predictions of the ML model and adjusts them such that the notion is fulfilled. For instance, one can calibrate an ML model after training toward equalized odd by applying Platt scaling (Pleiss et al. 2017).

Post-processing methods have several advantages in practice (Barocas et al. 2019, Chapter 3) that are relevant for the purpose of our work. First, post-processing methods are model-agnostic and are thus applicable to any ML model of interest. Second, post-processing methods do not tamper with the input data, which reduces the risk of introducing noise or reducing explainability. Third, post-processing methods check whether the predictions actually satisfy the notion, which is a necessary step for use in practice irrespective of the approach taken. Fourth, post-processing methods do not require that a model is re-trained and are thus computationally and operationally lightweight in settings such as digital platforms where training and deploying models is costly with respect to e. g., time, finances, and man-hours. To this end, we build upon post-processing methods and present a tailored framework to mitigate algorithmic bias in ML models for HTE prediction.

2.3. Calibration and Inference with ML

Our work also relates to methodological research in statistics that aim at calibration and statistical inference on ML predictions using auxiliary data. A recent line of work is the prediction-powered inference framework (Angelopoulos et al. 2023), which develops theory and methods for valid inference on predictions from any machine learning algorithm by correcting predictions with their prediction errors on hold-out experimental data. A related line of work stems from design-based inference and survey sampling, wherein the goal is to improve the estimation of and inference on population parameters by leveraging randomization or access to auxiliary population information. Most related to our work in this literature is recent research considering calibration and inference on estimates derived from ML predictions; see (see e. g. Breidt and Opsomer 2017, for an overview). Our work differs from prior research in these areas by considering the calibration and statistical inference on predictions of treatment effects aggregated to the group-level with the aim of detecting and mitigating disparities in prediction errors of treatment effects within and between groups.

2.4. Our Contributions

Our work aims to address gaps in the literature on HTE prediction and algorithmic bias. Most similar to our work are Ascarza and Israeli (2022), Huang and Ascarza (2023), and Leng and Dimmery (2024). The first proposes a tailored splitting criterion that ensures HTE predictions from deicsion trees are personalized while satisfying group and individual fairness, the second proposes how to correct a prediction model of HTEs for bias introduced by applying privacy-preserving methods, and the third proposes a method to calibrate HTE predictions using regression estimates from a randomized experiment. However, all three have notable differences to our work. The method of Ascarza and Israeli (2022) is an in-processing procedure, whereas we consider post-processing. Similarly, Huang and Ascarza (2023) differs from our work in that they use in-processing and target bias from injecting privacy-preserving noise in the predictions. Finally, the method of Leng and Dimmery (2024) differs from ours in that it does not aim at addressing algorithmic bias, i. e., *disparities* in predictions errors of treatment effects *between* groups.

3. Reasons for Algorithmic Bias in HTE Predictions

In the following, we discuss potential reasons why ML models predicting HTEs may exhibit algorithmic bias with respect to systematic disparities in prediction errors across different groups. As a running example, we motivate our discussion from the setting of our partner company *Booking.com* that seeks to avoid such disparities for website session with different country of origin. We categorize potential sources along the ML pipeline, namely, (1) bias in the data and (2) bias from modeling.

Bias in the Data

One reason for disparities in prediction errors of HTEs may be that the training data is imbalanced. One such imbalance relates to disparate sample sizes per group, which is also referred to as representation bias in the literature (Mehrabi et al. 2021). Such representation bias 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. Due to representation bias, the data contain more observations for high-income customers, which helps to learn more complex and thus more precise relationships between input data and outcomes with greater predictive power. Conversely, by having fewer observations, predictions for the underrepresented group will naturally be subject to more epistemic uncertainty and thus have a higher variance. This leads to greater prediction errors from underrepresented groups. Notwithstanding, representation bias is a known challenge in data-driven modeling, including standard ML with the aim of predicting outcomes (Campolo et al. 2017, Chouldechova and Roth 2020) and treatment effect estimation in general (Simester et al. 2020b). As such, it may apply to our setting due to that some nationalities travel relatively more, leading to disparities in prediction errors in HTEs across customers' countries of origin.

A second type of bias can originate from disparities in the data distributions across groups (also known as covariate shift). This can occur on travel platforms, as browsing sessions from some countries will be more heterogeneous than sessions from other countries in terms of the variation in the features and booking rates associated with a session. For instance, it is likely that large countries like the U.S. have a more heterogeneous customer base (with both rich and poor travelers) than small countries such as Lichtenstein, Luxembourg, and Monaco (with mostly rich travelers). If this is the case, then the conditional distribution of outcomes given predictors will be different across countries (which is also known as covariate shift). This conditional distribution provides the identifying variation in the data for estimating the HTE. Thus, differences in the conditional distribution across groups can lead to systematic disparities in prediction errors of HTEs across groups.

A third source of bias in training data is due to differences in the quality of the data across groups. For example, the training data from certain groups may have more missing variables or other measurement errors. This will naturally lead to disparities in prediction errors across groups, as then the data is less informative of the ground truth HTEs for certain groups. While missing values or measurement errors are rare in online marketing, it may be the case in offline marketing such as brick-and-mortar retailing where scanner data or checkout data is often only observed for those who purchased, but not for those who browsed the store without purchasing. Different customer groups may have different propensities of missing data (e. g., prospective customers with

less disposable income are more likely to browse the store without purchasing), potentially leading the ML model to have greater prediction errors in HTEs of offline marketing treatments such as in-store coupons for some customer groups. Even though our empirical application does not have missing data, our framework still allows for detecting algorithmic bias in the presence of missing data.

Bias from Modeling

Bias may also occur from modeling. Unless explicitly accounted for, an ML model will be optimized for overall prediction performance and thereby emphasize prediction performance for larger groups at the expense of a lower prediction performance for smaller groups (Chouldechova and Roth 2020). For travel platforms such as *Booking.com*, the data will feature relatively more website sessions from big, western, and economically developed countries such the U.S. simply because those are the countries where the platform tend to have the greatest brand recognition, inventory, and travel-able users. Hence, unless the training data is intentionally constrained to have equal sample sizes across groups, which may hurt overall prediction performance, a trained ML model may have disparate prediction performance for website sessions from different countries.

Another reason for bias from modeling is a form of omitted variable bias (Mehrabi et al. 2021) where the omission has a disparate effect on the prediction performance across groups. This can occur because variables that predict heterogeneity in treatment effects cannot be measured, are prohibited for legal reasons, or because they were dropped by regularization. As for legal reasons, many companies in the European Union, including *Booking.com* avoid the use of information about protected attributes in their model due to new privacy regulations such as those postulated by the General Data Protection Regulation. As for the regularization case, a variable indicating group membership may be dropped if, e.g., a sparsity constraint is used during model training. Irrespective of the reason, an ML model may not adequately capture heterogeneity in treatment effects between groups if the group variable is omitted from the model; instead, the model estimates will be biased towards the average treatment effect across groups for which, again, the majority groups contribute relative more. This is also known as aggregation bias (Mehrabi et al. 2021).[3]

[3] As an example of the latter, customers from low-income groups will in many empirical contexts have larger HTEs 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 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 incremental sales and low-income customers not receiving a beneficial offer, thus reducing both company profits and customer welfare. Importantly, even a model that does account for group-wise treatment effect heterogeneity may exhibit disparate prediction errors in treatment effects. This is because the predicted treatment effects per group may still not be equally accurate if, e.g., the sample sizes or structural treatment effect function vary across groups. Thus, simply including a variable of group membership in the prediction model (and ensuring that it is not dropped) does not guarantee that algorithmic bias is eliminated.

In sum, there are various potential sources of algorithmic bias, and these may interact in non-linear and unknown ways, leading to complex forms of algorithmic bias that may be highly difficult to characterize in practice. We thus design our framework to detect and mitigate systematic disparities in prediction errors regardless of the underlying source of bias to ensure that it is broadly applicable.

4. Setup of Framework

4.1. Preliminaries

We focus on a company that aims to remove bias in HTEs within and between groups due to some treatment (e.g., a coupon). The company has access to historical data collected through a randomized experiment designed for estimating the treatment effect and is able to collect additional experimental data for evaluation. Furthermore, there is a given sensitive attribute (e.g., country of origin) for which algorithmic bias should be removed. In practice, the sensitive attribute is often determined by regulatory frameworks, ethical concerns, or business considerations (De-Arteaga et al. 2022).

Let $i = 1, \dots, N$, $N \in \mathbb{N}$, refer to the units of analysis (e. g., individuals, customers, website sessions). The sensitive attribute partitions individuals into known groups $g \in \mathcal{G}$ each of size $n_g < N$, where $N = \sum_g n_g$ and $|\mathcal{G}| \geq 2$. Formally, the groups are defined by

$$\mathcal{G} \coloneqq \left\{ g_k \subseteq \{i, \dots, N\} \colon k \in \{1, \dots, |\mathcal{G}|\},\ |\mathcal{G}| \in \mathbb{N}, g_k \neq \emptyset,\ g_k \cap g_j = \emptyset \text{ for } k \neq j,\ \bigcup_k g_k = \{i, \dots, N\} \right\}, \tag{1}$$

i. e., a finite set of non-empty, pairwise disjunct subsets of the set of individuals. Thus, each individual belongs to exactly one group $g \in \mathcal{G}$. For each group, we let $(\mathbf{X}_i, T_i, Y_i) \overset{i.i.d.}{\sim} P_g$ be random variables for unit $i = 1, \dots, n_g$, where $\mathbf{X} \in \mathcal{X}_i \subseteq \mathbb{R}^p$ is a vector of pre-treatment covariates, $T_i \in \{0, 1\}$ is a binary treatment assignment, $Y_i \in \mathcal{Y} \subset \mathbb{R}$ is the outcome that may be continuous or binary, and P_g is an unknown joint distribution. The distributions $\{P_g\}_{g \in \mathcal{G}}$ can be arbitrarily different across groups. We assume without loss of generality that higher values of outcomes are preferred.

We use the potential outcomes framework (Rubin 1974). Let $Y_i(t)$ denote the potential outcome under treatment assignment $t = 0, 1$. Due to the fundamental problem of causal inference (Holland 1986), we only observe the potential outcome corresponding to the assigned treatment and not the counterfactual outcome had the other treatment been assigned. The observed outcome is thus given by

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

In the following, we will drop the unit subscript i, unless necessary.

4.2. Heterogeneous Treatment Effect

The HTE measures heterogeneity in an average treatment effect (ATE) with respect to the pre-treatment covariates. We let $\tau(\mathbf{X})$ denote the HTE and let τ denote an ATE obtained by marginalizing out heterogeneity. A HTE and ATE can be measured in several ways. In this work, we focus on relative HTE measure, which is commonly defined in unit-less terms as the ratio

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

The relative HTE measures heterogeneity in the expected percentage gain of treatment compared to control explained by covariates in the relative ATE $\tau = \mathbb{E}[Y(1)]/\mathbb{E}[Y(0)]$. Our framework and all of our result directly extend to the common magnitude measure of HTEs $\tau(\mathbf{x}) \coloneqq \mathbb{E}[Y(1) - Y(0) \mid \mathbf{X} = \mathbf{x}]$ and its corresponding ATE $\tau = \mathbb{E}[Y(1) - Y(0)]$. Nonetheless, we focus on the relative HTE as this is the measure of HTEs used in our empirical application and because it better highlights challenges in our framework. To save space, we will thus relegate results for the magnitude measure to the appendix.

The relative HTE apply to outcomes that are continuous (i.e., $\mathcal{Y} = \mathbb{R}$) or binary (i.e., $\mathcal{Y} = \{0, 1\}$). In marketing, an example of the former is the treatment effect on sales per customer, while an example of the latter is the treatment effect on whether a customer converted or not. For 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 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 requires $\mathbb{P}[Y(0) = 1 \mid \mathbf{X} = \mathbf{x}] > 0$ to be well-defined. For binary outcomes, relative HTE corresponds to a conditional estimand of causal relative risk (Hernán and Robins 2006), also known as uplift or lift factor.[4]

Due to the fundamental problem of causal inference (see Eq. (2)), the HTE is a counterfactual quantity that cannot be observed. Thus, the HTE must be obtained with a prediction rule (e.g., ML model) f trained on historical data, which maps a unit's covariates $\mathbf{X}$ to a prediction $f(\mathbf{X}) = \widehat{\tau}^f(\mathbf{X})$ of either the magnitude HTE or the relative HTE. In our framework, we make minimal assumptions on the prediction rule. We only assume that it fulfills standard regularity assumption and that it was trained on data with randomized treatment assignment, which is considered best practice in causal inference and is the default approach at *Booking.com*. In principle, the prediction rule can be chosen arbitrarily as long as it provides predictions for the HTE measure of interest. For example, one could use ML models such as an ensemble of uplift models (Sołtys et al. 2015), an uplift response transformation model (Jaskowski and Jaroszewicz 2012, Gubela et al. 2020), double

[4] For binary outcomes, the magnitude HTE becomes $\tau(\mathbf{x}) = \mathbb{P}[Y(1) = 1 \mid \mathbf{X} = \mathbf{x}] - \mathbb{P}[(0) = 1 \mid \mathbf{X} = \mathbf{x}]$, which corresponds to a conditional estimand of causal risk difference.

ML (Chernozhukov et al. 2018), or a meta-learning algorithm for HTE prediction such as, e.g., T-learner, S-learner, X-learner, or DR-learner (Künzel et al. 2019, Kennedy 2020)). The benefit of ML models is their data-driven and non-parametric form, which aids in prediction accuracy and avoids the need to imposeparametric assumptions. Notwithstanding, our frameworks allows the use of parametric or heuristic, rule-based models.

4.3. Formalizing Bias in HTE Predictions

We now formalize our notion of algorithmic bias based on systematic disparities in the errors of the HTE prediction across groups.

An immediate challenge is that we aim to address systematic disparities in prediction errors at the group level, but only have access to predictions at the personalized level. We thereby need an approach for aggregating HTE predictions to the group level. To formalize this idea, we first define collapsibility of treatment effect measures.

Definition 1 (Collapsibility (Colnet et al. 2023)). *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 $\tau(\mathbf{X})$, it holds that*

$$\mathbb{E}\big[W\tau(\mathbf{X})\big] = \tau \text{ with } w \geq 0, \text{ and } \mathbb{E}\big[W\big] = 1. \tag{4}$$

The definition states that an HTE measure is collapsible if it is possible to recover its corresponding ATE via a weighted average, where the weights depend on the density of the pre-treatment covariates $\mathbf{X}$ and the joint distribution of the controls, $P(\mathbf{X}, Y(0))$. The magnitude HTE and the relative HTE we consider in this work are both collapsible measures, but other common measures for treatment effects such as the odds ratio and the log-odds ratio are not (Colnet et al. 2023).

Using collapsibility, we can formalize group-level errors in predicted HTEs. Fix f to be an estimated prediction rule of a HTEs $\tau(\mathbf{X})$ and let W be defined as in Def. 1. The *true prediction error* (TPE) of f for group g is

$$b_g := \mathbb{E}_{P_g}\big[W\big(\widehat{\tau}^f(\mathbf{X}) - \tau(\mathbf{X})\big)\big], \tag{5}$$

We make five remarks on our definition. First, we condition on a prediction model of a HTE measure to mimic the real-world setting where the model is a given object to audit. Second, b_g depends on prediction rule f via its function class, estimation method, and training data. However, as f is fixed, we omit this dependence from our notation. Second, we denote the true error with a lowercase letter b_g to signify that it is a constant, not a random variable. This follows that conditional on $\mathbf{X}$

the true HTE $\tau(\mathbf{X})$ is a constant and so is the predicted HTE $\widehat{\tau}^f(\mathbf{X})$ for a given prediction model. Fourth, we define b_g with respect to the population and not a sample. The reason is that we wish to detect systematic bias in the prediction model in general, not just for a particular sample. Fifth, if we view the HTE $\tau(\mathbf{X})$ as a population parameter and the HTE prediction $\widehat{\tau}^f(\mathbf{X})$ as an estimate thereof, the true error is analogous to the canonical definition of bias of an estimator. We will thus use the terms true error and statistical bias interchangeably, where convenient.

For relative HTEs, 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).[5] Hence,

$$b_g = \mathbb{E}_{P_g}\left[\frac{\mathbb{E}[Y(0) \mid \mathbf{X}]}{\mathbb{E}[Y(0)]}\left\{\widehat{\tau}^f(\mathbf{X}) - \tau(\mathbf{X})\right\}\right]. \tag{6}$$

In short, the weighting ensures that HTE collapse to an ATE when the potential outcomes under no treatment are heterogeneous with respect to pre-treatment covariates, which is the general case in practice.

Key to our framework is that collapsibility enables us to decompose the statistical bias in the HTE predictions for a group to simply a difference between the predicted and the true ATE:

$$\begin{aligned} b_g &= \mathbb{E}_{P_g}\left[W\left(\widehat{\tau}^f(\mathbf{X}) - \tau(\mathbf{X})\right)\right] && (7)\\ &= \mathbb{E}_{P_g}\left[W\widehat{\tau}^f(\mathbf{X}) - W\tau(\mathbf{X})\right] && (8)\\ &= \mathbb{E}_{P_g}\left[W\widehat{\tau}^f(\mathbf{X})\right] - \mathbb{E}_{P_g}\left[W\tau(\mathbf{X})\right] && (9)\\ &= \widehat{\tau}^f_g - \tau_g, && (10) \end{aligned}$$

where the third equality follows by linearity of expectations and the fourth by collapsibility. Here, $\widehat{\tau}^f_g$ and τ^f_g is the predicted and true group ATE for group g, respectively. The above shows that after collapsing the HTE, the statistical bias b_g in the HTE model per group only depends on a single difference between two group-level quantities. In contrast, if we would evaluate statistical bias in terms of HTEs themselves as in Eq. (5), the statistical bias would depend on twice as many as differences as the number of unique covariate combinations in the data. Not only does the latter suffer from a combinatorial explosion in typical ML applications with high-dimensional covariates, it is also made challenging due to that it would require estimating the true HTEs, which is prohibitive since we are our prediction model thereof is presumed to be biased in the first place.

We use the above definition of true error to define our notion of algorithmic bias.

[5] For the magnitude HTE, the weight is $W = 1$ (Greenland et al. 1999). Thus, weighting is not needed and the true error coincides with statistical bias of an estimator.

Definition 2 (Algorithmic bias in HTE predictions). *Let b_{-g} refer to the true error across all observations except those from group g. A prediction rule f of an HTE measure has algorithmic bias against group $g \in \mathcal{G}$ if*

$$\left|b_g - b_{-g}\right| > 0. \tag{11}$$

Put simply, a prediction rule of HTEs has an algorithmic bias against a group if its group-level errors towards that group are, on average, larger than for the remaining observations across the other groups.

We make two remarks on our notion of algorithmic bias. First, it 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). As a result, our framework is thus broadly applicable in practice. Second, if the prediction rule is used to inform decisions, then fulfilling notion addresses the mechanism (i.e., the HTE predictions) producing the disparities in decisions across groups. However, it will *not* enforce that the HTE predictions themselves must be equal across groups. Thereby, the notion respects that, in the application areas we consider, different groups may respond differently to the same treatment (e. g., low-income groups may respond better to an offer than high-income groups) and, therefore, group-wise heterogeneity treatment effects is not necessarily evidence of algorithmic bias (e. g., following the previous example, many people would likely not object to hihg-income groups receiving less offers than low-income groups). Similarly, our notion correspond the principle that if the predictions are equally inaccurate for all groups, it is not necessarily a fairness problem, but rather a prediction or inference problem. In Appendix A, we discuss our notion to existing notions developed in the literature.

Remark 1. *An optimal prediction model in the collapsilibity sense $\widehat{\tau}_g^f = \tau_g$ has no statistical or algorithmic bias according for any group.*[6]

The statement hold by definition, and illustrates that both types of bias can be addressed in two ways: (i) by estimating the ML model to not make prediction errors at the group level in the first place, or (ii) by correcting its errors afterward. The problem with (i) is that, because the prediction errors depend on the true HTEs, we do not know whether an ML model estimated not to be biased actually is not biased. Moreover, the HTE predictions are at the personalized level whereas the bias is at the group level, thereby raising challenges for implementing such an approach. This motivates approach (ii), by detecting whether the ML model makes prediction errors per group and, if so, by how much such that they can be corrected. We next present our approach to this.

[6] A similar observation was made for algorithmic bias in predicting outcomes (Corbett-Davies et al. 2017), whereas we extend the idea to predicting HTEs.

5. Detection

5.1. Plug-in Estimator with Testable Implications

Based on the collapsibility result, we propose to estimate the prediction error per group by collapsing the predicted HTEs $\widehat{\tau}^f(\mathbf{X})$ within a group to a predicted ATE $\widehat{\tau}^f_g$ and taking its difference to an estimate $\widehat{\tau}_g$ of the true ATE for the group. The *estimated prediction error* (EPE) for group g by the prediction model f is thereby given by

$$\widehat{B}_g = \widehat{\tau}^f_g - \widehat{\tau}_g. \tag{12}$$

Recall that our notion of algorithmic bias is defined as a disparity in the true but unobservable prediction errors (cf. Def. 2), as this is what we aim to detect. The plug-in estimator $\widehat{B}_g$ leads to an empirical version of our notion with testable implications, given by

$$\left|\widehat{B}_g - \widehat{B}_{-g}\right| \geq \varepsilon. \tag{13}$$

where $\varepsilon \in \mathbb{R}^+$ is a threshold that accounts for that the estimated disparity $|\widehat{B}_g - \widehat{B}_{-g}|$ will generally not exactly equal zero even when the true disparity $|b_g - b_{-g}|$ does.

In principle, the plug-in estimator $\widehat{B}_g$ applies both experimental and observational settings, where for the latter, the true group ATE is estimated via e.g. regression discontinuity or instrumental variables techniques. In this paper, we focus on the gold-standard approach of using a randomized experiment to estimate the ATE. This is the the default approach in the application areas we consider (i.e., digital platforms and e-commerce companies) and facilitates our exposition by removing differences in estimands, identification, and estimation that arise when moving from experimental to observational settings. The next section explains our approach to identification and estimation based on data from a randomized experiment.

5.2. Identification and Estimation

We proceed in two steps. We first provide identification conditions and then present estimators.

Assumption 1. *For all $g \in \mathcal{G}$, $\widehat{\tau}_g$ and $\widehat{\tau}^f_g$ are calculated on an estimation set $\mathcal{E}_g$ and prediction set $\mathcal{P}_g$, respectively, that are independent splits of the data $\mathbf{D}_g \coloneqq (\mathbf{X}_i, T_i, Y_i)_{i=1}^{n_g}$ on group g.*

The assumption implies that estimated and predicted ATEs per group are independent. This facilitates statistical inference on the prediction error by avoiding inferential problem in using the same data twice and by making the covariance between the estimated and predicted group ATE zero (Breidt and Opsomer 2017, Angelopoulos et al. 2023). Our framework is agnostic to how the estimation and prediction sets are constructed. In practice, the prediction and estimation sets may be random splits of the same set of data of a group, or they stem from different datasets on a group.

For a digital platform, the prediction set may be log data containing the HTEs of the prediction model and the estimation set may be data from a portion of incoming traffic randomly allocated to an A/B test. Alternatively, both the prediction set and the estimation set could be random partitions of A/B test data for which the prediction set includes the HTE predictions.

We make the following assumptions that are standard in the potential outcomes framework for causal inference (Imbens and Rubin 2015, Hernan and Robins 2023).

ASSUMPTION 2. *For all $i \in g$ and $g \in \mathcal{G}$, we assume: (i)* Consistency*: $Y_i = Y_i(T)$. (ii)* No interference*: $T_i \perp\!\!\!\perp Y_j(T_j)$ for all $j \neq i$. (iii)* Strong exchangeability*: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i$. (iv)* Positivity*: $\mathbb{P}(T_i = t \mid \mathbf{X}_i = \mathbf{x}) > 0$ for $t = 0, 1$ and all $\mathbf{x}$ such that $p_{\mathbf{X}}(\mathbf{x}) > 0$.*

Consistency connects the potential outcomes to the observed outcomes and follows from $Y(T) = Y(1)T + Y(0)(1-T)$. No interference means that the potential outcomes of a unit do not depend on the treatment assignment to any other unit. Strong exchangeability states that a unit's potential outcomes are independent of its treatment assignment. Positivity states that all units had a positive probability of being assigned to treatment. This is always true for randomized experiments.

Given the identifying assumptions, we now consider estimation of the true error b_g in the group ATE. Recall that our plug-in estimator $\widehat{B}_g$ consists of two components. Hence, *statistically unbiased* estimation of the true prediction error b_g, i.e., that $\mathbb{E}[\widehat{B}_g] = b_g$, requires that both components – i.e., the predicted group ATE $\widehat{\tau}_g^f$ and the estimated true group ATE $\widehat{\tau}_g$ – are statistically unbiased of what they are *supposed to* identify. To clarify, note that

$$\mathbb{E}[\widehat{B}_g] = \mathbb{E}[\widehat{\tau}_g^f - \widehat{\tau}_g] = \mathbb{E}[\widehat{\tau}_g^f] - \mathbb{E}[\widehat{\tau}_g]. \tag{14}$$

Thus, $\mathbb{E}[\widehat{B}_g] = b_g$ will only reliably hold if (i) $\mathbb{E}\{\mathbb{E}[\widehat{W}\widehat{\tau}_g^f(\mathbf{X})]\} = \widehat{\tau}_g^f$, and (ii) $\mathbb{E}[\widehat{\tau}_g] = \tau_g$. Here, (i) requires correct collapsibility, whereas (ii) requires unbiased estimation. Thus, in the following, we present how to estimate both expectations. Estimators for magnitude the magnitude HTE and its associated ATE are provided in Appendix B

5.2.1. True group ATE τ_g. A key strength of our framework is that estimating the true error in the predicted HTEs does not require estimating the true HTEs, but only the true group ATEs. We can obtain a consistent estimate of the true group ATE by calculating a contrast in average observed outcomes between the treated and controls in a randomized experiment, where the randomization ensures that the identification assumptions are met. Let $\mathcal{E}_g$ be the estimation set for group g. The ratio-of-means

$$\widehat{\tau}_g = \frac{\left(\sum_{i \in \mathcal{E}_g} T_i\right)^{-1} \sum_{i \in \mathcal{E}_g} Y_i T_i}{\left(\sum_{i \in \mathcal{E}_g} (1 - T_i)\right)^{-1} \sum_{i \in \mathcal{E}_g} Y_i (1 - T_i)} \tag{15}$$

is a consistent estimator of the true ATE for a group that applies both when outcomes are binary or when they are continuous. The estimator has slight finite-sample bias vanishing asymptotically (Jewell 1986).

For small samples, one may wish to use more efficient estimators than contrasts in sample means. Examples of such estimators are parametric or non-parametric regression adjustment, inverse probability weighting, and augmented inverse probability weighting techniques (see, e. g., Imbens and Rubin 2015, Kennedy 2020). In randomized experiments, these estimators increase precision in the estimated average treatment effect by controlling for variation in the outcome not explained by treatment assignment.[7] Most regression-based estimators in the literature are designed for treatment effects measured in magnitude, so for relative measure one can either adapt existing estimators or use estimators designed for this case (see, e. g., Jewell 1986, Zhang and Kai 1998, Marschner and Gillett 2012, Richardson et al. 2017, for such estimators).

5.2.2. Predicted group ATE τ_g^f. Eq. (6) shows that collapsing relative HTEs to a relative ATE for a group involves weighting the relative HTEs with $W = \mathbb{E}[Y(0) \mid \mathbf{X}]/\mathbb{E}[Y(0)]$. This suggests two approaches to estimating the predicted relative ATE from the predicted relative HTEs. The first approach is to assume $Y(0) \perp\!\!\!\perp \mathbf{X}$, meaning that potential outcomes under no treatment are independent of pre-treatment covariates. Then $\mathbb{E}[Y(0) \mid \mathbf{X}] = \mathbb{E}[Y(0)]$ and so Eq. (6) reduces to a non-weighted contrast. Hence, if the assumption holds, the unweighted sample average of the predicted relative HTEs is an unbiased estimator of the predicted relative ATE. However, the assumption is unlikely to hold in practice, since it implies that heterogeneity in treatment effects with respect to covariates only stem from the covariates effect on heterogeneity in treated outcomes. The more realistic approach is thus to estimate the weights to directly target Eq. (6). Then, an unbiased estimator is the weighted sample average

$$\widehat{\tau}_g^f = \frac{1}{|\mathcal{P}_g|} \sum_{i \in \mathcal{P}_g} \widehat{w}_i f(\mathbf{X}_i) \quad \text{where} \quad \widehat{w}_i = \widehat{Y}_i(0) \times \left(\frac{1}{\sum_{i \in \mathcal{P}_g} (1 - T_i)} \sum_{i \in \mathcal{P}_g} (1 - T_i) Y_i \right)^{-1} \tag{16}$$

are plug-in estimates of the population weights $W = \mathbb{E}[Y(0) \mid \mathbf{X}]/\mathbb{E}[Y(0)]$. Here, $\widehat{Y}_i(0)$ is the predicted potential outcome of an individual as a function of the pre-treatment covariates from a suitable prediction model of outcomes.

The above estimator is applicable to outcomes that are continuous or binary. We now consider the special case that outcomes are binary and that we only have data for the positive outcomes ($Y = 1$). Examples of this are ubiquitous. In offline settings such as retailing, loyalty card and

[7] For instance, regression adjustment estimators use that, for $t = 0, 1$, 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 $\mathbf{X}$.

household scanner data often only contain data conditional on a purchase. In online setting such as digital platforms, conversion rates are typically very low, and so collecting or retaining data on only those who converted may be cost-effective. Irrespective of the reason, using data only on those with positive outcomes requires an additional assumption for collapsing the predicted HTEs to a predicted ATE for the group as a whole.

ASSUMPTION 3. *Let* $\widehat{\tau}^f(\mathbf{X}) = \widehat{\mathbb{P}}[Y(1) = 1 \mid \mathbf{X} = \mathbf{x}]/\widehat{\mathbb{P}}[Y(0) = 1 \mid \mathbf{X} = \mathbf{x}]$ *be the predicted relative HTE. Within each group, the predicted relative HTEs are independent of the outcome:*

$$\widehat{\tau}^f(\mathbf{X}) \perp\!\!\!\perp Y \mid G \quad \text{for all} \quad \mathbf{x} \in \mathcal{X}, G \in \mathcal{G}, \tag{17}$$

Under the assumption, estimating the predicted relative ATE only requires estimating two expectations: (i) the predicted ATE among the treated with positive outcomes, and (ii) the predicted ATE among the controls with positive outcomes. Both ATEs are then weighted to ensure the correct collapsibility to the overall ATE for a group. The predicted ATE among the treated in a group with positive outcomes is given by

$$\widehat{\psi}_g = \frac{1}{\sum_{i \in \mathcal{P}_g} T_i Y_i} \sum_{i \in \mathcal{P}_g} T_i Y_i f(\mathbf{X}_i). \tag{18}$$

The predicted ATE among the non-treated in a group with positive outcomes is analogously

$$\widehat{\lambda}_g = \frac{1}{\sum_{i \in \mathcal{P}_g} (1 - T_i) Y_i} \sum_{i \in \mathcal{P}_g} (1 - T_i) Y_i f(\mathbf{X}_i). \tag{19}$$

The predicted relative ATE for a group as a whole is then given by the following weighting formula:

$$\widehat{\tau}_g^f = \frac{\sum_{i \in \mathcal{P}_g} (1 - T_i) Y_i \widehat{\lambda}_g^2 + \sum_{i \in \mathcal{P}_g} T_i Y_i \widehat{\psi}_g}{\sum_{i \in \mathcal{P}_g} (1 - T_i) Y_i \widehat{\lambda}_g + \sum_{i \in \mathcal{P}_g} T_i Y_i}. \tag{20}$$

See Appendix D for details.

5.2.3. Choosing an Estimator. We thus have three estimators for collapsing HTEs to a ATE: (i) the sample average estimator in Eq. (56), (ii) the weighted average estimator in Eq. (16), and (iii) our special case estimator in Eq. (20). The choice between them boils down to ease of use and a bias-variance trade-off. On the one hand, the (non-weighted) 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 obtained by collapsing will be biased. On the other hand, the weighted sample average estimator in Eq. (16) and our estimator for the special case estimator in Eq. (20) are both unbiased irrespective of whether the assumption holds. Here, the former estimator will have a greater variance due to the additional estimation of the weights, whereas the latter estimator will have a larger variance due to the aggregation of two averages estimated on fewer observations. To choose an estimator, one should first judge the plausibility of the assumption of independence between outcomes and covariates.

5.3. Theoretical Results

Our main theoretical result for detection is that, if τ_g^f is a predicted ATE obtained by correctly collapsing the HTE predictions of a group and τ_g is a consistent ATE estimate for the same group, then the plug-in estimate of the prediction error in Eq. (12) is asymptotically normal with mean equal to the true error given by Eq. (5).

THEOREM 1. *Let $\widehat{B}_g$ be the estimated error for a group g obtained with any of our estimators and for either the magnitude HTE measure or the relative HTE measure. Let $Var[\widehat{\tau}_g^f], Var[\widehat{\tau}_g] < \infty$ so that $\sigma_g^2 := Var(\widehat{B}_g) = \mathbb{E}[(\widehat{B}_g - \mathbb{E}[\widehat{B}_g])^2] \in \mathbb{R}^+$. Then*

$$\sqrt{n_g}\,\widehat{B}_g \xrightarrow{d} \mathcal{N}\Big(b_g, \sigma_g^2\Big) \quad as \quad n_g \to \infty. \tag{21}$$

Proof. See Appendix C.1. □

Theorem 1 implies that, in finite samples, our bias estimate equals the true bias up to an error that vanishes with increasing group sample size, i. e., $\widehat{B}_g = b_g + o_p(n_g^{-1/2})$. Whereas central limit theorems typically concern the asymptotic convergence of an estimate to its limiting distribution, our theorem concerns the asymptotic convergence in distribution between a presumably biased estimate (the predicted ATE of the ML model) and a consistent estimate (the predicted true ATE) of the same estimand (the ATE). Our theorem may thus be of independent interest to other problems in statistical theory beyond the scope of this paper.

REMARK 2. *It follows immediately from Theorem 1 that*

$$\sqrt{n_g}\Big(\widehat{B}_g - b_g\Big) \xrightarrow{d} \mathcal{N}(0, \sigma_g^2) \quad as \quad n_g \to \infty,. \tag{22}$$

and, similarly, by Slutsky's theorem,

$$\sqrt{n_g}\Big(\widehat{\tau}_g^f - \widehat{B}_g\Big) \xrightarrow{d} \mathcal{N}(\tau_g, \sigma_{\widehat{\tau}_g}^2) \quad as \quad n_g \to \infty. \tag{23}$$

The first result in the remark provides a theoretical guarantee for inference methods based on normal approximation to detect algorithmic bias. The second result in the remark suggests a debiasing approach by which one subtracts the estimated error from the mean HTE prediction of a group as a means to recover the true error.

5.4. Inference

Using our theoretical result, we now present hypothesis tests that can be used to detect statistical bias and algorithmic bias according to our notion. Confidence intervals with correct coverage probability can also be constructed in a straightforward manner using normal approximation.[8]

[8] For the statistical bias, an asymptotically valid confidence interval for with coverage probability of $1-\alpha$, $\alpha \in [0,1]$ is simply given by $\widehat{B}_g \pm t_{\alpha/2}\widehat{\sigma}_g$ whereas for the algorithmic bias it is given by $\widehat{B}_g - \widehat{B}_{-g} \pm t_{\alpha/2}\widehat{\sigma}_{g,-g}$.

We first wish to test for the presence of true prediction errors per group. The null hypothesis to test is

$$H_0 \colon b_g = 0 \quad \text{vs.} \quad b_g \neq 0 \tag{24}$$

It follows immediately by Remark 2 that we can construct a t-distributed test statistic as

$$t = \frac{\widehat{B}_g}{\sigma_g} \sim t_{n_g - 2}. \tag{25}$$

where $\sigma_{g,-g} := \sqrt{\mathrm{Var}(\widehat{B}_g)}$. We thus reject the null hypothesis if $|t|$ exceeds the critical value $t_{n_g-2}(\alpha/2)$ at a pre-specified confidence level α. Thus, a simple t-test on the estimated error per group allows us to detect whether, on average, the HTE predictions are statistically biased relative to the true HTE for a group.

We then wish to test for the presence of algorithmic bias towards a group, i.e., the null hypothesis that the true error of a group deviates from that of the rest,

$$H_0 \colon b_g = b_{-g} \quad \text{vs.} \quad b_g \neq b_{-g}. \tag{26}$$

Again, by Remark 2,

$$t = \frac{\widehat{B}_g - \widehat{B}_{-g}}{\sigma_{g,-g}} \sim t_{n_g-2}, \tag{27}$$

where $\sigma_{g,-g} := \sqrt{\mathrm{Var}(\widehat{B}_g - \widehat{B}_{-g})}$. A rejection of the null implies that we have detected algorithmic bias for the group, where the size of the bias is the estimate $\widehat{B}_g - \widehat{B}_{-g})$. A failure to reject means that no bias is detected.

In practice, one can apply a correction for multiple testing if the sensitive attribute defines multiple groups (i.e., $|\mathcal{G}| > 2$) and it is deemed important control to the family-wise error rate. A simple approach is to use the classical Bonferroni correction (Dunn 1961), which simply scales the confidence level relative to the number of tests. For $\mathcal{G}$ groups, the confidence level for rejecting the null hypothesis then becomes $\alpha/|\mathcal{G}|$.

We finally state how the threshold ε that determines whether a difference in EPEs constitutes algorithmic bias is automatically set in our hypothesis tests.

PROPOSITION 1. *Let ε be defined as in Def. 2 and fix a confidence level $\alpha \in [0, 1]$. We yield*

$$\varepsilon \sim t_{n_g-2}(\alpha/2) \times \sigma_{g,-g}. \tag{28}$$

Proof. The proof follows immediately by Eq. (13) and the construction of the test in Eq. (26). □

Proposition 1 shows that the threshold for establishing algorithmic bias according to Eq. (11) is determined by the critical value of the test-statistic given a chosen confidence level and the standard error of the disparities in EPEs. Specifically, the lower we set the required confidence level, and the greater the variance in our EPEs, the larger the threshold. In practice, the sample size of the data used is fixed and so is the variance of the EPEs. Thus, in most settings, the confidence level is the parameter that a decision-maker can set to alter the threshold at which disparities in EPEs are labeled as evidence of algorithmic bias. Here, the decision-maker can either fix the confidence level at a common standard (e. g., $\alpha = 0.05$) or set it to optimally balance the costs of making type-I errors vs. type-II errors (i. e., rejecting a true null vs. failing to reject a false null) in the specific application at hand.

6. Mitigation

We now consider how to mitigate the bias in HTE predictions for new observations but, first, we outline the setup. At the time of mitigation, we observe covariate profiles $\widetilde{\mathbf{X}}$ of units drawn i.i.d. from the distributions $\{P_g\}$ of the different groups $g \in \mathcal{G}$, and based on those evaluate the prediction model to obtain new HTE predictions $\widehat{\tau}^f(\widetilde{\mathbf{X}}) = f(\widetilde{\mathbf{X}})$ for which we wish to mitigate the bias within and across groups. The mitigation does not require new data on treatment assignments T or outcomes Y for the observations that the mitigation should be applied to; these variables may be realized later and are only needed to detect if any bias remains after mitigation. This setup mimics practice by online platforms and e-commerce companies wherein HTEs are predicted in batches or in sequence as website or app sessions with observable features arrive, prior to assigning them to a treatment (e. g., recommendation, offer, incentive) and before observing their outcomes (e.g., booking, conversion, or engagement).

6.1. Motivation for Our De-Biasing Strategy

So far, we have presented methods for detecting algorithmic bias via estimation and inference of the true prediction error per group. Knowledge of the true error per group is useful for characterizing the algorithmic bias, as it tells us for which groups the HTEs are, on average, over or underestimated and by how much. By Theorem 1, the point estimate $\widehat{B}_g$ converges in the sample limit to a normal random variable centered at the true error b_g. Hence, a tempting but naïve strategy for debiasing when predicting HTEs for yet unseen units is to simply subtract the point estimate from the predicted HTEs per group. By Remark 2, we expect that the de-biased predicted ATE obtained by collapsing the de-biased HTE predictions per group would have no prediction error at the group level, implying no algorithmic bias if applied to each group and the corresponding remainder of observations associated with each group.

However, such a naïve de-biasing strategy is generally not appropriate. The reason is two-fold: First, in practice the true error b_g is estimated from a finite sample of data, and so the point estimate $\widehat{B}_g$ may not exactly equal b_g. If the point estimate is inaccurate or of the wrong sign, the naïve de-biasing strategy may increase algorithmic bias rather than reduce it. Even more worrying is that, since b_g is unobservable, we would not know whether the correction improves or worsen the algorithmic bias. Second, for mitigation, we are typically more concerned about the distribution of bias than its amount for any one particular group. Simply subtracting the point estimate of the estimated prediction error does not penalize greater deviations from the true error and does not account for its sampling variance had the data used for detection been different. The following example mathematically illustrates these problems.

EXAMPLE 1. *We omit the group subscript g to simplify notation. Let $\widehat{\tau}^{f^*}(\widetilde{\mathbf{X}}) \coloneqq \widehat{\tau}^f(\widetilde{\mathbf{X}}) - \widehat{B}$ be a HTE prediction for covariate profile $\widetilde{\mathbf{X}}$ in a given realization of mitigation data that is de-biased with the estimated error of their group. By Def. 1 of collapsibility and the definition in Eq.* (5)*, the finite-sample statistical bias that remain after debiasing the HTE prediction is given by*

$$\widehat{\tau}^{f^*} - \tau = \mathbb{E}[\widehat{W}\widehat{\tau}^f(\widetilde{\mathbf{X}})] - \widehat{B} - \tau \tag{29}$$

$$= \underbrace{\tau^f - \tau}_{=b} - \widehat{B} \tag{30}$$

$$= b - \widehat{B}. \tag{31}$$

By Remark 2 to Theorem 1, we have $\mathbb{E}[b - \widehat{B}] = 0$ since $\widehat{B}$ is a consistent estimate. As for the variance, we have

$$\mathrm{Var}[b - \widehat{B}] = \mathrm{Var}[\widehat{B}] = \underbrace{\mathbb{E}[\widehat{B}^2]}_{(a)} - \underbrace{(\mathbb{E}[\widehat{B}])^2}_{(b)}. \tag{32}$$

Here, $\widehat{B}^2$ is not a consistent estimator for b^2 even if $\widehat{B}$ is consistent for b. Thus, whereas the second expectation (b) vanishes in the sample limit for consistent point estimates $\widehat{B}$, the former expectation (a) does not. Hence, debiasing by simply subtracting the estimated error $\widehat{B}$ does not account for uncertainty in the detection of true prediction errors as reflected in variance in the point estimate.

Overall, the example speaks to the necessity of accounting for the variance and deviation in the estimated error $\widehat{B}_g$ relative to b_g, which may vary by group due to, e. g., different sample sizes, and suggests that we should debias relatively more for the groups with the larger prediction errors, or, only when we are confident that the true prediction error of a group is non-zero.

6.2. De-Biasing using an Uncertainty-Aware Approach

As shown in the previous section, the challenge in practice is that we cannot know the true error b_g, and the naïve de-biasing strategy by subtracting the point estimate $\widehat{B}_g$ does not account for

its variance. Hence, the naïve strategy may not reduce algorithmic bias when predicting HTEs for new observations. To address this, we approach de-biasing as an uncertainty-aware optimization problem. Here, the uncertainty is over the true error b_g, and the optimization problem is to de-bias HTE predictions toward new units so as to minimize the risk that they are subject to algorithmic bias. For this, we adopt statistical decision theory and define risk as the expected loss of prediction errors that may remain after de-biasing over hypothetical realizations of the data. The focus on hypothetical realizations of the data is crucial since, at the the time of mitigation, we do not yet have the data of the units for which we aim to debias the HTE predictions. We then seek to minimize this risk with respect to a correction factor $\gamma_g \in [0, 1]$ that specifies the amount by which we de-bias HTE predictions according to both the expected value and the variance of the point estimate $\widehat{B}_g$. This approach embeds a wide range of mitigation strategies, where the boundary cases $\gamma_g = 0$ corresponds to making no correction and $\gamma_g = 1$ corresponds to the naïve strategy if simply subtracting the full amount of the EPE. Overall, the strength of this approach is that it accounts for that that the estimated error may be inaccurate of the true error and that the sample of data we used to estimate it will be different from the sample of data for which we will make new HTE predictions after de-biasing.

Below, we first introduce the correction factor and then the objective function.

Correction factor: For our mitigation approach, we aim to find a correction factor $\gamma_g \in [0, 1]$ per group $g \in \mathcal{G}$ that adapts the de-biasing according to the expectation and variance in the point estimate $\widehat{B}_g$ so as to minimize the true error for new observations. The HTE prediction for a new observation drawn at random from group g after scaling the debiasing with γ is

$$\widehat{\tau}_g^f(\widetilde{\mathbf{X}}, \gamma) = \widehat{\tau}_g^f(\widetilde{\mathbf{X}}) - \gamma \widehat{B}_g, \tag{33}$$

The *remaining TPE* for group g, denoted b_g^γ, after such a correction is given by

$$\begin{aligned} b_g^\gamma &= \mathbb{E}[W \widehat{\tau}_g^f(\widetilde{\mathbf{X}}, \gamma)] - \tau_g && (34)\\ &= \widehat{\tau}_g^f(\gamma) - \tau_g && (35)\\ &= \widehat{\tau}_g^f - \gamma \widehat{B}_g - \tau_g && (36)\\ &= b_g - \gamma \widehat{B}_g, && (37) \end{aligned}$$

where we collapsed the debiased HTE predictions given and then subtracted the true ATE of a group to arrive at the remaining TPE. The lower boundary case $\gamma = 0$ corresponds to not making a correction, and the upper boundary case $\gamma = 1$ corresponds to simply subtracting the point estimate of the EPE.

The remaining true error $b_g^{\widehat{\gamma}}$ is the target parameter to minimize for mitigation; however, the derivation shows that it depends on b_g and is therefore is unobservable. We thus next present an optimization approach that minimizes the *risk* of remaining TPEs and later consider tests for post-mitigation inference on whether the remaining true error was eliminated.

Risk minimization: Let $L\colon \mathbb{R}\times\mathbb{R}$ be a loss function over the remaining true error by f for group g. To account for deviations in $\widehat{B}_g$ from b_g due to sampling variability, we follow statistical decision theory and define the frequentist risk as the expected loss, i. e.,

$$R(b_g^{\gamma}) = \mathbb{E}_{\mathbf{d}_g\sim P_g}\left[L\left(b_g, \gamma\widehat{B}_g(\mathbf{d}_g)\right)\right] = \int_{\mathcal{D}_g} L\left(b_g, \gamma_g\widehat{B}_g(\mathbf{d})\right)\mathrm{d}P_g(\mathbf{d}_g), \tag{38}$$

where the expectation is over the randomness in $\widehat{B}_g$ given possible realizations of data $\mathbf{d}_g = (\mathbf{x}_i, t_i, y_i)_{i=1}^{n_g}$ used to estimate it drawn i.i.d. from P_g. The decision-maker's objective is to, per group, choose a correction factor that minimizes the risk,

$$\gamma_g^* \in \operatorname*{arg\,min}_{\gamma\in[0,1]} R(b_g^{\gamma}) \quad \text{for all } g\in\mathcal{G}. \tag{39}$$

De-biasing with an optimal correction factor γ_g^* for a given group g minimizes the risk (according to the decision-maker's chosen loss function) of true errors in predicted HTEs towards new observations from the same group. Applying an optimal de-biasing for all groups minimizes the risk that new observations from different groups are subject to algorithmic bias.

6.3. Optimal Correction Factors

For ease of exposition, we focus on optimal correction factors under two standard loss functions: absolute error loss and squared error loss.

Example 2 (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 the mean absolute error (MAE)*

$$R(b_g^{\gamma_g}) = \mathbb{E}_{P_g}\left[\left|b_g - \gamma_g\widehat{B}_g\right|\right]. \tag{40}$$

The MAE places equal weight on all possible estimates of prediction errors. The following proposition shows how we can account for uncertainty in the detection in the use of the MAE risk function.

Proposition 2. *Let* $H_0\colon b_g = 0$ *as in the hypothesis test in Eq.* (24). *The* oracle-*optimal correction factor for Eq.* (38) *under MAE risk is given by*

$$\gamma_g^{\mathrm{MAE}} = \begin{cases} 1, & \textit{if} \quad H_0 \textit{ is false,} \\ 0, & \textit{else.} \end{cases} \tag{41}$$

Proof. The result follows immediately from that MAE risk is minimized in expectation by subtracting a consistent estimate $\widehat{B}_g$ of the true error b_g whenever b_g truly is non-zero. □

In practice, we do not know if the null is true and can at best empirically test it. Thus, a feasible plug-in estimator is

$$\widehat{\gamma}_g^{\mathrm{MAE}}(\alpha) = \begin{cases} 1, & \text{if} \quad \widehat{B}_g / \widehat{\sigma}_g > t_{n_g-2}(\alpha/2) \\ 0, & \text{else.} \end{cases} \tag{42}$$

This is implemented as a t-test with unknown, unequal variances; see Sec. 5.4. The estimated correction factor under MAE risk thus debiases a HTE prediction with the point estimate $\widehat{B}_g$ for a group if the null hypothesis that it is zero is rejected, otherwise no debiasing is applied. It is uncertainty-aware by depending on the decision of the hypothesis test, which incorporates the variance estimate $\widehat{\sigma}_g$ of $\widehat{B}_g$.

Conditional on having received a point estimate and its standard error, the confidence level $\alpha \in [0, 1]$ is the decision-maker's lever for sensitivity to rejecting the null. Specifically, the confidence level 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 whether the true statistical bias is non-zero.

The choice of the confidence level α 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 of the confidence level, 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. However, irrespective of what confidence level is chosen, it should be specified prior to performing the test. Otherwise the theoretical properties of the test are no longer guaranteed by statistical inference theory.

We now consider the risk-minimizing correction factor under squared error loss.

Example 3 (Squared error loss). *The loss is given by* $L(b_g, \gamma_g \widehat{B}_g) = (b_g - \gamma_g \widehat{B}_g)^2$. *The risk function is thereby the mean squared error (MSE)*

$$R(b_g^{\gamma_g}) = \mathbb{E}_{P_g}\left[(b_g - \gamma_g \widehat{B}_g)^2\right]. \tag{43}$$

The MSE risk decomposes into bias squared plus variance. It thereby penalizes larger deviations of the estimated error from the TPE, and more variability in the EPE. The MSE may thus be an appropriate risk function when we care about reducing large disparities relatively more or account for the influence of the sample size used for detection.

PROPOSITION 3. *The optimal correction factor for Eq.* (38) *under the MSE risk is*

$$\gamma_g^{\text{MSE}} = \frac{b_g^2}{\sigma_g^2 + b_g^2}, \tag{44}$$

which can be equivalently expressed as

$$\underbrace{\frac{b_g^2}{\mathbb{E}[\widehat{B}_g^2]}}_{(I)} = \underbrace{\frac{\mathbb{E}[\widehat{B}_g^2] - \sigma_g^2}{\mathbb{E}[\widehat{B}_g^2]}}_{(II)}, \tag{45}$$

where σ_g^2 is the population variance of the estimated error $\widehat{B}_g$.

Proof. See Appendix C.2. □

Proposition 3 presents three different expressions for the optimal correction factor under the MSE risk. The first expression in Eq. (44) shows that the oracle-optimal correction factor under MSE risk depends on population variance σ_g^2 of our estimated error $\widehat{B}_g$. Since variance is non-negative, a greater variance implies less debiasing. This corresponds to that greater uncertainty over the true bias makes us more risk-averse in terms of debiasing. In the theoretical best-case scenario that the variance is zero, meaning that there is no uncertainty about the true error b_g, the optimal correction factor collapses to 1, leading to perfect bias reductions. Hence, a procedure for mitigating algorithmic bias based on MSE risk adapts the debiasing relative to how uncertain we are about the accuracy of our point estimate $\widehat{B}_g$ relative its true value b_g. In contrast, the naïve approach that implicitly fixes the factor at $\gamma = 1$ does not account for this uncertainty.

Expression (I) in Eq. (45) further shows that the oracle correction is simply the ratio between the squared true bias and the expected value of the estimated squared bias. This is intuitive as it scales our correction according to how close our expectation of the magnitude of the estimate is to the truth. However, expression (I) is of limited practice use as it still depends on unobservable true bias b_g. Thus, in expression (II) in Eq. (45), we re-express the optimal correction factor to only depend on the expected value of the squared point estimate, which is observable.

The two alternative expressions in Proposition 3 lead to different plug-in estimators of the optimal correction factor implied by MSE risk. Expression (I) in Eq. (45) suggests replacing b_g^2 with the squared point estimate $\widehat{B}_g^2$ and the mean squared estimate $\mathbb{E}[\widehat{B}_g^2]$ with the empirical analog $\widehat{\mathbb{E}}[\widehat{B}_g^2]$, i.e., the sample average of squared bias estimates, for instance obtained via bootstrap (we later explained our procedure for this). The plug-in estimator is given by

$$\widehat{\gamma}_g^{\text{MSE-I}} = \frac{\widehat{B}_g^2}{\widehat{\mathbb{E}}[\widehat{B}_g^2]}. \tag{46}$$

This estimator may have slight statistical bias in finite samples and in the sample limit due to that $\widehat{B}_g^2$ is not a consistent estimator of b_g^2 even when $\widehat{B}_g$ is a consistent estimator of b_g.

As for expression (II) in Eq. (45), we replace the population variance σ_g^2 of $\widehat{B}_g$ with a sample estimate $\widehat{\sigma}_g^2$. Thus, the second expression in Eq. (44) suggests the alternative plug-in estimator

$$\widehat{\gamma}_g^{\text{MSE-II}} = \frac{\widehat{\mathbb{E}}[\widehat{B}_g^2] - \widehat{\sigma}_g^2}{\widehat{\mathbb{E}}[\widehat{B}_g^2]}. \tag{47}$$

The empirical mean $\mathbb{E}[\widehat{B}_g^2]$ and sampling variance $\widehat{\sigma}_g^2$ may have highly complicated functional form as they both depend on the predicted and estimated ATE for a group. Hence, instead of estimating them analytically from the realized detection data, we propose to approximate both in a data-driven manner using non-parametric bootstrap. We later present our approach to this.

6.4. Post-Mitigation Inference

We now consider the case that the mitigation procedure has been applied and that we are interested in testing whether it eliminated the bias. To do so, we will assume that we have access to experimental data $\widetilde{\mathbf{D}}_g = (\widetilde{\mathbf{X}}_i, \widetilde{T}_i, \widetilde{Y}_i)_{i=1}^{\widetilde{n}_g}$ that are independent of the detection data $\mathbf{D}_g$. The data $\widetilde{\mathbf{D}}_g$ may be a held-out portion of the detection data or collected concurrently with the mitigation by randomly allocating a portion of incoming observations to an A/B test. Post-mitigation inference (or, post-mitigation detection) on remaining bias then proceeds analogously to the hypothesis tests in Section 5.4 and has the same theoretical guarantees even though the inference now clearly depends on the data. In the following, we explain the post-mitigation tests and how the theoretical guarantees are retained.[9]

Consider the remaining true error in a given new sample conditional on an estimated correction factor and prediction error:

$$b_g^{\widehat{\gamma}_g} = b_g - \widehat{\gamma}_g \widehat{B}_g \tag{48}$$

This is unknown as it depends on the true error b_g. However, an estimate can be obtained simply by replacing b_g with a new estimated error $\widetilde{B}_g$ from the new sample $\widetilde{\mathbf{D}}_g$. That is, one uses the estimators in Sec. 5.2 on $\widetilde{\mathbf{D}}_g$ to obtain new estimates $\widetilde{\tau}_g$ and $\widetilde{\tau}_g^f$ of the true and predicted ATE, respectively, and take their difference to get a new estimated error $\widetilde{B}_g$ for the group. Then

$$\widehat{B}_g^{\widehat{\gamma}_g} = \widetilde{B}_g - \widehat{\gamma}_g \widehat{B}_g. \tag{49}$$

estimates how much of the prediction error remains for the observations in group g for which we applied the mitigation. We can use non-parametric bootstrap to approximate the sampling

[9] The following procedure can also be used to estimate confidence interval with correct coverage probability of containing the true remaining statistical and algorithmic bias in the usual manner via normal approximation but, to save space, we omit this from our exposition.

distribution of these post-mitigation estimates and then run a t-test to test the null hypothesis that no statistical bias remains given the debiasing strategy, i.e.,

$$H_0 \colon b_g^{\widehat{\gamma}_g} = 0 \quad \text{vs.} \quad b_g^{\widehat{\gamma}_g} \neq 0. \tag{50}$$

Similarly, we can test the null that no algorithmic bias remains, i. e.,

$$H_0 \colon b_g^{\widehat{\gamma}_g} = b_{-g}^{\widehat{\gamma}_{-g}} \quad \text{vs} \quad b_g^{\widehat{\gamma}_g} \neq b_{-g}^{\widehat{\gamma}_{-g}}. \tag{51}$$

These tests inherit the standard theoretical guarantees of t-tests and follow the same procedure as the test for detection without requiring adaptions. This follows from that the correction factor and prediction errors are estimated on the detection data, whereas the tests are carried out on mitigation data. Because the data sets are independent, standard inference theory applies even though the tested hypotheses depend on data via the estimated correction factor and the detected prediction error. Thus, the well-known problem of naively using the same data twice – first for selecting what to test and then testing it – is avoided, which invalidates the inferential guarantees of standard tests unless accounted for.

Our use of independence between the estimation and inference data is equivalent to the use of sample splitting in methods for post-selection inference (see e.g., Wasserman and Roeder 2009, Rinaldo et al. 2019, Kuchibhotla et al. 2022, Rasines and Young 2023). In that literature, a standard procedure to enable valid statistical inference on, e. g., regression parameters after data-driven variable selection is to use one set of data to perform the variable selection and another set of data to perform inference conditional on the selection. Borrowing from this literature, the theoretical basis our the post-mitigation tests outlined above follows the principles of Rinaldo et al. (2019), who show that sample splitting followed by bootstrap (or Normal approximation) provides near assumption-free and robust inference with asymptotic guarantees that are stronger than those of alternative approaches (see e. g. Kuchibhotla et al. 2022, for an overview) for non-dependent data sampled i.i.d., which is the data generating process considered in this work.

6.5. Price of Fairness

Essentially, our approach entails shifting personalized HTE predictions so that their (weighted) group average is calibrated to the true ATE. Given that the true ATE of any group cannot be known, it follows that applying the optimal correction factors per group leads to the theoretical best achievable minimization of prediction errors in ATEs. As a result, the risk that any one group is subject to larger prediction errors than the other groups is minimized. Thus, de-biasing with the optimal correction factors for all groups leads to HTE predictions that are Pareto-optimal at the group-level in terms of our notion of fairness.

However, the Pareto-optimality at the group-level has a price for personalization. By shifting HTE predictions such that their (weighted) average is calibrated to the true ATE, some degree of personalization in the predicted HTEs may be lost. Hence, in our framework, the concept of "price of fairness" pertains to the trade-off between the prediction performance of HTEs at the personalized level vs. the prediction performance and fairness of HTEs at the group-level. This is in contrast to previous works (e.g., Bertsimas et al. 2011, 2012), which concern fairness-accuracy trade-offs in prediction models within the same level of aggregation (i. e., at the personalized or group level, but not between the two).

7. Evaluation

We now present a general evaluation procedure for detecting and mitigating algorithmic bias using our framework under different mitigation factors. We first present the computational procedure and then present mitigation strategies that can be evaluated.

7.1. Evaluation Procedure

We consider offline evaluation on historical data. We use sample-splitting to divide the historical data per group into a detection set and a mitigation set, for which we apply the detection and the mitigation separately in a hold-out fashion. This way, the evaluation mimics the real-world application of post-processing methods to address algorithmic bias, where the bias is first detected on one set of observations, and then used to mitigate bias in the predictions for new observations. For both the detection and the mitigation step, we use non-parametric bootstrap for each group to estimate the statistics of the EPE (e. g., squared value and variance), which is required for correction factors and inference. As in cross-validation, we swap the roles of the detection and mitigation sets are after a full run and repeat the detection and mitigation procedures, including the bootstraps. The swap reduces the instability of the estimates to the random allocation of observations into detection or mitigation, conditional on the split. We finally repeat the whole procedure with a new random split (and swap) many times. This reduces the instability of the result to the splits themselves. Our procedure is as follows:

1. Randomly split the data per group in to a detection set and a mitigation set.
2. On the detection set:
   (a) Estimate the true ATE and collapse the HTE predictions to the predicted ATE given the HTE measure.
   (b) Take the difference in the the estimates per group to estimate the EPE.
   (c) Take the difference in the estimated error of a group and that of the rest to estimate the algorithmic bias.

(d) Estimate the sampling distribution of all estimates by repeat steps 2–4 many times via non-parametric bootstrap.
(e) Run the hypothesis tests in Sec. 5.4.
(f) Calculate mitigation factors.

3. On the held-out mitigation set:
   (a) Evaluate HTE predictions for all observations.
   (b) Debias the predictions by subtracting the EPEs of the groups multiplied by a correction factor.
   (c) Repeat steps 1 (a)–(e) on the debiased predictions to estimate and perform inference on the remaining statistical bias and algorithmic bias after debiasing.
4. Swap the roles of the detection and mitigation sets and repeat steps 1–3. Save all estimates.
5. Repeat step 1–4 many times with a new random split. Save all estimates.
6. Output summary statistics of the estimates.

Pseudo-code is provided in Appendix E.

The above procedure is designed for offline evaluation on historical data, and therefore provides counterfactual results had different possible mitigation strategies been applied. Moreover, as long as observations are i.i.d. within groups, sampled at random, and the identification assumptions hold, the counterfactual estimates are statistically indistinguishable from those we would obtain from an online evaluation (e. g., A/B test). Nonetheless, adapting the procedure for online evaluation (i. e., randomized experiment) is in principle straightforward. The only required change is that the mitigation data is sampled online (e. g., by directing a percentage of incoming traffic to the treatment condition) and that the de-biasing is applied in batches or on-the-fly to the HTE predictions for the sampled units. Once a pre-specified required sample size has been collected per group, our post-mitigation procedures can be applied just like in the offline case to detect if the online de-biasing was successful; see Step 3 (c).

A feature of an online evaluation is that it complicates the evaluation of different mitigation factors; in practice, one cannot apply different correction factors to the same observations as a means to find the best mitigation strategy. For online evaluation, this can be circumvented by randomly allocating incoming units to different experimental arms, where each arm uses a different correction factor for the de-biasing. Via the randomization to the arms, the performance of the different correction factors would be comparable, and a single mitigation strategy to deploy in the future can at the end of the experiment be chosen as the correction factor that eliminated the most algorithmic bias. It is typical for tech firms, and common procedure at *Booking.com*, to use offline evaluation prior to an online test as a means to set experimental design parameters, such as the minimum detectable effect, required sample size to obtain sufficient power, and which arms (i.e., correction factors) to test.

7.2. Mitigation Strategies

We consider the following correction factors:

1. **No correction:** Setting $\gamma = 0$ yields no bias correction. This strategy is the most conservative in the sense that not making a correction guarantees that we do not increase algorithmic bias.
2. **Mean error:** We set $\gamma = 1$ to yield the naïve strategy of subtracting the estimated error per group irrespective of the result from the hypothesis test and the variance in the point estimate. Thus, this strategy represents a baseline that neglects uncertainty.
3. **Mean error if rejected null:** This is the estimator of the optimal correction factor under the MAE risk provided by $\widehat{\gamma}_g^{\text{MAE}}$ in Eq. 42 where we set $\alpha = 0.05$. Hence, this correcgion factor is identical to the above strategy except that it subtracts the estimated error per group bias only if we reject the null hypothesis that it is zero.
4. **MSE plus approach:** This is the plug-in estimator of expression (I) in Eq. (45) provided by $\widehat{\gamma}_g^{\text{MSE-I}}$ in Eq. (46), where $\widehat{B}_g$ is estimated across the whole sample per group and we obtain the sample variance estimate $\widehat{\sigma}_g^2$ of $\widehat{B}_g$ via non-parametric bootstrap.
5. **MSE minus approach:** This is the plug-in estimator of expression (II) in Eq. (45) given by $\widehat{\gamma}_g^{\text{MSE-II}}$ in Eq. (47), where we estimate $\mathbb{E}[\widehat{B}_g^2]$ with the bootstrap average of the point estimates $\widehat{B}_g^2$ and use the same bootstrap variance estimate $\widehat{\sigma}_g^2$ as in the MSE-I correction approach.

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.

8. Empirical Application at Booking.com

8.1. Setup

We apply our framework to data from a randomized controlled experiment on *Booking.com* estimating the treatment effects of an offer on booking probability. Booking.com uses prediction models trained on data from randomized experiments to learn the heterogeneity in treatment effects across browsing sessions. While estimated HTEs should be personalized based on covariates, they should for fairness reasons not be biased with respect to the origin of a browsing session. However, a strong driver of the heterogeneity in treatment effects on Booking.com is the country of origin of a session; on average, users from some countries respond more to certain offers than users from other countries due to a variety of factors related to heterogeneity in willingness-to-pay, price-sensitivity, preferences, etc. Given such heterogeneity, it is thus important that the prediction models are not systematically worse at predicting the HTEs for some groups. Addressing systematic biases in prediction errors of HTEs across groups ensures that downstream inference and decisions are not systematically biased with respect to country of origin. Hence, the objective is to detect whether a prediction model of HTEs deployed at Booking.com was biased with respect to country of origin and, if so, mitigate it.

8.2. Data

The treatment of interest is the display of an offer of a free benefit that incentivizes a booking of a stay at a hotel. Examples of other benefits that are provided by hotels are free breakfast, late check-out, and room service. Website sessions that were randomly allocated to the treatment arm had the offer displayed next to other benefits on the web pages of the hotels included in the experiment. A hotel web page shows accommodation details such as the price of different rooms, the amenities provided, and benefits, and one navigates to a hotel web page by clicking on the hotel in the ranked list that appears after making a search on Booking.com for a stay, which is comprised of a destination and date range.

The experiment was designed as follows. Website sessions on the desktop version of Booking.com that met the eligibility criteria (explained below) were randomly assigned to be shown the offer or not, with 50% of website session allocated to in treatment arm and 50% to the control arm. The random assignment continued until the treatment arm and the control arm had an equal number of eligible sessions.

The eligibility criteria were as follows. First, the offer was only shown on the webpages of certain hotels selected by Booking.com prior to the experiments. Hence, a user was only eligible if they clicked on one of the included hotels. Second, the stay had to meet a minimum spend threshold pre-determined by Booking.com so as to offset the cost of providing the benefit. Third, the search had to be for a booking of at most six people. Website sessions in the treatment or the control arm that did not meet all eligibility criteria were discarded from the experiment. Thereby all observations are comparable pre-treatment.

The data are as follows. The unit of observation is a website session comprising pre-treatment covariates (i.e., a unique session-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), the random treatment assignment, a binary booking outcome, and a predicted HTE given the values of the covariates for a website session.

The HTEs were predicted using a causal ML model estimated on data from an identical experiment that ran one year earlier. The prediction model was fitted on the data from that experiment to estimate the HTE of the offer on the relative increase in the likelihood of a booking given the covariates. The HTE thereby measured the heterogeneity in the intent-to-treat effect, as users could choose to take up the offer or not, which was the relevant estimand to Booking.com for this treatment. The randomization of sessions into the treatment and control arm ensured overlap for the pre-treatment covariates, such that the HTE could be predicted for the website-sessions that were assigned to the second experiment that ran one year later, and which provided the data we

use. Details on the prediction model, experimental design, and data are provided in Goldenberg et al. (2020).

We omit the data from countries of origin with less than 10,000 observations as our simulation studies indicates that this is the sample size at which our asymptotic results approximately hold in practice. By randomized treatment assignment, each country has an equal number of treated and control observations, thus ensuring that estimates are not skewed towards either treatment or control groups. All identification assumptions listed in Sec. 5.2 hold per the design of the experiment.

8.3. Application of Methods

We apply our evaluation method in Sec. 7 to the data. Per country of origin, we use the ratio-of-means estimator in Eq. (15) to estimate the true ATE and apply our weighted average estimator in Eq. (20) to collapse HTE predictions to predicted ATE. Note that the estimation and collapsing is done both on the detection data and the mitigation data where, for the latter, the collapsing is done using the debiased HTE predictions to estimate the predition errors and algorithmic bias that remain for new observations after debiasing according to a correction factor. We evaluate all correction factors in Sec. 7.2. We run everything 10 times, each time with a new random split of the data into a detection set and mitigation set. For each random split, we swap the roles of the detection and mitigation data and repeat the whole procedure to reduce the instability of the results to the split, as in cross-validation. We draw 50 bootstrap replicates per country for both detection and mitigation, conditional on each split and the swap. This results in $2 \times 50 \times 10$ runs per group that we use estimate the sampling distribution of the estimates across groups. Further details are provided in Appendix F.

8.4. Empirical Results

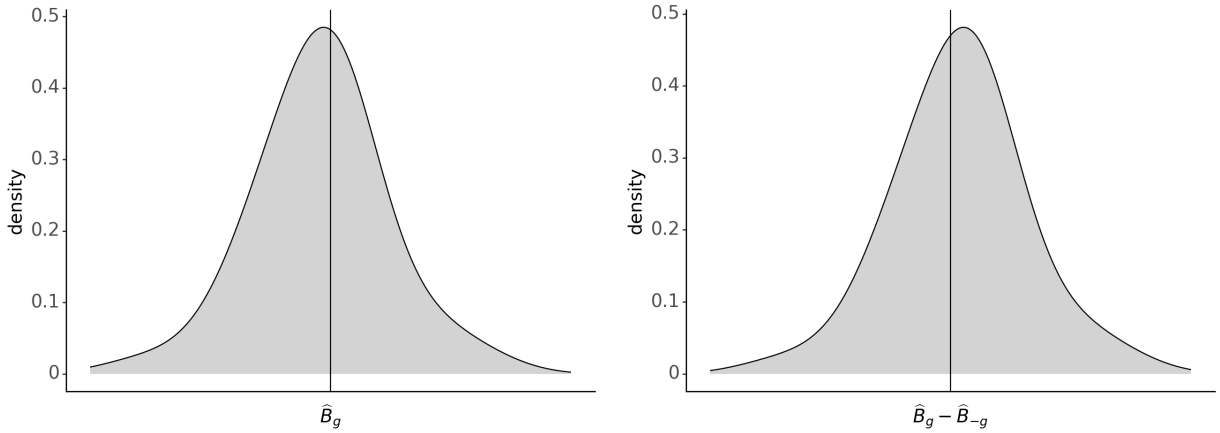
Fig. 1 shows the kernel density estimates of the sampling distribution of the prediction error and the algorithmic bias across countries prior to mitigation. All densities are estimated with a Gaussian kernel and bandwidth set according to Silverman's "rule of thumb" (Silverman 1986, p. 48). We multiply the bandwidths of all densities by 2 to smooth out wiggles and infer general patterns more easily. 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 of 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 observations 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. 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 the tests aligns with the theoretical distribution of the test statistics under the null hypotheses (Fig. 2). Here, the values of the t-statistic 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% confidence level, respectively.



(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.**

Next, we apply our evaluation procedure detailed in Sec. E to analyze how the different mitigation strategies in Sec. 7.2 affect the bias towards new observations from different countries of origin. We emphasize that the evaluation uses a train-test split, such that, per country of origin, the

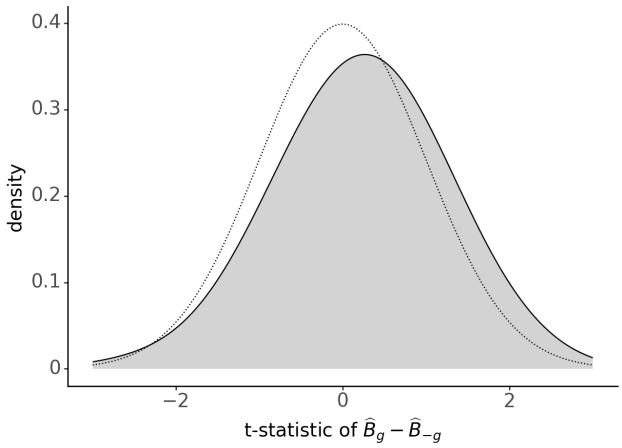


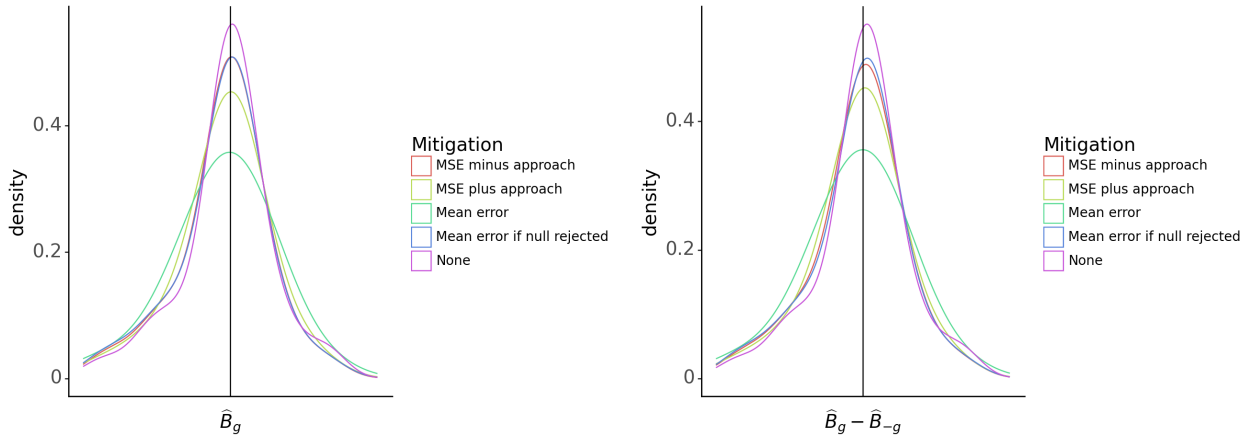
Figure 2 **The t-statistic of the hypothesis test for algorithmic bias across countries of origin with the theoretical density of the test statistic under the null hypothesis overlaid. The t-statistics are computed over 10-fold cross-validation each with 50 bootstrap runs for each country.**

observations 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 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 error without accounting for uncertainty and the MSE error approach that does not account for sampling variability. Finally, because of the success of our mitigation and the very few rejections of the nulls prior to the mitigation, we do not apply the post-mitigation tests as, in this case, the even lower rejection rate post-mitigation would be largely uninformative.

9. Discussion

Our work contributes to previous research along three dimensions. First, recent previous work in marketing and information systems has considered how to debias prediction models of heteroge-



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

Figure 3 **Empirical density of remaining (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.**

neous treatment effects. 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. Moreover, Leng and Dimmery (2024) propose how to calibrate HTEs in randomized experiments. Our work differs from those in that it does not require any additional modeling or data collection, but only inference on (group-wise) sample averages. This makes our framework computationally lightweight and statistically easy to use, thus facilitating its use at scale in ML-driven application areas such as digital platforms.

Second, research in algorithmic fairness has mostly considered algorithmic bias in binary classification in contexts such as hiring, law, and policing. Our framework addresses bias in HTE predictions from an ML model in a digital platform context. 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 fairness in causal inference.

Third, recent research 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, 2023). Here, we find that HTE estimates of an ML model may not recover ATE estimates from randomized experiments equally well for different groups, even though the ML model was trained on randomized experimental data and thereby satisfy the relevant identification assumptions for causal inference. The group-wise disparities in the calibration of predicted treatment effects may arise because of representation bias in the training data, heterogeneous bias from omitted variables or regularization in the ML model, or using global objective functions for

the training algorithm. Irrespective of the source of the bias, our framework can detect the systematic disparities in the prediction error per group and mitigate it. However, in doing so, our results point to a trade-off between personalized HTE predictions vs. group-level calibration and fairness; ML models for predicting HTEs may lose prediction performance at the personalized level when 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 notions of statistical and algorithmic bias in ML models of HTEs and provide methods to address them.

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 and resources for new data collection or modeling. Our framework only requires comparing sample averages and prediction averages for different groups on historical data. The designer typically already has such data for training and evaluating the prediction algorithm. Hence, external stakeholders who have been given the right to an independent audit can request such data and use the proposed framework to detect potential biases and explore the consequences of different mitigation strategies.

9.1. Concluding Remarks

In this paper, we have proposed a framework for detecting and mitigating bias in ML models for HTE prediction. We first presented a notion for algorithmic bias for HTE prediction whereby systematic disparities in prediction errors across groups should be reduced up to a threshold. We then presented a hypothesis test for detecting algorithmic bias and derived a data-driven procedure for de-biasing. We finally demonstrated the effectiveness of our framework in simulation studies and at a large-scale experiment from *Booking.com*.

Funding and Competing Interests

Authors 1, 4, and 5 have no competing interests. Authors 2 and 3 are employed at Booking.com.

References

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.

Angelopoulos AN, Bates S, Fannjiang C, Jordan MI, Zrnic T (2023) Prediction-Powered Inference. *Science* 382(6671):669–674.

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.

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.

Breidt FJ, Opsomer JD (2017) Model-Assisted Survey Estimation with Modern Prediction Techniques. *Statistical Science* 32(2):190–205.

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.

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.

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.

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

Dunn OJ (1961) Multiple Comparisons among Means. *Journal of the American statistical association* 56(293):52–64.

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.

Gordon BR, Moakler R, Zettelmeyer F (2023) Close Enough? A Large-scale Exploration of Non-experimental Approaches to Advertising Measurement. *Marketing Science* 42(4):768–793.

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 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.

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.

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

Kamishima T, Akaho S, Asoh H, Sakuma J (2012) Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, 35–50.

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)*.

Kuchibhotla AK, Kolassa JE, Kuffner TA (2022) Post-Selection Inference. *Annual Review of Statistics and Its Application* 9:505–527.

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.

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

Leng Y, Dimmery D (2024) Calibration of heterogeneous treatment effects in randomized experiments. *Information Systems Research* .

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).

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).

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

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

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

Rasines DG, Young GA (2023) Splitting Strategies for Post-Selection Inference. *Biometrika* 10(3):597—614.

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.

Rinaldo A, Wasserman L, G'Sell M (2019) Bootstrapping and Sample Splitting For High-Dimensional, Assumption-Lean Inference. *The Annals of Statistics* 47(6):3438–3469.

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.

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

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.

Wasserman L, Roeder K (2009) High Dimensional Variable Selection. *Annals of Statistics* 37(5A):2178.

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 (2019) Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 20(1):2737–2778.

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.

Zivich PN, Breskin A (2021) Machine Learning for Causal Inference: On the Use of Cross-fit Estimators. *Epidemiology* 32(3):393.

Online Appendix

Appendix A: Relation to Representative Criteria

Barocas and Selbst (2016) showed that most notions for group-level fairness in the literature belong to three criteria defined as conditional independence relations between predictions, outcomes, and group membership. These criteria are developed for binary classification of observed labels, which has been the canonical setting in the algorithmic fairness literature. In the following, we adapt the representative criteria to HTE prediction under collapsibility and explain their problems in this setting.

We first examine independence (also known as statistical parity (De-Arteaga et al. 2022)). For collapsed HTE predictions, it can be expressed as $\widehat{\tau}_g^f \perp\!\!\!\perp G$, where "$\perp\!\!\!\perp$" denotes statistical independence. The criteria thus requires that, per group, the predicted HTEs collapse to predicted ATEs that are independent across groups. In practice, this means that the prediction model cannot predict systematic heterogeneity at the group level. The problem with this criteria is that it neglects underlying heterogeneity, such as differences in base rates across groups. In many business contexts, treatment effects vary across groups due to underlying differences in behaviors or preferences. This is one motivation for the widespread practice of customer segmentation. What is troubling is that enforcing independence for HTE prediction may lead to a profit reduction without any gain in customer welfare. As an illustration, consider two groups where, in one group 5% of customers respond to a price discount, and in the other group 50% of customer do. A profit-maximizing firm could satisfy independence by providing the offer only to 5 percent of all customers irrespective of group membership. Worse, independence is also fulfilled by not offer the discount to anyone. According to the criterion, the business is acting fairly, even though it lowers customer welfare for one group without increasing it for the other. Hence, independence does not respect that businesses decision making should be Pareto-optimal.

The second representative criterion is sufficiency. For collapsed HTE predictions, it implies $\tau_g \perp\!\!\!\perp G \mid \widehat{\tau}_g^f$. Put simply, conditioning on the collapsed HTE predictions renders the true ATEs independent across groups. One trivial way to fulfil the notion is to inject a sufficient amount of random noise into the HTE predictions. Clearly, this defeats the purpose of HTE prediction for inference or decision-making. Thus, a more constructive solution is to calibrate the prediction model for each group. However, the criterion requires the independence to hold over all groups, and achieving joint calibration can be highly difficult or even impossible in practice without incurring an accuracy loss in the predicted HTEs that essentially makes them uninformative. This is especially challenging when the number of groups is large, the sample sizes vary by group, or the joint distribution vary by group, all of which is common and allowed in our framework. To address the problem with multiple groups, we define our notion as a contrast between a given group and the pooled rest. Our notion thus applies on a group-by-group basis. Moreover, instead of enforcing equal calibration for each group, we allow the disparity in prediction errors to meet an upper bound. Later, we show how this bound is automatically adapted to how well we can estimate the group-level prediction errors per group. Thereby, our notion respects the inferential challenges in detecting algorithmic bias from data, in particular for groups that deviate from the majority (in terms of, e. g., smaller sample size because they are a minority or underrepresented, or because their observations are more noisy due to less precise measurement).

The third representative criterion is called separation, and may for collapsed HTE predictions be stated as $\widehat{\tau}_g^f \perp\!\!\!\perp G \mid \tau_g$. In word, the error is independent across groups. In contrast to the independence criteria, this criterion acknowledges that different groups may have different treatment effects in aggregate. Unlike sufficiency, it does not require the collapsed HTEs to be accurate, but only that the accuracy does not systematically vary by group. Nonetheless, fulfilling this criterion for our setting poses the same challenges as fulfilling sufficiency.

Appendix B: Estimation under Magnitude Effect Measure

For the magnitude HTE measure, it is well known in the causal inference literature that the true ATE is identified by the difference in mean outcomes between the treated and the control group. Hence, to obtain the group ATE we can simply also condition on the group, i. e.,

$$\begin{aligned}\tau_g &= \mathbb{E}[Y(1) \mid G = g] - \mathbb{E}[Y(0) \mid G = g] && (52)\\ &= \mathbb{E}[Y(1) \mid T = 1, G = g] - \mathbb{E}[Y(0) \mid T = 0, G = g] && (53)\\ &= \mathbb{E}[Y \mid T = 1, G = g] - \mathbb{E}[Y \mid T = 0, G = g]. && (54)\end{aligned}$$

This leads to the within-group difference-in-means estimator

$$\widehat{\tau}_g = \frac{1}{\sum_{i \in \mathcal{E}_g} T_i} \sum_{i \in \mathcal{E}_g} Y_i \, T_i - \frac{1}{\sum_{i \in \mathcal{E}_g} (1 - T_i)} \sum_{i \in \mathcal{E}_g} Y_i \, (1 - T_i) \tag{55}$$

where $\mathcal{E}_g$ is the estimation set. As for the predicted ATE τ_g^f, identification follows directly from our collapsibility result and basic probability theory. Specifically, as $W = 1$ for magnitude measures and by linearity of expectations, we have $\mathbb{E}_{P_g}[\widehat{\tau}_g^f(\mathbf{X})] = \tau_g^f(\mathbf{X})$. Hence, an unbiased estimator of the predicted group ATE is given by simply averaging the HTE predictions within the group, i. e.,

$$\widehat{\tau}_g^f = \frac{1}{|\mathcal{P}_g|} \sum_{i \in \mathcal{P}_g} f(\mathbf{X}_i). \tag{56}$$

where $\mathcal{P}_g$ are the observations in the data $\mathbf{D}_g$ that are not in the estimation set $\mathcal{E}_g$.

Appendix C: Proofs

C.1. Proof of Theorem 1

We provide the proof for the relative HTE measure. The proof is analogous for the magnitude HTE measure. We omit the group subscript g to simplify notation.

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 population means of random variables Z_1 and Z_0, respectively. We then have

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

where $\mathbf{\Sigma}$ is the covariance matrix of τ and $\sigma_{0,1}$ is the covariance between Z_1 and Z_0. It follows by the central limit theorem and delta method (Doob 1935, Ver Hoef 2012) that

$$\sqrt{n}\left(\frac{Z_1}{Z_0}-\frac{\mu_1}{\mu_0}\right)\xrightarrow{d}\mathcal{N}\left(0,\sigma^2\right)\qquad\text{for } n\to\infty. \tag{57}$$

Now, let $Z_1=\widehat{\mathbb{P}}_n[Y=1\,|\,T=1]$ and $Z_0=\widehat{\mathbb{P}}_n[Y=1\,|\,T=0]$. Then $\widehat{\tau}=Z_1/Z_0$ and we have $\mu_1:=\mathbb{P}[Y=1\,|\,T=1]$, $\mu_0:=\mathbb{P}[Y=1\,|\,T=0]$, and thus $\tau=\mu_1/\mu_0$. It follows directly that

$$\sqrt{n}\left(\widehat{\tau}-\tau\right)\xrightarrow{d}\mathcal{N}\left(0,\sigma^2\right)\qquad\text{for } n\to\infty. \tag{58}$$

By the unbiasedness of $\widehat{\tau}^f$ and Eq. (5), we have

$$\sqrt{n}\left(\widehat{\tau}^f-\tau\right)\xrightarrow{p}b\qquad\text{for } n\to\infty. \tag{59}$$

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

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

This concludes the proof. □

C.2. Proof of Proposition 3

We first derive how the squared prediction error after de-biasing depends on the correction factor. For ease of notation, we omit the group index g. Recall that the prediction error after de-biasing is $b_\gamma = (b - \gamma\widehat{B})$. Then its squared value is

$$b_\gamma^2 = (b - \gamma\widehat{B})^2. \tag{60}$$

For any consistent estimator, $\widehat{B}$ is a normal random variable. Using the definition of expectated value of squared random variables, we get

$$\mathbb{E}[b_\gamma^2] = \gamma^2\sigma^2 + b^2(\gamma - 1)^2. \tag{61}$$

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

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

Setting the derivative to 0 and solving for γ yields

$$2\gamma\sigma^2 + 2b^2(\gamma - 1) = 0 \tag{62}$$
$$\iff \gamma\sigma^2 = b^2 - \gamma b^2 \tag{63}$$
$$\iff \gamma(\sigma^2 + b^2) = b^2 \tag{64}$$

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

$$\gamma^* = \frac{b^2}{\sigma^2 + b^2} \tag{65}$$

This is the expression in Eq. (44) of Proposition 3. Using that $\mathbb{E}[\widehat{B}^2] = b^2 + \sigma^2$ and swapping terms in the denominator of Eq. (65) yields expression (I) in Proposition 3, i. e.,

$$\gamma^* = \frac{b^2}{\sigma^2 + b^2} = \frac{b^2}{\mathbb{E}[\widehat{B}^2]} \tag{66}$$

Similarly, using that $b^2 = \mathbb{E}[\widehat{B}^2] - \sigma^2$ and swapping terms in the numerator of Eq. (66) gives expression (II) in Proposition 3. We thus yield

$$\gamma^* = \frac{b^2}{\sigma^2 + b^2} = \frac{b^2}{\mathbb{E}[\widehat{B}^2]} = \frac{\mathbb{E}[\widehat{B}^2] - \sigma^2}{\mathbb{E}[\widehat{B}^2]}. \tag{67}$$

This concludes the proof. □

Appendix D: Derivation of Estimator

We now show the derivation of the estimator in Sec. 5.2.2. We omit the group subscript g as the results apply to each group analogously. We first have that

$$\widehat{\tau}_g = \frac{\sum_{i\in g}(1-T_i)Y_i f(\mathbf{X}_i)\widehat{\lambda}_g + \sum_{i\in g} T_i Y_i \widehat{\psi}_g}{\sum_{i\in g}(1-T_i)Y_i\widehat{\lambda}_g + \sum_{i\in g} T_i Y_i \frac{1}{f(\mathbf{X}_i)}\widehat{\psi}_g} \tag{68}$$

$$= \frac{\sum_{i\in g}(1-T_i)Y_i f(\mathbf{X}_i) \times \frac{1}{\sum_{i\in g}(1-T_i)Y_i}\sum_{i\in g}(1-T_i)Y_i f(\mathbf{X}_i) + \sum_{i\in g} T_i Y_i \frac{1}{\sum_{i\in g} T_i Y_i}\sum_{i\in g} T_i Y_i f(\mathbf{X}_i)}{\sum_{i\in g}(1-T_i)Y_i \frac{1}{\sum_{i\in g}(1-T_i)Y_i}\sum_{i\in g}(1-T_i)Y_i f(\mathbf{X}_i) + \sum_{i\in g} T_i Y_i \frac{1}{f(\mathbf{X}_i)} \times \frac{1}{\sum_{i\in g} T_i Y_i}\sum_{i\in g} T_i Y_i f(\mathbf{X}_i)} \tag{69}$$

$$= \frac{\frac{1}{\sum_{i\in g}(1-T_i)Y_i}\left(\sum_{i\in g}(1-T_i)Y_i f(\mathbf{X}_i)\right)^2 + \sum_{i\in g} T_i Y_i f(\mathbf{X}_i)}{\sum_{i\in g}(1-T_i)Y_i f(\mathbf{X}_i) + \sum_{i\in g} T_i Y_i} \tag{70}$$

$$= \frac{\sum_{i\in g}(1-T_i)Y_i\widehat{\lambda}_g^2 + \sum_{i\in g} T_i Y_i \widehat{\psi}_g}{\sum_{i\in g}(1-T_i)Y_i\widehat{\lambda}_g + \sum_{i\in g} T_i Y_i}. \tag{71}$$

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 sessions belonging to a group g (i.e., the HTE of a user-session does not depend on her covariates $\mathbf{X}$), then the ATE for the treated (i.e., Eq. (18)), the ATE for the untreated (i.e., Eq. (19)), and the ATE across both the treated and non-treated (i.e., Eq. (20)) 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

$$\widehat{\psi}_g = \frac{\sum_{i\in g} T_i Y_i}{\sum_{i\in g} T_i Y_i \times \frac{1}{\tau_g}} = \frac{1}{\sum_{i\in g} T_i Y_i}\sum_{i\in g} T_i Y_i \tau_g = \tau_g. \tag{72}$$

For the ATE for the controls, we also get that

$$\widehat{\lambda}_g = \frac{1}{\sum_{i\in g}(1-T_i)Y_i}\sum_{i\in g}(1-T_i)Y_i\tau_g = \tau_g. \tag{73}$$

Finally, plugging these into the estimator for the ATE in Eq. (71) yields

$$\frac{\sum_{i\in g}(1-T_i)Y_i\tau_g^2 + \sum_{i\in g} T_i Y_i \tau_g}{\sum_{i\in g}(1-T_i)Y_i\widehat{\tau}_g + \sum_{i\in g} T_i Y_i} = \tau_g \times \frac{\sum_{i\in g}(1-T_i)Y_i\tau_g + \sum_{i\in g} T_i Y_i}{\sum_{i\in g}(1-T_i)Y_i\tau_g + \sum_{i\in g} T_i Y_i} = \tau_g, \tag{74}$$

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

Appendix E: Pseudo-Code of Evaluation Procedure

Our evaluation follows best practices for evaluating ML models for causal inference (Zivich and Breskin 2021) and uses a general procedure based on multiple runs of random sample-splitting and bootstrapping. Here, the multiple runs of random sample-splitting reduces the instability of the estimates to a split, the sample splitting allows us to evaluate the mitigation for new observations not used for detection, and bootstrapping allows us to estimate the variance in the estimated error (which are required for the detection and mitigation of algorithmic bias). We also use random splitting of the data into an estimation set and a prediction set such that the covariances between the predicted and the estimated ATEs can be ignored for estimating the variance of the EPE. For additional robustness, cross-validation can be applied simply by for each split into a detection and a mitigation set, swapping the roles of the detection and mitigation sets and taking the average over the swaps. Likewise, an outer cross-validation is applied by swapping the roles of the detection and mitigation sets and averaging the post-mitigation estimates across the now two runs per random split. Although this nested cross-validation can reduce the instability of the procedure, it incurs a computational cost of quadrupling the number of executions.

Algorithm 1 provides pseudo-code. For simplicity, we omit the multiple runs from the pseudo-code. It can be incorporated simply by calling the algorithm many times and computing a summary statistic (e.g., mean or median) of the outputs over the runs.

Algorithm 1: Offline evaluation for detection and mitigation

Input:
- Number of bootstrap runs Z;
- Confidence level α;
- Collection of mitigation strategies $\mathcal{C}$;
- Data $\mathbf{H}_g = (\mathbf{X}_i, T_i, Y_i)_{i=1}^{n_g}$ per group $g = 1, \ldots, |\mathcal{G}|$

Output: Estimates of remaining statistical and algorithmic bias per group

for $g = 1, \ldots, |\mathcal{G}|$ **do**
- $\mathbf{H}_{-g} \leftarrow \bigcup_{j \neq g} \mathbf{H}_j$;
- Randomly split $\mathbf{H}_g$ into detection data $\mathbf{D}_g$ and mitigation data $\widetilde{\mathbf{D}}_g$;
- Repeat for $\mathbf{H}_{-g}$ to obtain $\mathbf{D}_{-g}$ and $\widetilde{\mathbf{D}}_{-g}$;
- Randomly split $\mathbf{D}_g$ and $\mathbf{D}_{-g}$ into a prediction sets $\mathcal{P}_g$, $\mathcal{P}_{-g}$ and estimation sets $\mathcal{E}_g$, $\mathcal{E}_{-g}$;
- Collapse the HTE predictions in $\mathcal{P}_g$ and $\mathcal{P}_{-g}$ to $\widehat{\tau}_g^f$ and $\widehat{\tau}_{-g}^f$ following Sec. 5.2.2;
- Calculate $\widehat{\tau}_g$ and $\widehat{\tau}_{-g}$ on $\mathcal{E}_g$ and $\mathcal{E}_{-g}$. resp., using an estimator in Sec. 5.2.1;
- $\widehat{B}_g \leftarrow \widehat{\tau}_g^f - \widehat{\tau}_g$;
- $\widehat{B}_{-g} \leftarrow \widehat{\tau}_{-g}^f - \widehat{\tau}_{-g}$;
- $\widehat{A}_g \leftarrow \widehat{B}_g - \widehat{B}_{-g}$;
- Bootstrap $\mathbf{D}_g$ and repeat steps 5–9 Z times to estimate the sampling distributions;
- Test the null hypotheses in Eq. (24) and (26);
- **for** each correction factor $c \in \mathcal{C}$ **do**
  - Calculate $\widehat{\gamma}_g^c$ and $\widehat{\gamma}_{-g}^c$;
- Repeat steps 5–7 on $\widetilde{\mathbf{D}}_g$ and $\widetilde{\mathbf{D}}_{-g}$ to update the ATE estimates;
- **for** $c \in \mathcal{C}$ **do**
  - $\widehat{B}_g^c \leftarrow \widehat{\tau}_g^f - \widehat{\gamma}_g^c \widehat{B}_g - \widehat{\tau}_g$;
  - $\widehat{B}_{-g}^c \leftarrow \widehat{\tau}_{-g}^f - \widehat{\gamma}_{-g}^c \widehat{B}_{-g} - \widehat{\tau}_{-g}$;
  - $\widehat{A}_g^c \leftarrow \widehat{B}_g^c - \widehat{B}_{-g}^c$;
  - Bootstrap $\widetilde{\mathbf{D}}_g$ and repeat steps 17–19 Z times to estimate the sampling distributions;
  - Test the null hypotheses in Eq. (24) and (26);

return $\widehat{B}_g^c$ and $\widehat{A}_g^c$ for all $g \in \mathcal{G}$ and $c \in \mathcal{C}$;

Appendix F: Evaluation Procedure for Empirical Application

We first randomly split the data in half into detection data and mitigation data. On the detection data, we use the ratio-of-means estimator in Eq. (15) to estimate the true ATE per country of origin and apply our weighted average estimator in Eq. (20) to collapse the prediction model's relative HTE predictions to a predicted relative ATE per country of origin. For each group, we also do this on all observations not belonging to the group, so that we can construct the algorithmic bias estimate following Eq. (11). We repeat the whole procedure conditional on the detection-mitigation split 50 times via non-parametric bootstrap, and use the bootstrap estimates to construct the correction factors for the mitigation strategies.

On the held out mitigation data, we then evaluate all mitigation strategies from Sec. 7.2. We use the correction factors and estimated prediction errors calculated from the detection split to debias the HTE predictions in the mitigation data, as shown in Eq. (33). We repeat this for all correction factors separately and save the corresponding debiased HTE predictions. For each debiasing strategy, we then apply our weighted average estimator in Eq. (20) to collapse the debiased relative HTE predictions to a debiased predicted relative ATE per country of origin. We estimate the true ATEs by applying the ratio-of-means estimator per group on mitigation data. For each debiased predicted ATE corresponding to a correction factor, we take the group-wise differences between the debiased predicted ATE and the estimated ATE per group. This yields estimates of the prediction errors that remain after debiasing; see Eq. (49). For each group, we repeat all steps on the pooled observations in the mitigation data not belonging to that group, so that we can construct an estimate of the algorithmic bias that remain after debiasing according to each correction factor. We repeat all of steps on the mitigation data 50 times via non-parametric bootstrap, and save the bootstrap estimates of the prediction errors and algorithmic bias that remain per group. We then swap the roles of the detection and mitigation set and repeat the detection and mitigation procedures.