Fairness in Online Ad Delivery

Joachim Baumann
baumann@ifi.uzh.ch
University of Zurich
Zurich University of Applied Sciences
Zurich, Switzerland

Piotr Sapiezynski
p.sapiezynski@northeastern.edu
Northeastern University
Boston, USA

Christoph Heitz
christoph.heitz@zhaw.ch
Zurich University of Applied Sciences
Zurich, Switzerland

Anikó Hannák
hannak@ifi.uzh.ch
University of Zurich
Zurich, Switzerland

ABSTRACT

Advertising funds a number of services that play a major role in our everyday online experiences, from social networking, to maps, search, and news. As the power and reach of advertising platforms grow, so do the concerns about the potential for discrimination associated with targeted advertising. However, despite our ever-improving ability to measure and describe instances of unfair distribution of *high-stakes* ads—such as employment, housing, or credit—we lack the tools to model and predict the extent to which alternative systems could address such problems. In this paper, we simulate an ad distribution system to model the effects that enforcing popularly proposed fairness approaches would have on the utility of the advertising platforms and their users. We show that in many realistic scenarios, achieving statistical parity would come at a much higher utility cost to platforms than enforcing predictive parity or equality of opportunity. Additionally, we identify a tradeoff between different notions of fairness, i.e., enforcing one criterion leads to worse outcomes with respect to other criteria. We further describe how pursuing fairness in situations where one group of users is more expensive to advertise to is likely to result in "leveling down" effects, i.e., not benefiting any group of users. We show that these negative effects can be prevented by ensuring that it is the platforms that carry the cost of fairness rather than passing it on to their users or advertisers. Overall, our findings contribute to ongoing discussions on fair ad delivery. We show that fairness is not satisfied by default, that limiting targeting options is not sufficient to address potential discrimination and bias in online ad delivery, and that choices made by regulators and platforms may backfire if potential side-effects are not properly considered.

CCS CONCEPTS

• **Information systems** → **Online advertising**; • **Computing methodologies** → **Machine learning**; *Modeling and simulation*; • **Social and professional topics** → *Computing / technology policy*.

This work is licensed under a Creative Commons Attribution International 4.0 License.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658980

KEYWORDS

algorithmic fairness, online advertising, bias mitigation, leveling down

ACM Reference Format:
Joachim Baumann, Piotr Sapiezynski, Christoph Heitz, and Anikó Hannák. 2024. Fairness in Online Ad Delivery. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3630106.3658980

1 INTRODUCTION

Advertising is the backbone of a variety of online services, particularly those available to users for free. Websites and apps fund their existence by offering screen space to parties that leverage vast amounts of data to present the most profitable ads for each given user. Search engines such as Google and Bing, as well as social network platforms such as Meta and X (formerly known as Facebook, Inc. and Twitter, respectively), have grown to control much of this process by owning the screen real estate, data collection, and ad matching processes. This accumulation of personal data has obvious implications for privacy as well as trust and safety. The functionality built atop this rich data, especially with respect to detailed ad targeting, can be used by malicious actors. For example, advertisers may choose to target their ads in discriminatory ways, for example, by prohibiting older users from seeing job ads [4], or non-white users from seeing housing opportunities [5].

Limiting the targeting options does not fully mitigate the potential for discriminatory effects in advertising. Harmful effects may stem from the process of optimizing ad *delivery*, even for non-malicious advertisers: ads appearing in Google search are more likely to suggest a criminal history when users search for individuals with Black-sounding names, even if those individuals have no such history [32, 43]; women may see fewer opportunity ads, especially if the advertisers operate on smaller budgets, because of the so-called *competitive spillover* effects [31]; similarly, the higher costs of running informative online ads for SNAP (short for the Supplemental Nutrition Assistance Program, which provides food benefits to low-income families) targeting Spanish speakers, compared to English speakers, mean that allocating funds to reach Spanish-speaking audiences reduces the overall number of individuals who can be informed and enrolled in the program [30]. Even with appropriate budgets, Meta may deliver different job and housing opportunities based on the gender and race of the recipient, in

pursuit of maximizing relevance [3]. All these phenomena translate both to individual and societal harms. Those who are not presented with opportunities within a particular field are less likely to seek employment there. And even if they do apply, they are less likely to be hired, if the employers see an indication of criminal history while searching for their names.

The problem of unfair ad delivery has been recognized in both the United States and the European Union. Meta has been forced to address biases in housing ad delivery as part of a settlement with the U.S. Department of Justice [47]. With the advent of the Digital Services Act (DSA) in the EU, online platforms will need to increase the transparency of online advertising and will be subject to third-party audits for algorithmic biases. Additionally, the Digital Markets Act (DMA) is already in effect, with initial investigations into non-compliance by designated gatekeepers underway [20], and the recently enacted AI Act classifies "AI systems [...] to place targeted job advertisements" as *high-risk* [21, p. 425]). Furthermore, studies such as that by Koenecke et al. [30] suggest that there is broad public support for prioritizing fairness over efficiency in ad delivery, indicating a societal shift towards valuing equity in online advertising practices.

Despite the growing awareness of the discriminatory effects of online advertising and a building consensus on the need to address it, there is much less agreement on what an *unbiased* system would look like, and how exactly it should be implemented. The settlement between Meta and the U.S. Department of Justice establishes a fair system as one that achieves *statistical parity*, such that the distribution of gender, age, and race among the users who were shown an ad (i.e., the *actual audience*), closely resembles the demographic distribution of the users in the *targeted audience*, who were active on the platform during the lifetime of that ad (i.e., the *eligible audience*). However, while Meta has claimed it now adheres to statistical parity, they do not provide enough information to the external auditors to actually verify this [2, 24]. Furthermore, this approach is only one of possible views of "fairness", and others have been proposed [8, 35, 49]. Importantly, if there are population-level differences in interests, qualifications, or propensity to click on advertising, enforcing multiple fairness definitions at the same time may be mathematically impossible [29].

The lack of transparency surrounding online advertising platforms makes selecting appropriate measures even more difficult. Researchers have tried investigating user experiences through automated sock puppet accounts [17] as well as studying spillover effects [31] and ad delivery optimization effects [3, 41] by running ads themselves and leveraging the platforms' reporting tools. Unfortunately, such approaches do not allow us to understand the implications of alternative solutions. To this end, several previous papers focused on modeling the core underlying mechanism of online advertising: the real-time bidding auctions [11, 36]. The fair machine learning literature's proposed fairness criteria and their implications are yet to be fully understood within the complex domain of online programmatic advertising.

1.1 Our contributions

Conducting real-world studies to evaluate the impact of fairness-enhancing interventions on online ad delivery system presents significant challenges, as outside researchers cannot directly experiment with the platforms' advertising algorithms. While experimental evaluations of unfairness in ad delivery systems are suited to detect platform-specific problems, they are limited in scope. In this work, we use simulations to observe general trends regarding the utility for advertising platforms and their users under various popularly proposed fairness approaches. Simulations are the only method to measure counterfactual outcomes—such as whether someone would have clicked on an ad had they seen it—and to control and manipulate various factors, including platform utility, user benefits associated with an ad impression, and group-level differences in ad clicking probabilities. This approach enables us to abstract away the effects of the ad content and advertiser account properties (which is impossible in experiments that involve running particular ads on specific accounts) and to test a broad range of reasonable parameter values and combinations. Our findings demonstrate that discriminatory effects persist and are not merely artifacts of how platforms manage specific ads or accounts.

In this paper, we conduct a simulation study to explore the impact of enforcing different fairness notions in the context of high-stakes online ad delivery. Our results are predicated on minimal assumptions that accurately reflect real-world advertising systems and include a sensitivity analysis. Consequently, the following findings are expected to hold generally for different online advertising algorithms, even when abstracting away the auction process:

- **The delivery of opportunity ads on online platforms is inherently unfair.** This is attributed to the platforms' efficiency-driven operations aimed at maximizing their utility.
- **Limiting the targeting options for advertisers does not ensure equitable online advertising.** Using personalized ML-based predictions, the resulting ad impressions can still be unfair for certain demographic groups.
- **Enforcing fairness often reduces platform utility.** Specifically, we demonstrate that achieving predictive parity or equality of opportunity generally incurs a lower utility cost for platforms than achieving statistical parity.
- **Tradeoffs exist between different notions of fairness.** Enforcing a specific fairness criterion can inadvertently worsen outcomes concerning other criteria.
- **Enforcing fairness can result in potentially undesirable side-effects.** For example, we find that "leveling down" scenarios occur in the presence of large spillover effects.
- **Negative effects can be prevented.** We show that it is crucial to ensure that the platforms carry the cost of fairness interventions instead of passing it on to end-users and advertisers, in order to limit the undesirable side-effects.

Our research offers valuable implications for the regulation, monitoring, and compliance of ad delivery system (summarized in Table 2). Additionally, our approach facilitates thoughtful consideration and evaluation of alternative solutions when imposing new fairness constraints, thereby helping ensure fair ad delivery and overall equity on online platforms.

Code. To ensure the reproducibility of the experiments and results, we have made our code publicly available at https://github.com/joebaumann/fair-online-ad-delivery.

2 BACKGROUND AND RELATED WORK

This paper draws on research in online ad auction design and algorithmic fairness. Below we introduce the important aspects of both of these streams of inquiry.

2.1 Ad auctions and auto-bidding

There is a complex ecosystem around matching available ads to the individuals among the targeted audience who ultimately see them. This matching process is backed by expansive data collection that allows both for fine-grained segmentation of audiences and for immediate feedback on each ad's performance. Whenever a user sees an ad on a website, they witness a result of an elaborate process involving multiple parties. First, the *publisher*, i.e., the entity that controls the ad space, announces to a supply-side platform (SSP) that an ad slot is available to be shown to a user identified with a cookie. The SSP may combine the cookie with all their other information about that user (behaviors, demographics, etc.) and share it with an *ad exchange*. The ad exchange hosts an ad auction where demand side platforms (DSPs), acting on behalf of advertisers, choose whether, and how much to bid for the opportunity to show their ad. The DSP who places the winning bid gets to present their ad to the user. The advertiser is charged per thousand impressions (cost per mille, or CPM), or only whenever the user actually clicks on the ad (cost per click, or CPC). In the latter scenario, the DSP is particularly motivated to bid in auctions where they estimate the user will find the ad *relevant*. The entire process happens programmatically, in real-time, and appears seamless to the user.

More recently, major online platforms such as Meta or Google have been taking over the roles of both SSPs and DPSs. Rather than relying on showing ads on third-party webpages through SSPs, these platforms can show ads within their own products. They also hold the auctions themselves, eliminating the need for third-party ad exchanges. Finally, they replaced much of what DSPs would offer with *auto-bidding*. In the auto-bidding scenario, the advertiser specifies the characteristics of their target audience and sets the budget. The advertiser also selects their goal, e.g., maximizing impressions, clicks, or conversions. The platform then bids on the advertiser's behalf in an attempt to maximize their stated goal. Of course, maximizing that goal is subject to the constraint of simultaneously optimizing the platform's revenue; if the platform overwhelms the user with ads, or shows ads that are upsetting, or even uninteresting, the user might stop browsing, thus reducing the number of future ad slots. Therefore, the platforms aim to show users relevant ads. This motivation is enacted through an important change to the auction system. Namely, the platform *subsidizes* bids from ads predicted to be relevant to the user [33], i.e., the advertiser actually pays less to show their ads to users who the platforms deem well-matched. Conversely, ads predicted to have less relevance to the particular user are less likely to win an auction even if they have budgets comparable to those of relevant ads.

2.2 Fairness and discrimination in online ads

At first glance, the developments in auto-bidding appear to benefit all stakeholders: they lower the knowledge barrier to entry for advertisers, who can now rely on the platform's algorithms to optimize the matching and bidding; at the same time, delivery optimization promises a better user experience through ads that are more likely to be of interest. Unfortunately, certain problems become apparent when looking at the advertising platforms through the lens of fairness. First, the advertiser may choose targeting criteria that are discriminatory, e.g., by excluding individuals of particular gender or age from seeing the ads. Additionally, there are two important phenomena that can lead to apparently discriminatory or harmful outcomes without the advertiser's intent: *competitive spillovers*, and *optimization for relevance*.

Discriminatory targeting. Until recently, advertisers on Meta could choose to use gender, age, and "Ethnic affinity" for targeting and excluding audiences from seeing opportunity advertising [5]. Following 2019 settlement between Meta and a number of civil rights organizations [46] these criteria have been removed. Despite that, a malicious advertiser can still discriminate by targeting interests whose distribution is skewed between genders or racial groups [42] or by targeting a biased Custom Audience [41]. Importantly, in this work, we focus on situations where *the advertiser selected no discriminatory targeting options*.

Competitive spillovers are rooted in the auction mechanism. Imagine there are two advertisers, one targeting men and women, the other targeting only women. Because of that targeting, they will be competing in the ad auctions for women, thus raising the price necessary to win the auction. As a consequence, if the advertiser who targets men and women uses the same bids for both, they are more likely to lose auctions for impressions to women, and end up showing their ad predominantly to men, despite inclusive targeting [31]. One possible remedy to this problem is for the advertiser to dynamically adjust the bid based on the gender of the user whose ad slot is auctioned [36].

Optimization for relevance aims to maximize the utility both for the advertisers and the users. The allocation predicted to be optimal may, however, mean that the subsets of the audience that are eventually shown the ad are skewed along the lines of gender, race, age, or political inclination [30]. Ali et al. [3] have shown that, despite inclusive targeting, the job ads they ran were shown to skewed audiences depending on the advertised job: Meta showed opportunities in the lumber industry predominantly to white men; openings for taxi drivers and janitors went mostly to Black users; offers for supermarket cashiers were presented disproportionately to women. The effect is particularly troubling because it appears to replicate stereotypes even if the qualifications (and lack thereof) to perform the advertised job are evenly distributed among the targeted audience [28]. Even though the source of algorithmic bias is unrelated to the competitive spillovers, the proposed solutions also rely on modifying the bidding strategy in a similar fashion [11, 12, 38]. However, in the auto-bidding scenario, it is the platform's responsibility rather than each individual advertiser's.

2.3 Side-effects of enforcing fairness constraints

Many different notions of fairness exist in fair machine learning literature, none of which is universally acceptable [35, 49]. Skewed ad delivery corresponds to a violation of the fairness criterion called statistical parity. Notice that statistical parity can be violated even in the case of a perfectly accurate classifier (i.e., one that perfectly predicts which users are interested in an ad), e.g., if the base rates

(BR, the proportion of individuals interested in an ad, also called *prevalence*) differ largely across groups. The issue of BR disparities is frequently discussed as a reason for unfair outcomes in utility-optimizing decision systems [13, 14, 29, 40].

Alternative notions of fairness, which are based on the actual outcome (i.e., measuring clicks rather than impressions), have been proposed [8, 49]. For example, predictive parity, also referred to as PPV (positive predictive value) parity [10], is closely related to the notion of calibration for the case of continuous-valued predicted scores [6, 18]. In the context of serving online ads with binary outcomes and decisions, a system satisfying predictive parity would indicate that the click-through rates (CTR) are balanced between groups. Proponents of this solution may claim it reflects the underlying interests in the advertised content. Finally, equality of opportunity (also referred to as TPR—true positive rate—parity) requires an equal share of user groups to see the ads, among all those who would click on the ad [25].

The choice of a particular fairness metric is crucial since it is mathematically impossible to satisfy all parity metrics simultaneously (apart from degenerate cases) [13, 29]. Furthermore, Friedler et al. [22] show that there is a trade-off between accuracy and fairness,[1] although this trade-off may be negligible in practice when equality of opportunity is the metric of choice [39]. Finally, enforcing group fairness for ad impressions on online platforms may lead to worse outcomes for all groups [26]. This "leveling down objection" challenges the idea of egalitarianism [16, 37], which is the basis for most group fairness criteria discussed in the fair ML literature, as they typically aim to minimize inequality [34]. This raises the question of whether we should demand equality even if it does not benefit anyone, emphasizing that the group-specific utility of fairness-enhancing techniques must be evaluated.

2.4 Legal responsibility

The U.S. Department of Justice found that demographic skews in ad delivery constitute a violation of the Fair Housing Act and sued Meta in the landmark case alleging algorithmic discrimination [47]. As part of the settlement, Meta pledged to address this problem using the Variance Reduction System (VRS) [44]. VRS is to be applied to ads in protected categories (housing, employment, and credit) and ensure *statistical parity* between the targeted and actual audience. Under this system, the demographic distribution (initially by gender and race, later also by age) of the actual audience of each ad in the protected category will be continuously monitored, and the bidding strategy adapted. For example, if the fraction of men in the actual audience is higher than in the eligible audience, the bidding strategy will be modified in an attempt to slow down the delivery to men and/or speed up the delivery to women. The system is designed to not explicitly rely on demographic attributes when making the bidding choices, but—if the system achieves its goal—the end effect is not different than if these variables were used directly.

3 A SIMPLE ONLINE AD DELIVERY MODEL

In this section, we introduce a model to simulate the delivery of ads on online platforms. We describe the overall setup and introduce the used notation before presenting optimal ad delivery strategies with and without fairness constraints.

3.1 Setup and notation

The auction mechanism used by online ad platforms is a complex process involving many factors, such as bid price, ad quality, and relevance. However, online platforms place bids on behalf of advertisers through auto-bidding in auctions whose mechanism they design and control, and can thus decide which users see which ads. Thus, we may abstract away the underlying auction mechanism and model the platform's ad delivery problem as a simple binary decision problem from the perspective of a given high-stakes ad: for all ad slots where the high-stakes ad is a competitor, the platform must decide whether or not to show it. This allows us to investigate the effect of introducing different fairness constraints to the ad delivery system on different demographic groups.

The decision variable is denoted by D, where $D = 1$ indicates showing the high-stakes ad and $D = 0$ indicates showing another ad. After being shown the ad, a user either clicks on it ($Y = 1$) or not ($Y = 0$).[2] The revenue for the given high-stakes ad is denoted by α, which is the utility gained if the user clicks on the ad, i.e., the advertiser's bid amount. The utility gained if the high-stakes ad is not shown is denoted by β, which could represent the expected utility of the best other ad or the utility gained by not showing any ad to the user. We assume that advertisers are willing to pay for their ads to be shown, i.e., $\alpha > 0$ and $\beta > 0$. This assumption holds as long as there are other ads to show in a given slot, or if there is non-paid content to be shown instead that would encourage the user to stay on the platform and browse to the next ad slot.[3] The platform may draw utility not just from showing an ad in a particular ad slot, but also from the downstream effects of retaining a user's attention by showing them a non-paid piece of content.

Online ad platforms often refer to the (dis)utility that users experience when viewing an ad as the ad's quality. Though, they do not estimate ad quality on a user-specific basis, which is why it can be modeled by simply adding constants to α and β [44]. We measure what we call the user utility of a high-stakes ad as the number of individuals that get to see the ad, which we denote by $V(d) = \sum d$.[4] This allows us to evaluate the benefits and harms of a specific intervention, which may be desirable from a societal perspective. This is based on the assumption that, for those types of ads, users benefit solely from seeing them, e.g., because it exposes them to some kind of opportunity or resource. Furthermore, we assume that the online platform accurately estimates the user's probability of clicking on the ad, which we denote by $p = P[Y = 1]$.[5] Advertisement clicks

[1] In this context, the term accuracy is used to refer to any type of performance metric such as the utility of the platform in the case of online ad delivery.

[2] It is worth noting that the model follows a CPC model, but the outcome Y represents whether a user takes the advertiser's desired action, which is why our model could easily be generalized to other business objectives, representing any type of conversion.

[3] This formulation resembles a Vickrey-Clarke-Groves auction that includes organic content [48].

[4] We consider high-stakes ads to be those for which seeing them is clearly desirable (such as employment, housing, or financial opportunities) or undesirable (such as predatory lending services) [19, 45]. Hence, the user utility could easily be defined as a disutility if a high-stakes ad is harmful to see. For simplicity, we consider desirable high-stakes ads that provide an opportunity in the remainder of this paper.

[5] In contrast to other online ad simulations (such as [23]), our proposed model is implementation-agnostic: Our model is based on probabilities, while in reality, platforms estimate this probability using machine learning (ML) models, leveraging various user-specific data, such as online behavior, ad content, and temporal patterns and interactions [44].

follow a power-law distribution, and users' overall click rates are typically very low, usually less than 10% [1, 7]. We assume that fairness constraints are imposed based on the sensitive attribute A, which is sometimes referred to as the protected attribute. For simplicity, we consider fairness w.r.t. two groups, but our model generalizes to non-binary sensitive attributes.

3.2 Optimal ad delivery strategies

Following a CPC model, the platform's utility u for applying a decision rule d to a single user is determined by whether the high-stakes ad is displayed ($d \in \{0, 1\}$). If the ad is shown, the platform's utility can be calculated as the product of the probability of the user clicking on the ad and the payment made by the advertiser in the event of a click:

$$u(d) = \begin{cases} \alpha \cdot p, & \text{for } d = 1 \\ \beta, & \text{for } d = 0. \end{cases} \tag{1}$$

We use the capital letter $U(d)$ to denote the platform's total utility achieved for some decision rule d. A rational decision maker selects the decision rule that maximizes their expected utility: $\arg\max_d E(U(d))$. The platform's optimal unconstrained decision rule d^* takes the form of a uniform threshold rule:

$$d^* = \begin{cases} 1, & \text{for } p > \frac{\beta}{\alpha} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This shows that the threshold increases with β and decreases with α, i.e., the expected utility of showing the high-stakes ad relative to the alternative of not showing it determines the optimal decision threshold.

3.3 Group fairness criteria

Here, we provide a basic overview of each fairness constraint considered in this work. For a detailed description of the fairness constraints, see Appendix A. We refer to Baumann et al. [9] for an extensive overview of various sources of bias and their relation to unfairness.

Statistical parity requires that the proportion of individuals who receive a positive decision ($D = 1$, here: being shown the high-stakes ad) is the same across different groups defined by a sensitive attribute. **Equality of opportunity**, also known as TPR (true positive rate) parity, requires that the TPR (here: the share of people receiving a specific ad among all those that would click on it) is equal across different groups. **False positive rate** (FPR) parity is conceptually similar, but it measures on the proportion of people shown the ad among all those that would not click on it. **Predictive parity**, also known as PPV (positive predictive value) parity, requires that the positive predictive value (here: the proportion of individuals who click on the ad among those that are shown the ad, i.e., the click-through-rate (CTR)) is equal across different groups. **False omission rate** (FOR) parity is conceptually similar to predictive parity, but it focuses on the proportion of people who would click on the ad among all those that do not receive it.

3.4 Optimal ad delivery strategy under fairness constraints

We formulate algorithmic fairness as a constrained optimization problem, where an ad platform optimizes its utility while satisfying a predefined fairness constraint (FC) that represents a societal fairness desideratum:

$$\arg\max_d \quad E(U) \quad \text{subject to } FC. \tag{3}$$

If resources are limited, the total number of positive decisions ($\sum d$) can form a second constraint to Eq. (3).

The form of optimal fairness-constrained decision rules (denoted by d_c^*) depends on which FC is enforced. Hardt et al. [25] and Corbett-Davies et al. [15] show that optimal decision rules satisfying statistical parity, equality of opportunity, or FPR parity take the form of group-specific threshold rules:

$$d_c^* = \begin{cases} 1, & \text{for } p \geq \tau_a \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where τ_a denotes group-specific constants. On the other hand, Baumann et al. [10] showed that optimal decision rules satisfying predictive parity or FOR (false-omission rate) parity take the form of group-specific upper- or lower-bound decision rules.

$$d_c^* = \begin{cases} \left.\begin{cases} 1, & \text{for } p \geq \tau_a \\ 0, & \text{otherwise} \end{cases}\right\} \text{for } \bar{p} > BR_{A=a} \\ \left.\begin{cases} 1, & \text{for } p \leq \tau_a \\ 0, & \text{otherwise} \end{cases}\right\} \text{for } \bar{p} < BR_{A=a}, \end{cases} \tag{5}$$

where τ_a denote different group-specific constants, $\bar{p}$ denotes the optimal PPV in the case of the predictive parity requirement (1 minus the optimal FOR when enforcing FOR parity, respectively), and $BR_{A=a}$ denotes group a's base rate (BR)—which is defined as the ratio of individuals belonging to the positive class ($Y = 1$) from a set of individuals S_a belonging to a group $A = a$, also called *prevalence*: $BR_{A=a} = P[Y = 1|A = a] = \frac{1}{n_{A=a}} \sum_{i \in S_a} p_i$.

4 EXPERIMENTS

4.1 Assumptions

For each scenario, we generate synthetic datasets for two groups $a \in A$, (representing a binary sensitive attribute A, men (m) and women (w)) by sampling the active users' click probabilities P_a from a power-law distribution.[6] That is, P_a follows a power-law distribution with a shape parameter k_a: $P_a \sim \text{powerlaw}(k_a)$. We assume that the platform's utility for a user clicking on a high-stakes ad (α) is constant, i.e., it is the same for all individuals across all groups A. This resembles advertisers that are oblivious to the gender of the platform users they target for high-stakes ad impressions.[7] In contrast, we model β_a as a group-specific constant for individuals of a group $A = a$, which means that β_a is assumed to be the same for all individuals within a group a.

[6] The probability density function for the power-law is $f(x, k) = kx^{k-1}$. Usually, the power-law shape parameter is denoted by a, but here we use the alternative notation of k to avoid confusion with the sensitive attribute.

[7] For simplicity, we do not consider competitive spillovers for the high-stakes ad in question even though this may theoretically occur in practice.

4.2 Scenarios

We now describe the four different scenarios that we simulate using our proposed model. The clicking probability distribution of men and the utility of not showing them the high-stakes ad is fixed for all scenarios with $k_m = 0.05$ (see Fig. 4 in Appendix B.1) and $\beta_m = 0.03$. This corresponds to an average clicking probability of around 4.8% for men. On the other hand, the clicking probability distribution and the non-clicking utility of women vary across scenarios, as does the utility of clicking on a high-stakes ad. We use a grid of parameters shown in Table 1 to simulate the four scenarios. For each scenario, we sweep over different values of the parameters of interest and simulate the scenario for each of those values. Each simulation samples 1,000 men and 1,000 women and is repeated 30 times.

4.2.1 ***A: reference case*** *(no gender differences).* In this scenario, we consider a reference case where there are no gender differences. Both men and women have the same clicking probability distribution P with $k_m = k_w = 0.05$, and the platform's expected utility for users not seeing the high-stakes ad is the same irrespective of their group membership ($\beta_m = \beta_w = 0.03$). We simulate this scenario for ten different values of α, ranging from 0.03 to 1.

4.2.2 ***B: competitive spillovers*** *($\beta_m \neq \beta_w$).* The clicking probability distribution for men and women remains the same as in Scenario A ($k_m = k_w = 0.05$). However, in this scenario, we fix $\alpha = 0.2$ and simulate ten different values of β_w, ranging from 0.03 to 1. Recall that β denotes the utility gained by the platform if the high-stakes ad is not shown ($D = 0$), e.g., if the best alternative ad is shown. For men, this is fixed at $\beta_m = 0.03$, as in Scenario A. The platform's higher utility for women than for men if the high-stakes ad is not shown ($\beta_w < \beta_m$) represents competitive spillovers [31].

4.2.3 ***C: base rate (BR) differences*** *($P_m \neq P_w$).* In this scenario, we introduce BR differences between men and women, i.e., that there are gender differences in the probability of clicking an ad. The clicking probability distribution for men remains the same as in Scenario A ($k_m = 0.05$), while we simulate ten different clicking probability distributions for women using values ranging from 0.05 to 0.005 for the power-law distribution's shape parameter k_w. This corresponds to an average clicking probability in the range between 0.5% (for $k_w = 0.005$) and 4.8% (for $k_w = 0.05$) compared to a fixed probability of 4.8% for men, as visualized in Fig. 5 in Appendix B.1. There are no group differences for the remaining parameters with $\alpha = 1$ and $k_m = k_w = 0.03$.

4.2.4 ***D: competitive spillovers & BR differences*** *($P_m \neq P_w$ and $\beta_m \neq \beta_w$).* In this scenario, we combine the competitive spillovers and base rate (BR) differences between men and women from Scenarios B and C. We use the values $k_m = 0.05$ and $k_w = 0.01$ for the click probability distribution for men and women. This means that the women's BR is lower, i.e., they are less likely to click on the high-stakes ad on average (1% of women and 4.7% of men). Similar to Scenario B, we then simulate different values ranging from 0.01 to 0.1 for the platform's expected individual utility when they decide not to show the high-stakes ad to women.

For each scenario, we first simulate the optimal ad delivery strategy from the perspective of the online ad platform. This represents the situation of an ad platform that wants to maximize its total utility without considering any potential unfairness for the platform's users, resulting in the optimal unconstrained decision rule d^* (see Eq. (2)). In a second step, we add different FC to those ad delivery strategies, resulting in d_c^*, as described in Section 3.4. We compare the outcomes resulting from applying these strategies for men and women and measure fairness using the metrics introduced in Section 2.2.

Table 1: Grid of parameter used for the four different experiment scenarios

Scenario	α	men		women	
		k_m	β_m	k_w	β_w
A	0.03–1	0.05	0.03	0.05	0.03
B	0.2				0.03–1
C	1			0.05–0.005	0.03
D	0.2			0.01	0.03–1

Without a fixed number of impressions, adding a FC may result in the high-stakes ad being shown to fewer individuals, compared to the unconstrained case. Thus, we additionally run each scenario with the added constraint of $\sum d^* = \sum d_c^*$, requiring the number of impressions to remain constant, irrespective of whether the platform's decision rule is constrained or not—the reasons for this choice will be described in more detail in Section 5.3.

5 RESULTS

5.1 Enforcing fairness reduces platform utility

Fig. 1 visualizes the cost of fair ad delivery associated with enforcing different FC. We investigate the sensitivity of those costs by sweeping over different parameters, whose values are shown on the x-axis—as described in Table 1. The y-axis shows the costs of fairness for different FC. Those costs are presented in % of the utility achieved in the unconstrained case, i.e., $\frac{U(d_c^*)}{U(d^*)} \cdot 100$, for any given simulation setup. Thus, it is always 100% for the unconstrained case.

In scenario A, where there are no gender differences whatsoever, fairness can be achieved at no cost. This is because there are no group-specific differences to take into account. In scenario B, where there are competitive spillovers, there is a cost in utility to ensure fairness, even though the click probability distributions of men and women are the same. However, since satisfying any type of fairness also ensures all other types of fairness, the cost in terms of utility is the same for any type of fairness. In contrast to the reference case (scenario A), the presence of competitive spillovers leads to women having a higher average utility gain from showing any other ad (other than a high-stakes ad) than men. Therefore, a higher fraction of men see the high-stakes ad compared to the fraction of women seeing it—even absent any discriminatory ad targeting, i.e., even if price bids are the same for users of both groups. Mitigating this disparity requires counteracting the competitive spillovers to ensure that the same fraction of men and women see the high-stakes ad. Notice that fair ad delivery costs increase with a larger magnitude of competitive spillovers (i.e., with larger differences between β_m and β_w) but then slowly decrease again. This is due to the fact that the optimal threshold, and thus also the optimal acceptance rate,

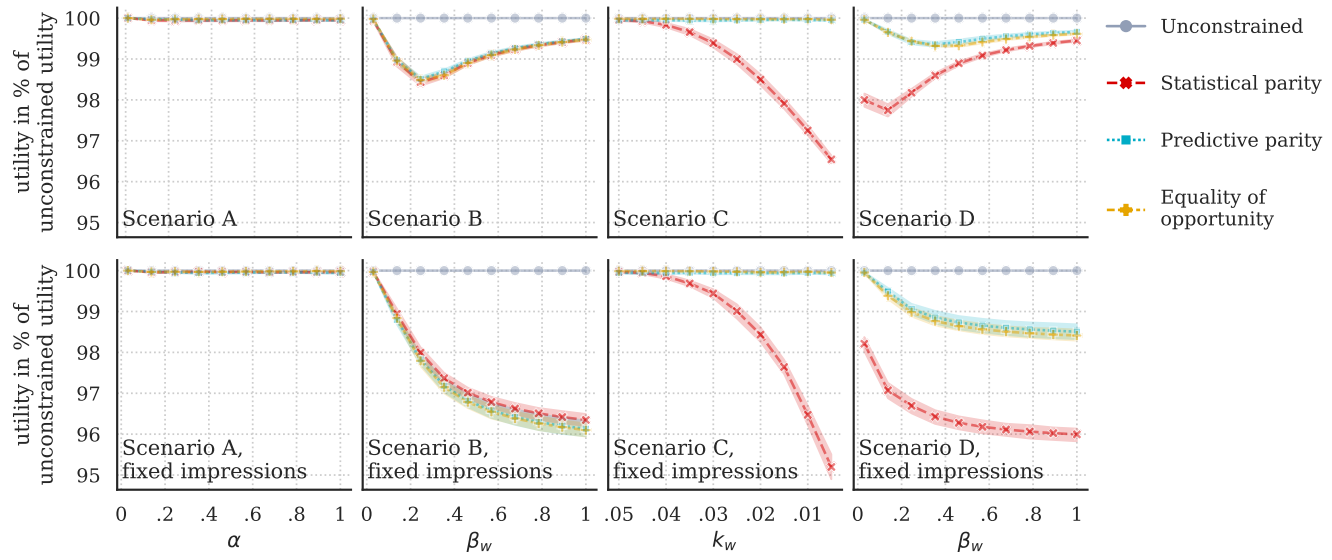


Figure 1: Average cost of different fairness criteria (in % of utility achieved absent any fairness constraint, i.e., $\frac{U(d_c^*)}{U(d^*)} \cdot 100$) for the four scenarios: with (below) and without (above) a fixed impression constraint. Shaded intervals reflect 95% confidence intervals from variation across repeated simulations.

is strictly tied to the group-specific utility gained by not showing the high-stakes ad β_a, all else equal. The acceptance rate of men remains constant, whereas women's acceptance rate decreases with increasing values for β_w. In contrast, under any FC, it is optimal to decrease the acceptance rates for m and w with increasing values for β_w, and for $\beta_w \geq 0.353$ it is even optimal not to show the high-stakes ad to anyone. This is visualized in detail in Fig. 7b

In scenario C, the click probability distributions differ between the two groups, with women being less likely to click on the high-stakes ad, on average, i.e., $k_w < k_m$. We find that the FC predictive parity and equality of opportunity can be achieved at almost no cost. For $k_w = 0.005$, the utility decreases by just $\sim 0.01\%$ when enforcing predictive parity compared to the unconstrained case. The biggest cost is observed for statistical parity, which measures the disparity in the overall percentage of high-stakes ad delivery across sensitive groups. For $k_w = 0.005$ the platform loses 4% of its utility, on average, when ensuring statistical parity compared to the unconstrained case. This represents a 400-fold increase in costs compared to the fairness criteria equality of opportunity and predictive parity (i.e., equal CTRs). Although the 4% reduction may appear to be small, it is important to consider that this decrease corresponds to the revenue generated from all ad auctions where the high-stakes ad was a contender. As a result, the costs of achieving fairness could be substantial for an industry that generates nearly $100 billion in revenue annually [27].

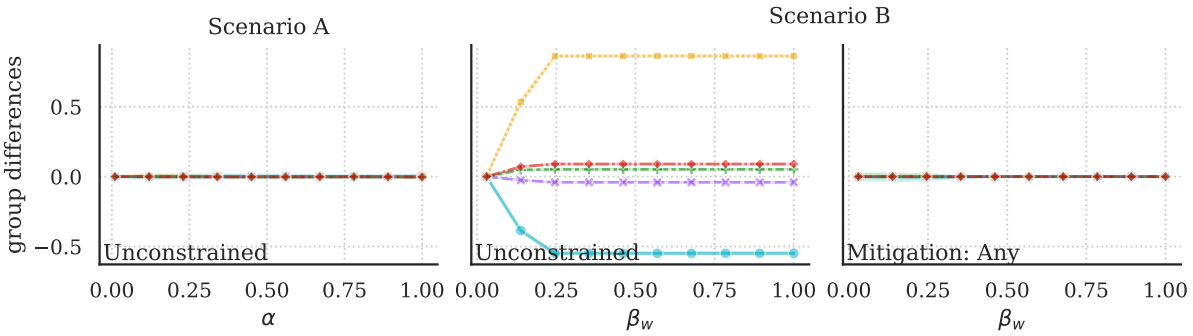
The results for Scenario D, which combines BR differences and competitive spillovers, are similar to those of Scenarios B and C. For $\beta_w = 0.03$, there are no competitive spillovers, resulting in higher costs for FPR parity and especially for statistical parity. Just as in Scenario B, for small competitive spillovers, the costs increase, but for very large competitive spillovers, they decrease since the fair dissemination of the high-stakes ads becomes less and less lucrative relative to the high utility that can be gained by not showing the high-stakes ad ($D = 0$).

The results presented in this section are sensitive to the platform user groups' BRs. For very low BRs, adjusting decision thresholds for the high-stakes ad is cheap as there a very few users that would click on the high-stakes ad anyway. However, for larger BRs and BR differences between groups, ensuring fairness is much more costly. We provide the results for a batch of simulations using parameters k_a that represent this case in Appendix B.2.

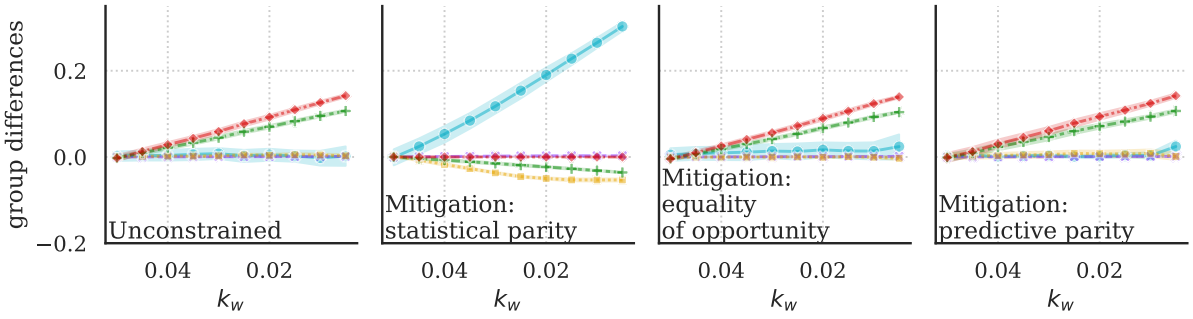
5.2 Tradeoffs between fairness criteria: enforcing one constraint conflicts with others

By definition, all fairness criteria are met in scenario A since there are no group-specific differences, as demonstrated in Fig. 2a. In cases where there is no difference in the click probability distributions between the two groups, satisfying any type of fairness ensures all other types of fairness at no additional cost. I.e., the different FC can be achieved at the same time since the BR does not differ across groups. Thus, for scenario B, enforcing any type of fairness through post-processing produces fairness with respect to all metrics.

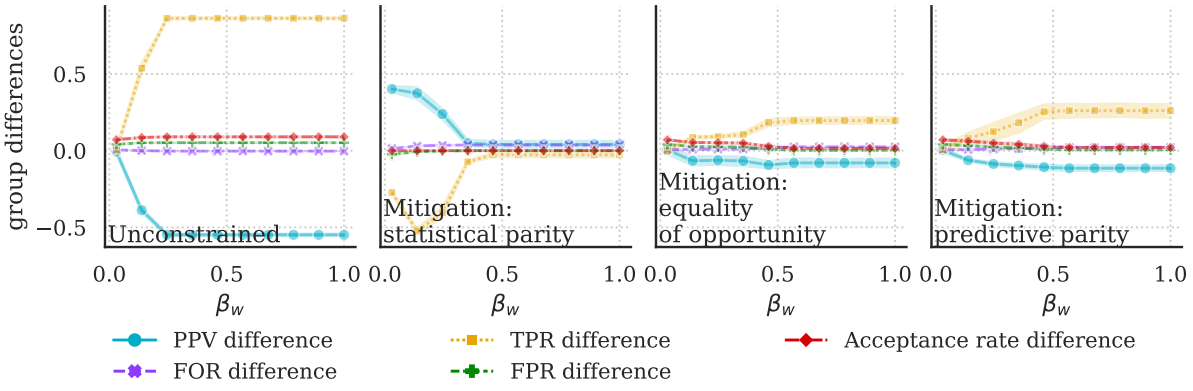
For scenario C, in addition to the fairness-utility tradeoff, there are tradeoffs between different notions of fairness due to the difference in BRs, as shown in Fig. 2b. The optimal post-processing techniques provided by [10, 15, 25] work effectively, but depending on the specific constraint that is enforced, the side-effects in terms of other notions of fairness differ. This is in line with the theoretical fairness impossibility results [13, 29]. Enforcing statistical parity comes at a huge cost in predictive parity, as can be seen in Fig. 2b. In the unconstrained case, PPVs are similar for men and women due to the similar shape of the tail of the click probability distribution



(a) Scenario A: Fairness is always satisfied without any group differences. Scenario B: Any mitigation will lead to satisfying all FC.



(b) Scenario C



(c) Scenario D

Figure 2: Between fairness tradeoffs when enforcing different fairness criteria. Positive values represent higher rates for men than for women; e.g., the PPV difference is calculated by subtracting the women's PPV from the men's PPV. The legend in (d) applies to all panels. Shaded intervals reflect 95% confidence intervals from variation across repeated simulations.

(as visible with the almost identical blue lines in the top left panel of Fig. 7c). As this is not the case for acceptance rates and TPR, a larger deviation from the unconstrained optimum (i.e., a larger change in the group-specific threshold away from the single uniform threshold) is necessary to ensure statistical parity —as visible with the diverging blue lines in the top right and bottom right panels of Fig. 7c. This is the reason why it is not only more costly (in terms of platform utility) to ensure those metrics (as explained in detail in Section 5.1) but also in conflict with predictive parity. Another side-effect of ensuring statistical parity is that the resulting TPR of women (as well as their FPR) is slightly higher than the one of men. For this same reason, ensuring predictive parity or equality of opportunity is not only cheap but also has a marginal effect on other fairness notions (as can be seen with the almost identical plots in Fig. 2b). The results of scenario D are conceptually similar to those of scenarios B and C, as there are both competitive spillovers and BR differences across groups.

5.3 Enforcing fairness leads to "leveling down" in the presence of competitive spillovers

Scenarios B and D both suffer from "leveling down": Introducing a FC requires the platform to deviate from its optimal unconstrained decision rule in order to achieve equality w.r.t. a certain parity metric. "Leveling down" is defined as the situation in which enforcing fairness harms some group without benefiting any group [16, 34, 37], i.e., if the values are smaller than or equal to zero in Fig. 3. For large competitive spillovers (simulated with increasing values for β_w to increase the difference between β_m and β_w), the optimal fair solution for the platform is to reduce the number of high-stakes ad impressions for men without changing the number of high-stakes ad impressions for women, compared to the unconstrained case. This leads to a decrease in utility for users of the platform. Specifically, men experience a decrease in utility while women's utility remains unchanged. This is visualized in the light blue and light orange lines in Fig. 3, where the y-axis represents the difference between the utility in the unconstrained vs. in the fairness-constrained, i.e., $V(d_c^*) - V(d^*)$. The rationale for the "leveling down" effect is the fact that with large competitive spillovers, achieving equality by harming the better-off group becomes optimal from the platform's perspective, resulting in a situation where no one benefits from the added FC.

To avoid "leveling down", we reran all the simulations described in Section 4.2 with an additional constraint of a fixed total number of high-stakes ad impressions for the fairness-constraint simulations, depending on the number of users who would get to see the high-stakes ad in the unconstrained case: $\sum d^* = \sum d_c^*$. As can be seen in Fig. 3, this prevents the platform from achieving equality by harming the better-off group and ensures that the burden of the added FC is on the platform. By maintaining a constant number of high-stakes ad impressions, we can ensure that the added FC benefits at least one of the groups. In this case, ensuring fairness while keeping the total number of impressions constant increases the user utility for women and decreases the utility for men. However, preventing the platform from reducing the total number of high-stakes ad impressions comes at a cost in terms of platform utility, shifting the cost of fairness from the users to the platform. This new cost in terms of platform utility is shown in Fig. 1 (below), showing that adding an impression constraint shifts the cost of fairness from the users to the platform. Consequently, it leads to an increase in the cost of fairness in terms of the platform's utility, relative to the utility derived in the unconstrained case, for scenarios B, C, and D. However, the increase is more pronounced for the scenarios including competitive spillovers, i.e., scenarios B and D.[8] This observation highlights the tradeoff between utility-maximizing platforms that simply aim to meet an egalitarian notion of fairness and society, which aims to ensure fairness in a way that is beneficial for historically disadvantaged user groups.

[8]Notice that the added impression constraint yields a monotonically increasing cost of fairness with increasing competitive spillovers. This occurs because the platform cannot just show fewer ads to all users in order to to achieve a given fairness objective, which would be the optimal strategy in the case of large competitive spillovers without the additional impression constraint, as shown in Fig. 7b and 7d.

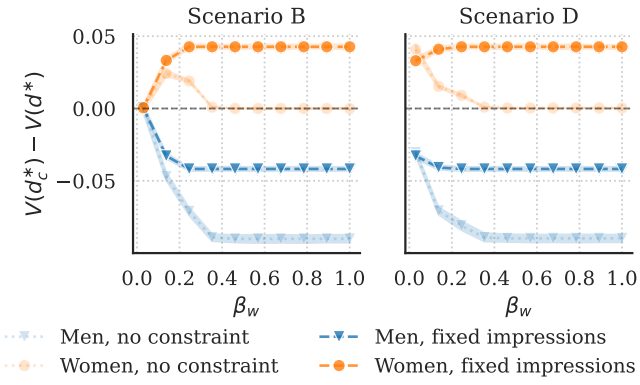


Figure 3: "Leveling down" user utility when enforcing fairness in the presence of large competitive spillovers. Without impression constraints (transparent lines) men's utility is decreased while women's utility is unchanged. Forcing the platform to retain the same number of impressions as without a fairness intervention (opaque lines) ensures that no leveling down occurs.

6 DISCUSSION

Algorithmic advertisement (ad) delivery systems determine which ads users see, forming the backbone of a multibillion-dollar industry that powers major tech platforms such as Meta, Google, Bing, and X. These platforms use vast amounts of collected user data to steer ads towards strategically selected sub-groups among the targeted audiences. They have recently faced criticism for their ad systems' emphasis on utility-maximization, which can lead to unfair access to opportunity for historically disadvantaged communities.

In this work, we investigated the downstream effects of introducing a variety of fairness interventions in an attempt to remedy discriminatory ad delivery effects, such as entrenching housing segregation or gender differences in access to certain career paths. The results of our simulated experiments show that applying the optimal unconstrained decision rule (maximizing the platform utility) does not *fully* satisfy any definition of fairness. Instead, fairness constraints must be explicitly enforced to ensure any given notion of fairness. This can be achieved using post-processing approaches [10, 15, 25].

Several tradeoffs emerge when enforcing fairness constraints. First, there exists a tradeoff between the ad platform's utility and fairness for the users. The cost of fairness in terms of the online platform's utility depends on the type of fairness considered. We find that platform's costs of ensuring statistical parity are significantly higher than other criteria. On the other hand, our results show that this tradeoff is negligible when equality of opportunity is the only criterion considered. These findings also highlight the 'impossibility of fairness' effect, which means there is a tradeoff between different notions of fairness [13, 29]. This means that enforcing a fairness notion may have positive effects on some other fairness notions, but negative effects on others. For example, ensuring statistical parity results in a violation of predictive parity if the clicking probability distributions differ across groups.

Table 2: Summary of the key insights and policy recommendations to improve the fairness of algorithmic ad delivery systems.

Key insights:		Policy recommendations:
Efficiency-driven online ad systems inherently exhibit unfairness, even with restricted targeting options.	→	As harmful effects may stem from the process of optimizing ad delivery, regulation must go beyond limiting the use of sensitive information for ad targeting. Independent audits of ad delivery systems should be mandated.
Prioritizing fairness can lead to unintended consequences, such as a leveling down effect (achieving fairness by harming the privileged group without uplifting others).	→	As side-effects of fairness-enhancing interventions can be undesirable from a societal perspective, any audits introduced as part of settlements or regulation (like the DMA, the DSA, or the AI Act in the European Union) must also monitor (1) by what means exactly fairness is achieved and (2) whether platforms are passing the cost onto the end-users.
Inherent differences among targeted user groups often preclude satisfying multiple fairness metrics simultaneously.	→	The selection of a fairness metrics determines who benefits from the intervention. Thus, the choice of a metric requires substantial evaluation.

Our findings also contribute to the debate on the "leveling down" objection, which challenges the idea of egalitarianism and suggests that enforcing group fairness for ad impressions on online platforms may lead to worse outcomes for all groups [16, 26, 34, 37]. We demonstrate that enforcing fairness in the presence of large competitive spillovers can lead to the leveling down effect, as achieving equality by harming the better-off group becomes optimal from the platform's perspective. We show that this scenario can be prevented if the platform enforces a fixed number of high-stakes ad impressions in both unconstrained and constrained scenarios. As a result, the cost of the fairness intervention would be carried by the platform, rather than by its users. While our results imply that the costs of fairness can be shifted from the users to the platform, in practice, platforms could still pass them on to the advertisers, unless specifically prevented from doing so by regulation. Therefore, any proposed policy should consider the downstream effects of enforcing any fairness criteria on the cost, and availability of high-stakes advertising.

Taken together, our findings are crucial in light of the current discussion of what a fair ad delivery system should look like, which fairness criteria are defensible and what downstream effects they would lead to, as well as who should carry the burden of introducing the constraints.

Policy recommendations. Significant regulatory responses have been initiated in the US (in the form of the Department of Justice's lawsuit against Meta [47]) and Europe (in the form of the Digital Services Act, the Digital Markets Act, and the recently approved Artificial Intelligence Act). However, our findings indicate that existing measures may not suffice. Effective external monitoring is essential to ensure fair ad delivery, and potential side-effects must be carefully considered and addressed. Table 2 details specific policy recommendations derived from our insights.

Limitations. Our results suffer from several limitations: While abstracting away the auction mechanism allows us to investigate the effects of enforcing fairness of online ad delivery in isolation, in practice, the exact implementation of the auction mechanism might have an effect on the platform's room for maneuver. Further, in our simulation, we do not model within-group differences in impression costs; in practice, the actual utility costs of enforcing fairness might be lower. In addition, more research is needed to investigate the long-term effects of enforcing fairness on the online ad market and whether advertisers can game the system if they know of the platform's fairness adjustments in their implementation.

7 CONCLUSION

Unfair delivery of high-stakes ads online has significant societal effects. It can perpetuate and exacerbate existing inequalities and discrimination by reinforcing stereotypes and biases that have far-reaching consequences for individuals and society as a whole. In this paper, we focus on the goal of ensuring equitable access to relevant information that is shared through online advertisement (such as housing or job opportunities). In conclusion, our study highlights the need for explicit enforcement of fairness in online ad delivery systems. Fairness is not satisfied by default, and simply restricting advertisers' targeting options does not ensure fairness. Our study indicates that while ensuring fairness in online ad systems reduces platform utility, achieving predictive parity or equality of opportunity generally incurs a lower utility cost than achieving statistical parity. Our findings reveal tradeoffs between different notions of fairness; for example, enforcing statistical parity leads to larger disparities in CTRs between groups. Moreover, pursuing fairness can result in potentially undesirable side-effects, such as "leveling down" effects that occur in the presence of competitive spillovers, which must be actively mitigated. As online programmatic advertising continues to evolve, our findings are important for ensuring fairness of ad impressions for historically marginalized groups on online platforms.

ACKNOWLEDGMENTS

We would like to thank Aleksandra Urman and Stefania Ionescu for their invaluable feedback throughout this work. We also thank the four anonymous reviewers for their helpful suggestions.

This work was supported by the National Research Programme "Digital Transformation" (NRP 77) of the Swiss National Science Foundation (SNSF)—grant number 187473, and by the National Science Foundation (NSF)—grant CNS-2318290.

REFERENCES

[1] Deepak Agarwal, Andrei Zary Broder, Deepayan Chakrabarti, Dejan Diklic, Vanja Josifovski, and Mayssam Sayyadian. 2007. Estimating Rates of Rare Events at Multiple Resolutions. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 16–25.

[2] John Albert. 2023. Not a solution: Meta's new AI system to contain discriminatory ads. *AlgorithmWatch* (2023). https://algorithmwatch.org/meta-discriminatory-ads

[3] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proc. ACM Hum.-Comput. Interact.* 3 (2019).

[4] Julia Angwin, Noam Scheiber, and Ariana Tobin. 2017. Dozens of Companies Are Using Facebook to Exclude Older Workers From Job Ads. *ProPublica* (dec 2017).

[5] Julia Angwin, Ariana Tobin, and Madeleine Varner. 2017. Facebook (still) letting housing advertisers exclude users by race. *ProPublica* 21 (2017).

[6] Chloé Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, et al. 2021. Fairness on the ground: Applying algorithmic fairness approaches to production systems. *ArXiv preprint arXiv:2103.06172* (2021).

[7] Eytan Bakshy, Dean Eckles, Rong Yan, and Itamar Rosenn. 2012. Social Influence in Social Advertising: Evidence from Field Experiments. In *Proceedings of the 13th ACM Conference on Electronic Commerce*. 146–161.

[8] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

[9] Joachim Baumann, Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. 2023. Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. 1002–1013.

[10] Joachim Baumann, Anikó Hannák, and Christoph Heitz. 2022. Enforcing Group Fairness in Algorithmic Decision Making: Utility Maximization Under Sufficiency. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. 2315–2326.

[11] Elisa Celis, Anay Mehrotra, and Nisheeth Vishnoi. 2019. Toward Controlling Discrimination in Online Ad Auctions. In *Proceedings of the 36th International Conference on Machine Learning*, Vol. 97. 4456–4465.

[12] Andrea Celli, Riccardo Colini-Baldeschi, Christian Kroer, and Eric Sodomka. 2022. The parity ray regularizer for pacing in auction markets. In *Proceedings of the ACM Web Conference 2022*. 162–172.

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

[14] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. The Measure and Mismeasure of Fairness. *Journal of Machine Learning Research* 24, 312 (2023), 1–117.

[15] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 797–806.

[16] Roger Crisp. 2003. Equality, Priority, and Compassion. *Ethics* 113, 4 (2003), 745–763.

[17] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.

[18] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*. Technical Report. Northpoint Inc. https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/

[19] Emily Dreyfuss. 2019. Facebook Changes Its Ad Tech to Stop Discrimination. WIRED. https://www.wired.com/story/facebook-advertising-discrimination-settlement

[20] European Commission. 2024. Remarks by Executive-Vice President Vestager and Commissioner Breton on the opening of non-compliance investigations under the Digital Markets Act. https://ec.europa.eu/commission/presscorner/detail/en/speech_24_1702

[21] European Parliament. 2024. Artificial Intelligence Act. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf

[22] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 329–338.

[23] Lodewijk Gelauff, Ashish Goel, Kamesh Munagala, and Sravya Yandamuri. 2020. Advertising for demographically fair outcomes. *ArXiv preprint arXiv:2006.03983* (2020).

[24] Guidehouse Inc. 2023. VRS Compliance Metrics Verification. (2023). https://www.justice.gov/usao-sdny/file/1306631/dl

[25] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*. 3323–3331.

[26] Lily Hu and Yiling Chen. 2020. Fair Classification and Social Welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 535–545.

[27] IAB. 2018. Digital Ad Spend Hits Record-Breaking $49.5 Billion in First Half of 2018, Marking a Significant 23% YOY Increase. https://www.iab.com/news/digital-ad-spend-hits-record-breaking-49-5-billion-in-first-half-of-2018/

[28] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2021. Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the Web Conference 2021*. 3767–3778.

[29] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *ArXiv preprint arXiv:1609.05807* (2016).

[30] Allison Koenecke, Eric Giannella, Robb Willer, and Sharad Goel. 2023. Popular Support for Balancing Equity and Efficiency in Resource Allocation: A Case Study in Online Advertising to Increase Welfare Program Awareness. *Proceedings of the International AAAI Conference on Web and Social Media* 17, 1 (2023), 494–506.

[31] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science* 65, 7 (jul 2019), 2966–2981.

[32] Anja Lambrecht and Catherine E Tucker. 2020. Apparent algorithmic discrimination and real-time algorithmic learning in digital search advertising. *Available at SSRN 3570076* (2020).

[33] Meta. 2023. About Ad Delivery. https://www.facebook.com/business/help/1000688343301256

[34] Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default. *ArXiv preprint arXiv:2302.02404* (2023).

[35] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*.

[36] Milad Nasr and Michael Carl Tschantz. 2020. Bidding Strategies with Gender Nondiscrimination Constraints for Online Ad Auctions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 337–347.

[37] Derek Parfit. 1995. *Equality or priority*. Department of Philosophy, University of Kansas.

[38] Alexander Peysakhovich, Christian Kroer, and Nicolas Usunier. 2022. Parity in Markets–Methods, Costs, and Consequences. *ArXiv preprint arXiv:2210.02586* (2022).

[39] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence* 3, 10 (2021), 896–904.

[40] Kit T. Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. 2020. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. 142–153.

[41] Piotr Sapiezynski, Avijit Ghosh, Levi Kaplan, Aaron Rieke, and Alan Mislove. 2022. Algorithms That "Don't See Color": Measuring Biases in Lookalike and Special Ad Audiences. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 609–616.

[42] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for discrimination in online targeted advertising. In *Conference on fairness, accountability and transparency*. PMLR, 5–19.

[43] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[44] Aditya Srinivas Timmaraju, Mehdi Mashayekhi, Mingliang Chen, Qi Zeng, Quintin Fettes, Wesley Cheung, Yihan Xiao, Manojkumar Rangasamy Kannadasan, Pushkar Tripathi, Sean Gahagan, Miranda Bogen, and Rob Roudani. 2023. Towards Fairness in Personalized Ads Using Impression Variance Aware Reinforcement Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*. 4937–4947.

[45] Ariana Tobin. 2019. HUD sues Facebook over housing discrimination and says the company's algorithms have made the problem worse. *ProPublica* (2019).

[46] American Civil Liberties Union. 2019. Facebook Agrees to Sweeping Reforms to Curb Discriminatory Ad Targeting Practices. https://www.aclu.org/press-releases/facebook-agrees-sweeping-reforms-curb-discriminatory-ad-targeting-practices

[47] U.S. Department of Justice. 2022. United States of America v. Meta Platforms, Case 22 Civ. 5187.

[48] Hal R Varian and Christopher Harris. 2014. The VCG auction in theory and practice. *American Economic Review* 104, 5 (2014), 442–445.

[49] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness (FairWare '18)*. 1–7.

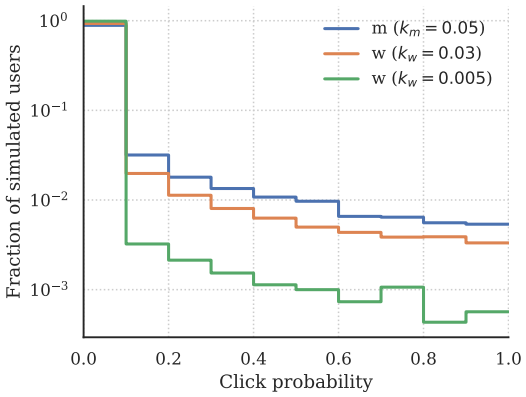


Figure 4: Click probability distributions P_a (log scale)

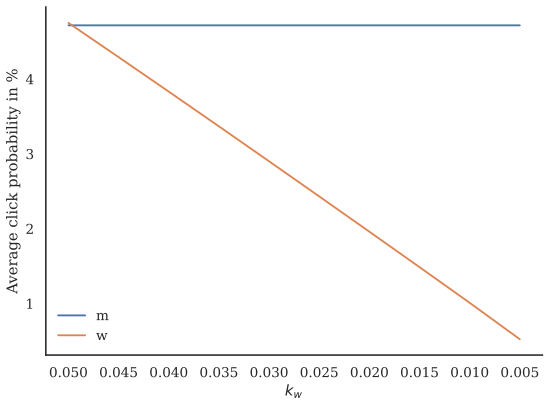


Figure 5: Scenario C: average probabilities of clicking on ad for men (m) and women (w)

A GROUP FAIRNESS CRITERIA

Table 3 lists the mathematical constraints associated with the fairness criteria we consider in this paper. Additionally, it provides a reference for the optimal post-processing unfairness mitigation solutions that exist for each of those criteria.

Statistical parity requires that the proportion of individuals who receive a positive decision ($D = 1$) is the same across different groups defined by a sensitive attribute A (e.g., race or gender). In the context of online ad delivery, this means that the proportion of individuals who are shown an ad should be the same across different groups defined by A.

Equality of opportunity, also known as TPR parity, requires that the true positive rate (TPR) is equal across different groups. The TPR is the proportion of individuals who are correctly identified as belonging to the positive class ($D = 1$) out of all individuals who actually belong to the positive class ($Y = 1$). In the context of online ad delivery, the TPR refers to the share of people seeing a specific ad ($D = 1$) among all those that would click on the ad ($Y = 1$). False positive rate (FPR) parity is conceptually similar to equality of opportunity, but it focuses on the proportion of people seeing the ad among all those that would not click on it (i.e., that belong to the negative class $Y = 0$).

Predictive parity, also known as PPV parity, requires that the positive predictive value (PPV) is the same across different groups A. The PPV is the proportion of individuals who actually belong to the positive class ($Y = 1$) out of all individuals who are identified as such ($D = 1$). In the context of online ad delivery, the PPV corresponds to the proportion of individuals clicking on an ad ($Y = 1$) among all those that got to see it ($D = 1$). Hence, enforcing predictive parity is equivalent to equalizing click-through rates (CTR). False omission rate (FOR) parity is conceptually similar to predictive parity, but it focuses on the proportion of people who would click on the ad ($Y = 1$) among all those that do not see it ($D = 0$).

B ADDITIONAL EXPERIMENTAL RESULTS

B.1 Additional experimental details

Fig. 4 shows the click probability distributions P_a (following a power-law distribution) on a log scale for different values of k_a for $a \in \{m, w\}$. The plot shows that decreasing the value of the shape parameter k_a leads to a more right-skewed distribution with a long tail on the right side, indicating users' low clicking probabilities.

Fig. 5 visualizes the average probabilities of clicking on an ad for men (m) and women (w) associated with a given value for k_a on the x-axis (as by the range of values simulated in scenario C).

Fig. 7 visualizes group-specific metrics when enforcing different fairness criteria for all four scenarios. This figure is similar to Fig. 2, but instead of visualizing differences, it shows the specific values for men and women. For example, instead of visualizing the difference between the acceptance rates of men and women, it visualizes the acceptance rates of men and women separately. As in Fig. 2, the x-axis represents the values of a certain scenario-specific parameter we sweep over to investigate the sensitivity (see Table 1).

B.2 Larger BR differences between men and women increase the cost of fairness

We run a second batch of simulations, using the same scenarios as in Section 4.2 but with bigger base rate (BR) differences between men (m) and women (w). Table 4 list the parameters used for these scenarios, which we denote by A2, B2, C2, and D2. Notice that those simulations only differ in their shape parameters for the clicking distribution. In particular, we use a larger shape parameter k_m, which represents a larger average probability of clicking on the high-stakes ad for men ($\sim 28\%$), and a different range of simulated k_w for the scenario C2. All other parameters are equivalent to the corresponding scenarios A, B, C, and D—see Table 1.

Fig. 6 shows the cost of different fairness criteria (in % of total utility) for the four scenarios A2, B2, C2, and D2. As can be seen on the y axis, the effect size increases compared to the scenarios A, B, C, and D, where men's average clicking probability for the high-stakes ad is lower. If men are, on average, more likely to click on the high-stakes ad, men are more likely to see this high-stakes ad (e.g., since advertisers are willing to pay more for their high-stakes ad to be shown to men). The increased cost of fairness is a direct consequence of this increased BR differences between men and women, as this means that enforcing some FC results in changing the high-stakes ad impression decision for more platform users.

Table 3: Group fairness criteria

Fairness criterion	Mathematical constraint	Post-processing unfairness mitigation
Statistical parity	$P(D=1 \mid A=0)=P(D=1 \mid A=1)$	Corbett-Davies et al. [15]
Equality of opportunity	$P(D=1 \mid Y=1, A=0)=P(D=1 \mid Y=1, A=1)$	Corbett-Davies et al. [15], Hardt et al. [25]
FPR parity	$P(D=1 \mid Y=0, A=0)=P(D=1 \mid Y=0, A=1)$	Hardt et al. [25]
Predictive parity	$P(Y=1 \mid D=1, A=0)=P(Y=1 \mid D=1, A=1)$	Baumann et al. [10]
FOR parity	$P(Y=1 \mid D=0, A=0)=P(Y=1 \mid D=0, A=1)$	Baumann et al. [10]

Table 4: Parameter grid for scenarios A2, B2, C2, and D2.

Scenario	α	men		women	
		k_m	β_m	k_w	β_w
A2: Reference case	0.03–1	0.4	0.03	0.4	0.03
B2: competitive spillovers	0.2				0.03–1
C2: base rate (BR) differences	1			0.4–0.01	0.03
D2: competitive spillovers & BR differences	0.2			0.01	0.03–1

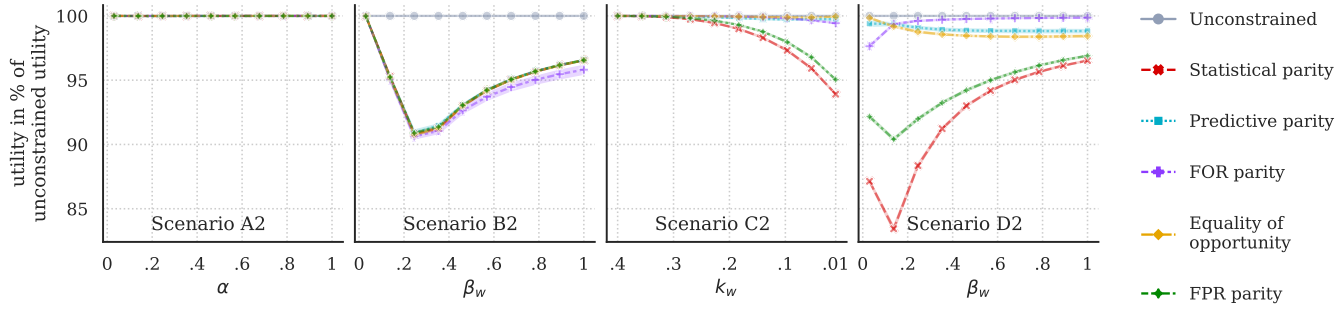
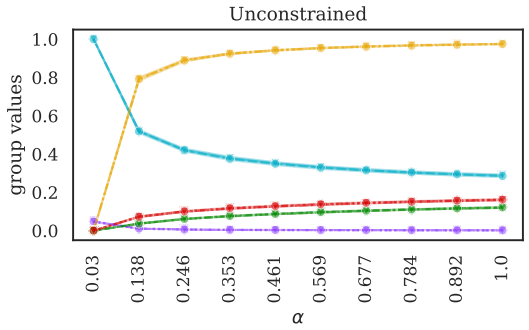
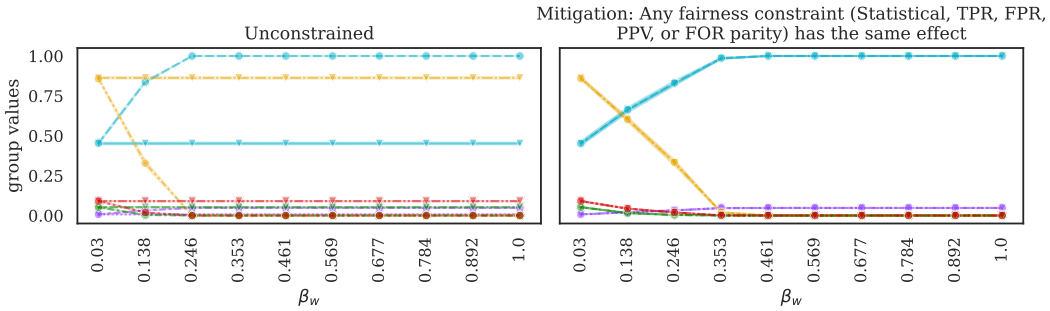


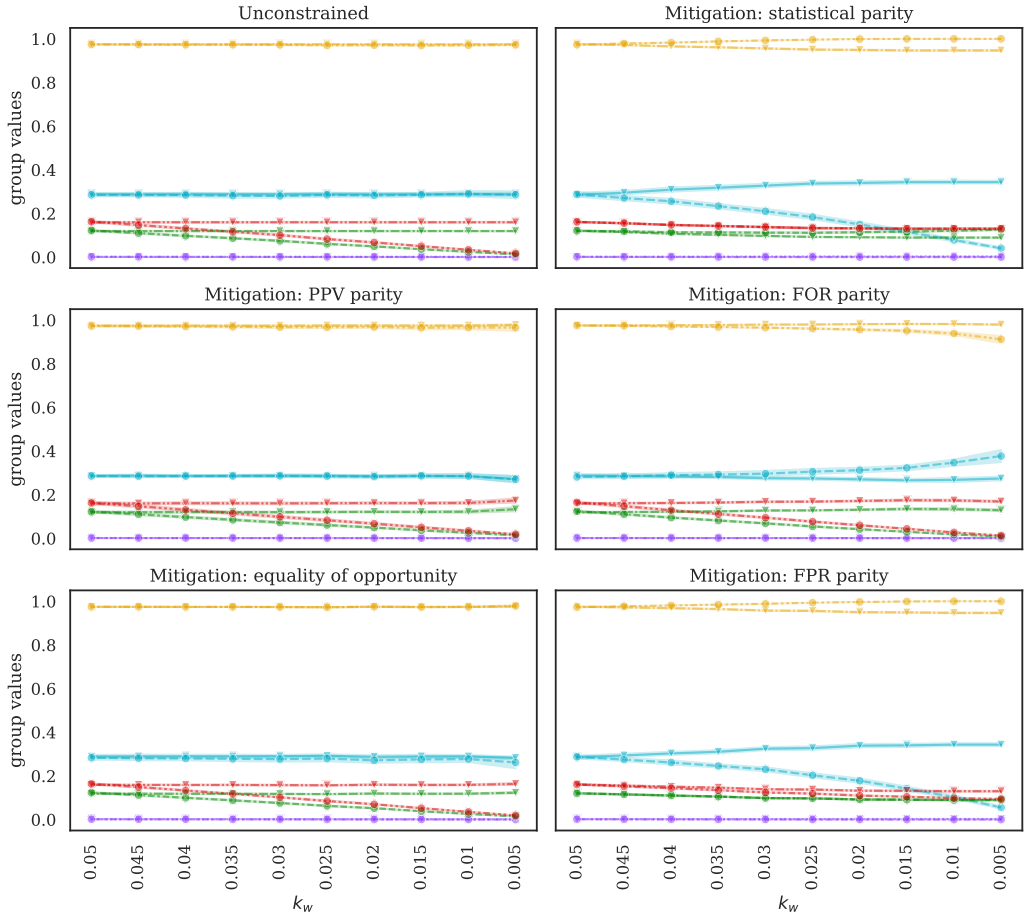
Figure 6: Average cost of different fairness criteria (in % of utility achieved absent any fairness constraint, i.e., $\frac{U(d_c^*)}{U(d^*)} \cdot 100$) for the four scenarios A2, B2, C2, and D2. Shaded intervals reflect 95% confidence intervals from variation across repeated simulations.



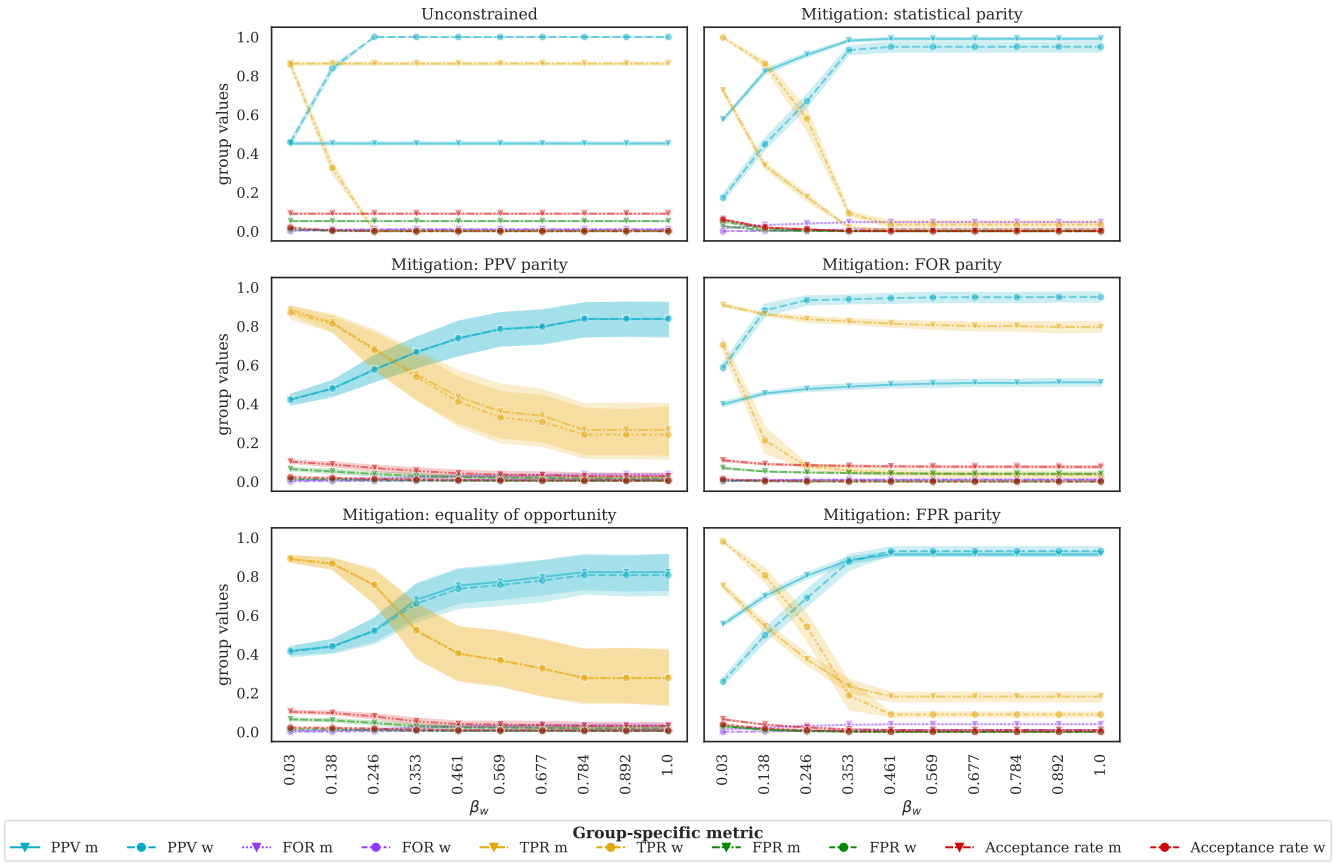
(a) Scenario A: without any differences between m and w, fairness is always satisfied



(b) Scenario B



(c) Scenario C



(d) Scenario D

Figure 7: Group-specific metrics when enforcing different fairness criteria. The legend in (d) applies to all panels. Shaded intervals reflect 95% confidence intervals from variation across repeated simulations.