External Evaluation of Discrimination Mitigation Efforts in Meta's Ad Delivery

Basileal Imana
imana@princeton.edu
Princeton University
Princeton, New Jersey, USA

Zeyu Shen
zs7353@princeton.edu
Princeton University
Princeton, New Jersey, USA

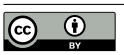
John Heidemann
johnh@isi.edu
USC/Information Sciences Institute
Los Angeles, California, USA

Aleksandra Korolova
korolova@princeton.edu
Princeton University
Princeton, New Jersey, USA

Abstract

The 2022 settlement between Meta and the U.S. Department of Justice to resolve allegations of discriminatory advertising resulted in a first-of-its-kind change to Meta's ad delivery system aimed to guard against algorithmic bias in housing ad delivery. However, the actual reduction in discrimination resulting from the settlement's choice of fairness metrics has not been explored. In this work, we explore direct and indirect effects of both the settlement terms and the resulting Variance Reduction System (VRS) implemented by Meta. We first show that the settlement terms allow for an implementation that does not meaningfully improve access to opportunities for individuals. It measures impact as impressions, instead of unique individuals reached by an ad, it allows the platform to level down access, achieving fairness by decreasing overall access to opportunities, and it allows the platform to selectively apply VRS to only small advertisers. We then conduct experiments to evaluate VRS's implementation with real-world ads, and show that while VRS does reduce variance, it also raises advertiser costs (measured per-individuals-reached), therefore decreasing exposure of opportunity ads to users for a given ad budget. VRS thus *passes the cost of decreasing variance to advertisers*. Finally, we explore an alternative approach to achieve the settlement goals, that is significantly more intuitive and transparent than VRS's implementation. We show our approach outperforms VRS by both increasing ad exposure to users to *all* groups and reducing cost to advertisers. Our methodologies use a black-box approach that relies on capabilities available to any regular advertiser, rather than on privileged access to data, allowing others to reproduce or extend our work.

ACM Reference Format:
Basileal Imana, Zeyu Shen, John Heidemann, and Aleksandra Korolova. 2025. External Evaluation of Discrimination Mitigation Efforts in Meta's Ad Delivery. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3715275.3732170



FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/2025/06
https://doi.org/10.1145/3715275.3732170

1 Introduction

The 2022 agreement between the US Department of Justice (DoJ) and Meta to implement a Variance Reduction System (VRS) has been widely celebrated as a groundbreaking step in regulating ad delivery algorithms in social media platforms to mitigate discrimination in domains of important life opportunities, such as housing and employment [7, 45]. The need for changes in ad delivery algorithms due to their biases was demonstrated by investigative journalism [5, 48], civil-rights audits [29, 44], and academic research [3, 21, 27, 43]. These works highlighted the significant role that machine learning used in platforms' advertising systems can play in perpetuating discriminatory access to economic opportunities. Although the settlement focused exclusively on housing ads, Meta has since voluntarily expanded the deployment of VRS to cover employment and credit ads [7].

Any fairness intervention should consider two questions: First, is the metric of algorithmic fairness effective at assessing a reduction in harm? Second, what are the trade-offs for other objectives, such as utility [8, 20, 39]? The DoJ/Meta settlement suggests minimizing *variance* between the demographic distribution of an ad's actual audience compared to the demographic distribution to the ad's *eligible audience*. For metrics, the settlement establishes a compliance goal: VRS will ensure that the variance of *coverage*, a specified proportion of ads, does not exceed a certain threshold [18]. (we expand on settlement terms, Meta's implementation, and compliance requirements in §2). However, both the chosen metrics and their implementation could result in trade-offs, with the ad reaching fewer recipients at a higher per-recipient cost, perhaps not achieving the best possible reduction in harm [8].

The first contribution of our work is to demonstrate that the settlement terms allow for an implementation that does not improve exposure to opportunity ads for individuals (§3). We first show that the settlement's goal is to balance ad impressions, not ad *reach*, actual number of unique recipients (§3.1). Variance in impressions can be reduced by repeatedly showing an ad to the same users, making the actual societal benefit unclear. Second, the settlement specifics that coverage is measured by unique ads, not user impressions or reach, so ads with, say, 1k and 1M impressions count equally to the target. We show that this metric allows for selective application of VRS to small advertisers (§3.2). By omitting large advertisers, the platform will generate more revenue, but mitigation applies to fewer people, resulting in greater discrimination. We quantify the

potential effect of such selective application on access to opportunities by using public data from Meta on advertising budgets and reach. Finally, we show the settlement requirements allows for a leveling down effect, which risk achieving fairness by decreasing overall access to opportunities (§3.3).

Our second contribution is to show that VRS make ads more expensive to advertisers, therefore effectively reducing how many recipients see opportunities for a fixed ad budget (§4). The result is *leveling down*, where fairness is accomplished by reducing the outcomes of higher-performing groups to down to the level of lower-performing groups, with no demographic group benefiting in the process [35]. We demonstrate this outcome through the first independent evaluation of VRS impact on delivery of real-world ads. We conduct our experiments using a novel black-box methodology that isolates VRS's role by running paired campaigns of the same ad with and without VRS applied. Our findings show that while VRS reduces variance according to the legal compliance metrics for housing ads compared to a case without VRS intervention, fewer unique users are reached by the ad versions to whom VRS is applied and the cost of achieving fairness is passed on to advertisers, making it more expensive for economic opportunity ads to reach a wider audience.

Our final contribution is to explore a new budget-splitting approach that outperforms VRS by increasing exposure of opportunity ads to *all* demographic groups and reducing cost to advertisers (§5). Our findings show VRS's reduced utility for users and advertisers is an artifact of Meta's implementation choices and not an inherent limitation for mitigating discrimination in ad delivery. We explore budget-splitting as an example strategy that addresses the shortcomings of VRS while better meeting fairness and utility goals, and, unlike Meta's VRS implementation, having the benefit of transparency and explainability.

We make data from all our experiments publicly available at [24].

2 Background

We first summarize the terms of the legal settlement between Meta and the DoJ, the VRS implementation to fulfill these terms, and Meta's compliance reporting. We base this overview on the public settlement terms [45], Meta's white-paper and academic publication [34, 47], and documents provided by the DoJ and external reviewer [46].

2.1 Settlement Agreement Between DoJ and Meta

VRS to Reduce Variance in Ad Delivery. A key settlement requirement is that Meta will implement a system to reduce variance by race and gender by aligning the demographics of an ad's *actual audience* with that of an ad's *eligible audience*. This baseline is central to VRS's guarantees; we define it here and analyze its implications for fairness in ad delivery in §3.

The *eligible audience* is the set of all users who fit the targeting criteria chosen by the advertiser and have received one or more impressions of any type of ad on Meta during the last thirty days [45]. The baseline against which the VRS system measures variance is defined via the *eligible ratio*, which relies on the eligible audience. Specifically, the eligible ratio for a specific demographic group g for a particular ad is calculated using the proportion of impressions received by users from the ad's eligible audience who belong to g compared to impressions received by users from the ad's entire eligible audience from any advertiser on Meta in the last 30 days. Mathematically, for a specific *ad* and demographic group g, the eligible ratio is given by: $\textbf{Eligible Ratio}_{g,ad} = (\sum_{u \in ad\_eligible\_aud \cap g} \textbf{Imps}_u)/(\sum_{u \in ad\_eligible\_aud} \textbf{Imps}_u)$, where $\textbf{Imps}_u$ is the number of impressions that the user u received from all advertisers on Meta over the last thirty days [34, 47].

After an ad starts running, VRS measures the demographic distribution of the *actual audience* of the ad, which is the set of all users in the eligible audience to whom at least one impression of this ad is displayed. VRS then calculates a *delivery ratio* for each demographic group g, which is the fraction of total impressions for an ad that were shown to members of g. Mathematically, $\textbf{Delivery Ratio}_{g,ad} = \frac{\textbf{Imps}_{g,ad}}{\textbf{Imps}_{ad}}$, where $\textbf{Imps}_{g,ad}$ is the number of impressions delivered to users in group g, and $\textbf{Imps}_{ad}$ is the total number of impressions for the ad.

Using the Eligible and Delivery Ratios, VRS then aims to reduce the *variance* between the eligible and actual audiences of an ad, aiming for this variance to be below a 10% or 5% threshold. The variance for an *ad* is defined separately for race and gender as follows:

$$\textbf{Variance}(Race)_{ad} = \frac{1}{2} \sum_{g \in \{\text{African American, Hispanic, White, Other}\}} |\textbf{Eligible Ratio}_{g,ad} - \textbf{Delivery Ratio}_{g,ad}|$$

$$\textbf{Variance}(Gender)_{ad} = \frac{1}{2} \sum_{g \in \{\text{Male, Female}\}} |\textbf{Eligible Ratio}_{g,ad} - \textbf{Delivery Ratio}_{g,ad}|$$

where "Other" refers to all races other than African American, Hispanic and White [34, 47]. If an ad impression is delivered to a user with 'unknown' gender, the impression is omitted in calculation of variance for gender [18].

The variance metric, roughly-speaking, represents the minimum fraction of impressions that need to be moved between the groups for the delivery ratio to match the eligible ratio. For example, if the delivery ratio is (0.4, 0.6) for males and females, respectively, and the eligible ratio is (0.5, 0.5), then the variance is $\frac{1}{2}(|0.5 - 0.4| + |0.5 - 0.6|) = 0.1$, indicating VRS needs to move a 0.1 fraction of delivery impressions from females to males to match the eligible ratio.

Users' gender for the computations is based on user self-report on their profiles. Meta infers user race using Bayesian Improved Surname Geocoding (BISG), a public method that gives probabilistic estimates that a person is of a given race based on their surname and zip code [1, 16]. We discuss open questions about how the use of BISG may impact VRS's performance in Appendix B.

Compliance Metrics and Coverage. The extent to which the variances need to be reduced is specified in agreed upon *compliance metrics* [18]. *Coverage* is a key metric, defined as percentage of housing ads among all housing ads run over the compliance reporting period of 4 months whose delivery variance falls below the

thresholds of 5% or 10%. Coverage targets are defined separately for gender and race and for ads that receive more than 300 and more than 1,000 impression. The precise targets agreed upon in [18] are given in Table 1. For example, for housing ads that received at least 1,000 ad impressions, VRS must ensure that the variance by gender is below 10% for 91.7% of ads and the variance by race is below 10% for 81% of the ads.

2.2 Meta's Implementation of VRS

VRS is invoked for any ad the advertiser self-identifies as housing, employment, or credit (HEC) [7, 34]. Meta examines ads on these topics and refuses to run them without HEC tagging. Once an ad tagged as HEC starts running, VRS periodically measures variance and *adjusts the bids* for users on the advertiser's behalf, so as to increase delivery rate to a group currently under-served and/or decrease delivery rate to a group that is over-served.

Specifically, when an ad has a chance to be shown to a user who is using one of Meta's platforms, Meta's ad delivery system runs an ad auction between all ads targeting that user. In this auction, the ads compete based on their *total value*, which is calculated using the advertiser's bid, the *estimated action rate*, which is how likely a user is to take the advertiser's desired action such as clicking on the ad, and *ad quality score*, which estimates overall quality of the ad's content such as its image and text [18]. The total value is given by the following formula:

$$\textbf{Total Value} = \textbf{Advertiser Bid} \times \textbf{Estimated Action Rate} + \textbf{Ad Quality Score}.$$

VRS introduces a new parameter called *VRS multiplier* that modifies the advertiser bid component to change the likelihood of an ad winning the auction [34]. The direction of adjustment is chosen by a machine learning module trained on past data that takes as an input the latest measurement of variance, and produces either an *adjust up* or *adjust down* action that aims to shift the delivery ratios towards the eligible ratios.

The machine learning module is trained in a offline environment to learn which actions have the potential to reduce variance across demographic groups. The module does not receive individual-level demographic information such as the gender or estimated race. Instead, it is trained to take as an input an embedding that summarizes the potential ad viewer along with the latest variance measurement to predict the most likely action that reduces variance for all demographic groups [47].

Lack of Transparency and Clarity on Implementation's Effectiveness. Meta's ad platform generally uses auto-bidding where advertisers specify a budget and the platform bids for each user on the advertiser's behalf; the bids and individual user costs are not reported to the advertiser. The adjustments VRS makes and the resulting changes in costs are also not reported to the advertiser; and are thus completely opaque. In §4.2.3, we show this implementation approach increases costs for advertisers, resulting in lower exposure to economic opportunity ads for users.

2.3 External Verification of Compliance

The settlement requires that a third-party entity serves as an external reviewer to confirm that Meta meets the compliance metrics [18]. The reviewer (currently, Guidehouse) is proposed and paid by Meta, but is subject to consent by the DoJ.

The external reviewer performs its analysis based on aggregate housing data that Meta reports to it every four months, using the data schema shown in Figure 1 (as documented by the reviewer [18]). Specifically, for each housing ad with 300+ impressions, identified using a hashed ad id, Meta includes: *potential impressions*, i.e. the number of impressions each demographic group in the eligible audience[1] received in the last 30 days; *actual impressions* broken down by demographic group; and Meta's estimate of *variance* by gender and estimated race[2].

To verify compliance, the external reviewer simply computes the variance and coverage based on the aggregated data provided by Meta and using the variance formula given in §2.1, and compares it with the variance and coverage metrics reported by Meta. The reviewer also compares the coverage against the thresholds agreed upon in the settlement (Table 1).

Notably, the reviewer has no means to get privileged access to Meta's internal data, and since the reviewer does not run test ads, the reviewer's ability to independently verify accuracy of potential or actual impressions and their breakdown by demographics is limited. Furthermore, the reviewer does not receive any information on costs. Finally, as we discuss in Appendix A, the information reviewer receives is privacy-protected, which further interferes with its accuracy.

While our work also operates without access to internal data, we suggest that our methods using test ads can strengthen an external audit and that such tests are critical to provide a more independent and broader scope verification of compliance that is not limited to verification of coverage computation formulas.

2.4 Related Work

Prior to our work, the only type of external audit VRS has undergone, to our knowledge, is by the reviewer receiving the periodic compliance reports mandated by the settlement [18] (see §2.3). In addition, a 2023 article by a European non-profit, AlgorithmWatch, identified important information gaps in the compliance reports, called into question the scalability of VRS to other domains with risks of bias, and underscored the need for "adversarial audits" that are fully independent of Meta [2]. In contrast, our work provides the first such fully independent and systematic critique of the settlement terms and VRS's design from the perspective of ability to mitigate discrimination and identifies more gaps in information needed for assessing alignment with the fairness goals of the system. We also conduct the first black-box audit of VRS's impact on the delivery of real-world opportunity ads and their cost of delivery, filling some of the gaps and identify areas for improvement.

Separately, Sapiezynski et al. [41] conducted an audit of Meta's Lookalike and Special Ad Audiences tools. Their audit was motivated by sources of bias that remained unaddressed following a 2019 legal settlement between Meta and the National Fair Housing

[1] Meta uses only a sample of the full eligible audience for estimating compliance metrics. The external reviewer reports Meta's system has a target sample size of 6,000 users [19].

[2] In our description of VRS we omit details that relate to measures taken by Meta for privacy reasons, that include adding noise to the mechanism used to measure actual impressions and variance. We expand on these details in Appendix A.

	Coverage for ads that received:	
Variance	≥ 300 Impressions	≥ 1,000 Impressions
Gender (≤ 10%)	90.2%	91.7%
Gender (≤ 5%)	78.3%	84.5%
Estimated race (≤ 10%)	80.1%	81.0%
Estimated race (≤ 5%)	56.8%	61.0%

Table 1: Coverage Requirements for Housing Ads, which vary for ads with 300+ or 1000+ impressions [18].

						Inputs to Calculate Variance													
						Potential Impressions						Actual Impressions							
				Impression Bucket		Sex		Estimated Race/Ethnicity				Sex		Estimated Race/Ethnicity					
#	Hashed Ad ID	Ad Start Date	Ad End Date	300-1000	>1000	Male	Female	White	Hispanic	African American	Other	Male	Female	White	Hispanic	African American	Other	Variance (Sex)	Variance (Estimated Race / Ethnicity)
1																			
2																			
3																			
...																			
n																			

Figure 1: Meta VRS Compliance Metrics Reporting Schema [18]

Alliance. Similarly, our work is motivated by the need for independent audits to confirm stated legal compliance metrics actually lead to better outcome for users.

We next discuss related work on methods for externally auditing discrimination in ad delivery, solutions for algorithmic discrimination in ad delivery and their trade-offs with utility of ads for users and advertisers.

External Auditing Discrimination in Ad Delivery. The methodology we develop for our experiments is inspired by prior audits that rely on paired ads to evaluate discrimination in ad delivery algorithms [3, 21, 23]. Ali and Sapiezynski et al. [3, 4] were the first to develop a paired ads methodology that isolates the role of platforms' algorithms from other confounding factors to show ad delivery outcomes that are skewed by gender and race. Subsequent studies by Imana et al. showed such skewed outcomes maybe in violation of anti-discrimination laws for employment and education domains [21, 23]. The methodology we develop in this work builds on the paired ads methodology to isolate and measure the effect of VRS on delivery of opportunity ads.

Tradeoffs Between Fairness and Utility. Our evaluation of the tradeoff between VRS's fairness guarantees and utility for users and advertisers builds on prior body of work that study similar tradeoffs in both ad delivery and other algorithmic decision making systems [8, 14, 17, 20, 25, 39, 40]. Closest to our study is the work by Baumann et al. [8] that showed through simulations that fairness interventions in ad delivery can lead to a leveling down effect, unless platforms explicitly share the cost of ensuring fairness. Mittelstadt et al. define leveling down [35], a concept we further explore. Hu and Chen similarly show enforcing group fairness metrics may not translate to improved outcomes to previously disadvantaged groups [20]. Pesysakhovich et al. show that fairness constraints in two-sided markets such as ad delivery can discourage advertiser participation, highlighting the trade-offs involved in designing effective fairness interventions [39]. Our research questions are motivated by these known tradeoffs between fairness and utility guarantees of algorithmic decision-making. Our work is the first to study these tradeoffs in VRS's context and test how variance reduction affects utility for advertisers and users.

Balancing the interests of different stakeholders while achieve fairness is also an open research area in the broader field of recommender systems. Deldjoo et al. survey approaches to fairness within recommender systems and identify the need to consider how normative claims underlying chosen fairness metrics apply to particular domains and stakeholders [12]. Similarly, Stray et al. evaluate how different values, including fairness, have been operationalized in the context of recommender systems. They identify open challenges, one of which is balancing trade-offs between different stakeholders [42]. Our findings show that these trade-offs can cause VRS, a system designed to mitigate disparities in ad delivery, to inadvertently lower access to opportunities. This result shows the need for platforms such as Meta to be transparent about how they manage these trade-offs.

Strategies for Mitigating Discrimination in Ad Delivery. A number of approaches have been proposed for mitigation discrimination in targeted advertising systems [10, 14, 37]. Dwork and Ilvento et al. demonstrate that fairness guarantees for individual components of a complex system, such as in ad delivery, do not necessarily translate into fairness for the entire system, and propose methods for combining seemingly unfair components to achieve fair outcomes [14]. Celis et al. proposes imposing fairness constraints on ad auctions to ensure balanced exposure across demographic groups [10]. Another study proposes an alternative approach that modifies the bids set by advertisers to mitigate discrimination without changing the underlying auction mechanism [37]. While we do not propose a concrete and final solution for mitigating discrimination in ad delivery, we explore a budget-splitting approach that can lead towards an alternative to VRS that is more effective at equitable delivery, transparent and explainable.

3 Analysis of Meta/DoJ Settlement Terms

We next identify gaps in the settlement for mitigating discrimination in ad delivery: it focuses on impressions, not individuals, so it may not increase the number of distinct individuals reached by an ad; it requires using a baseline for fairness that factors in prior impressions on the platform, so it is dependent on Meta's possibly biased algorithms and it is not visible to external auditors, making independent audits challenging; its coverage requirement treats all ads above a threshold equally, so it allows for selective application of VRS to small ads impacting few users while continuing the standard algorithm for large ads; and its requirements can be satisfied by leveling down access to opportunities, so no demographic group may benefit from the fairness intervention.

3.1 Focused on Impressions, not Individuals

The first gap is that the settlement's compliance metrics are all defined in terms of impressions rather than individuals reached (§2.1). Ad platforms typically use two types of metrics to evaluate the performance of an ad: *impressions* and *reach*. Impressions represent the number of times a given ad was shown overall whereas reach represents to how many unique user accounts the ad was shown to [32]. Thus, an ad's delivery can meet the variance threshold by showing it repeatedly to the same individuals, providing no increase in how many people see the opportunity. As an example, according to the chosen metrics, an ad shown once to each of 100 unique men and show 100 times to one woman has zero variance, and thus perfect equity, even though many more men are exposed to the opportunity.

Recommendation: We recommend that variance should be measured with respect to reach, i.e. the number of individuals from each demographic group to whom an ad is shown, rather than in terms of impressions received by members of each demographic group.

3.2 Coverage and Selective Application

A third gap is that the coverage requirement allows selective application of VRS to exclude large campaigns. Coverage is defined as the fraction of housing ads for which VRS reduces variance below a certain threshold. Meta has full leeway to choose the subset of ads for whom to meet the variance threshold. In a hypothetical scenario, Meta might choose to meet the variance threshold for ads with small budgets or small audiences while allowing ads with large budgets or large audiences to fall into the (permitted) fraction that does not need to meet the threshold. This decision could result in a much greater impact on the individuals and ad impressions excluded from the settlement constraint than the coverage metric alone might indicate. For example, ads with 1k and 1M impressions count equally to the coverage target, but the latter affects $1,000\times$ more people.

We illustrate the magnitude of spend and impressions that could be excluded from the fairness intervention given this slack in the settlement using the public data Meta provides on political advertising budgets and reach[3]. We obtain a representative sample of 32,867 political ads ran in the US in 2024 using the Meta Ad Library and a methodology developed in prior work [36]. Figure 2 shows the distribution of ad spend and number of impressions for the sample of political ads obtained (excluding ads that received fewer than 300 impressions as settlement terms do not apply to such ads). Both figures suggest a power-law relationship [30, 50], where most of ads spend a small amount and receive relatively few impressions, but a few ads spend a large amount and receive a very large number of impressions.

To quantify the implications of this power-law relationship under selective application of VRS we examine the actual guarantees the compliance metrics provide depending on which ad campaigns are selected and omitted. We compare selective application that excludes large ad campaigns with randomly selecting campaigns for which not to reduce variance below the threshold. For each coverage level, we repeat our random sampling 100 times and report the average. Based on the 81% coverage requirement at 10% variance threshold for race (from Table 1), excluding from variance reduction the largest 19% of the ads excludes 78.9% of the 1.3 billion impressions from the compliance requirement, compared to the 18.9% of impressions that are excluded by randomly selecting which campaigns not to cover. Similarly, exclusion of the largest 39% of ads (based on coverage requirement at 5% threshold for variance by race) would exclude 90.3% of impressions, again far more than the 39.01% that would be excluded using random selection. Excluding the largest 8.3% and 15.5% of ads (based on the coverage requirements for gender corresponding to 10% and 5% variance thresholds) would exclude 57.9% (rather than 8.31%) and 75.6% (rather than 15.51%) of impressions, respectively.

Recommendation 1: We recommend enforcing the coverage requirement within stratified tiers based on audience size and ad spend levels to reduce the uncertainty in how broadly VRS applies. One starting point could be the categories of "small" and "large" advertisers that Meta already uses internally to classify advertisers [49].

Recommendation 2: We recommend for the external reviewer to check for selective application of VRS by conducting an analysis of the distribution of spend and impressions for ads that meet and do not meet the variance reduction thresholds.

3.3 Risk of Leveling Down

Finally, a fourth drawback of the settlement is that it is possible to satisfy the compliance metrics by leveling down access to opportunities. Leveling down is defined as achieving fairness by bringing down the performance for better performing groups down to the level of worse performing groups [35]. Recent prior work has shown through simulations on simplified models that fairness interventions in ad delivery specifically run that risk, unless one explicitly constrains the space of solutions to those where the total number of ad impressions does not decrease [8]. VRS's compliance metrics and the implementation have no such constraint on the number of impressions. In §4.2.3, we run experiments with real ads that show how this limitation can lead to leveling down in practice.

Recommendation: We recommend future efforts to regulate ad delivery algorithms explicitly address the risk of leveling down. One potential approach to address leveling down is to add to VRS's

[3]Our illustration assumes distributions of impressions and spending for ads for economic opportunities follows similar patterns to those of political ads. This assumption is also consistent with data published by Meta [49].

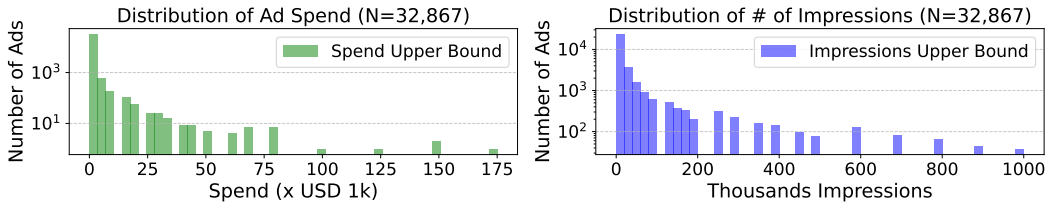


Figure 2: Distribution of ad spend and number of impressions for a sample of political ads from Meta's ad library.

requirements a constraint that ensures the overall number of impressions does not decrease due to variance reduction efforts [8].

4 A More Complete Independent Evaluation of VRS's Performance

We next present the first external validation of VRS's ability to mitigate discrimination in the delivery of real-world ads. Our findings reveal that while VRS reduces variance as measured by its compliance metrics, it uses a baseline that is skewed by race and gender, and increases cost to advertisers, decreasing exposure of opportunity ads to recipients as a result.

4.1 Methodology

Our black-box methodology isolates the effect of VRS by running the same ad twice: once with and once without VRS. We then evaluate the relative performance of the ads by the demographic attributes of interest, such as race and gender. We run both copies with the same targeting parameters, including the ad creative, the budget, and targeted audiences, so that the only difference between the two ad configurations is the presence or absence of VRS application to their delivery. This approach builds on prior work that used paired ads for isolating the role of the ad delivery algorithm from other factors when auditing for discrimination [3, 4, 21, 23], but is the first to use it to evaluate a system built to mitigate discrimination.

4.1.1 Isolating the Effect of VRS. To measure the effect of VRS, we enable or not enable VRS for each of our ad campaigns by declaring them or not declaring them as belonging to the special ad category of housing, the only category for which Meta is legally required to reduce variance. We infer from Meta's and DoJ's public statements that VRS is automatically enabled for any ad an advertiser self-declares as housing [7, 45]. We use non-housing ads in our experiments instead of real housing ads because housing ads are required to be declared as such, and not doing so risks being rejected from running the ad and future ads. To our knowledge, Meta does not enforce restrictions on non-housing ads that are labeled as housing ads. Our declaration of a non-housing ad as a housing ad should not lead to any user harm, as such a declaration should lead towards a more equitable delivery per VRS goals.

We use ads whose delivery may be skewed towards a particular demographic group when VRS is not active to clearly see the effect when VRS is enabled. The three types of ads creatives we use and their rationale are summarized in Table 2. First, we use two non-opportunity ads that are stereotypically associated with a particular demographic group: an ad for a hair product that is stereotypically associated with Black women and an ad for golfing that is stereotypically associated with White men. We expect their delivery to be skewed towards users from those groups when VRS is not enabled. Second, we examine two ads for education opportunities, an economic opportunity that prior research has shown is vulnerable to discriminatory ad delivery [23]. Third, we analyze two ads for insurance and financial products, which are domains not studied in previous research but also have risks of discrimination and are planned to be added to categories of ads for which Meta voluntarily applies VRS beginning in January 2025 [31].

We measure variance by race and gender for both the VRS and no-VRS ads, and test whether enabling VRS reduces variance. From the racial and gender groups in VRS's scope (summarized in §2.1), we measure variance using both gender groups and the two largest racial groups most well-represented in our source dataset for building ad audiences: Black and White. We describe how we build ad audiences in §4.1.3.

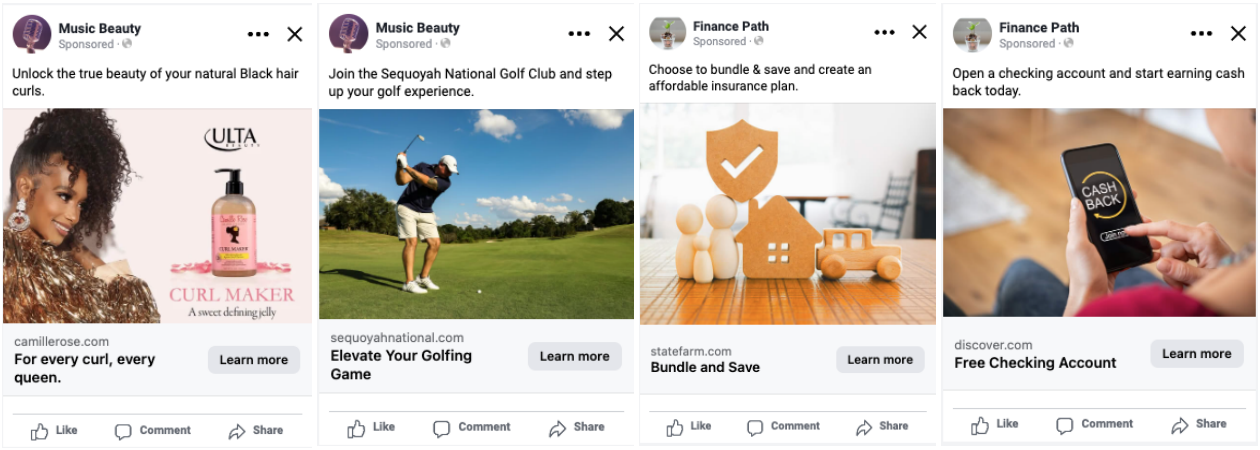
4.1.2 Estimating Eligible Ratio. As defined in §2.1, VRS uses the eligible ratio as a baseline for reducing variance. We estimate this baseline to measure variance using the metric that is agreed upon in the settlement to verify compliance.

We estimate the eligible ratio used by Meta for our ads by taking an average of all delivery ratios we observe for each attribute across all VRS-enabled ads we run. As external auditors, we do not have access to data about past ad impression breakdowns by demographic group, which is the data used by Meta to determine the eligible ratio (as defined in §2.1). However, assuming that VRS works as intended and by the law of large numbers, a reasonable estimate of the eligible ratios VRS aimed for can be obtained by taking an average of all delivery ratios observed across many VRS-enabled ads we run in §4.2. We note that our estimate of eligible ratio does not account for the up to 10% variation that the compliance metric allows between eligible ratio and delivery ratio, and the noise from VRS's reliance on BISG and privacy-protecting measures introduced. We do not see a path to a more precise estimate with only external information.

4.1.3 Building Ad Audience. For our experiments, we use audiences that are demographically balanced by both race and gender. We specify the audiences using Meta's Custom Audience feature that allows us to upload a list of individuals' personal information such as their names and location, which then Meta matches to real user accounts. We select individuals using North Carolina's

Ad ID	Ad creative	Description
HA	Hair product ad	Stereotypically skewed: Black, Women
GA	Golfing ad	Stereotypically skewed: White, Men
EA	Education ad: Arizona State	Evidence of bias from [23]
EB	Education ad: Colorado State	Evidence of bias from [23]
IA	Insurance ad	VRS may apply starting 2025
FA	Financial ad	VRS may apply starting 2025

Table 2: List of categories of ad creatives we use in our experiments. Example screenshots are given in Figure 3



(a) Hair product ad **(b) Golfing ad** **(c) Insurance ad** **(d) Financial ad**

Figure 3: Example screenshots of ad creatives we used in our experiments.

(NC) voter dataset, a public source of location, gender and race of individuals [38].

For experiments where we study variance by race, we build our ad audiences in a way that allows us to infer race from the location of ad recipients, following the approach in prior external audits of ad delivery [3, 21, 23]. Meta reports breakdown of ad recipients by gender, age and District Market Areas (DMAs), but not race. Therefore, an auditor can build audiences using (race, DMA) pairs from NC voter data so that one can infer the racial breakdown of ad impressions based on the DMA breakdown reported by the platform. The list of DMAs in NC split in to two parts, and include only Black individuals from one subset of DMAs and only White individuals from the other subset of DMAs. Depending on which DMA an ad was shown, one can then infer whether it was shown to a White or Black person. We use this same approach in our methodology to determine the racial breakdown of ad impressions. To ensure location does not skew results, prior work replicates all experiments on "flipped" audiences, where the DMAs used to select groups are reversed [3, 23]. We omit this step to due to the cost of doubling our experiments, and because prior work found that flipping audiences produced similar results.

To validate our results with multiple replications, we reproduce each experiment on three audience partitions randomly sampled from the NC voter dataset. We avoid test-retest bias by ensuring the partitions are disjoint. In each audience partition, we include a list of 30k individuals balanced by both race and gender: 7.5k White-male, 7.5k Black-male, 7.5k White-female, and 7.5k Black-female individuals.

4.1.4 Other Campaign Parameters. We run both the VRS-enabled and no-VRS copies of an ad for a 24-hour period with a total budget of $20 per ad. We find these audience sizes, campaign durations and budgets are large enough to generate sufficient ad impressions for our analyses. In our experiments in §4.2, we observe that VRS begins to take effect after a few hours of delivery. Meta's settlement requires VRS to reduce variance for all ads that receive at least 300 impressions (see §2.1), and our ads generate approximately 1,500 impressions on average (all get at least 1,100 impression)—well above this threshold.

To avoid the two ads with the same visual elements competing for the same set of users, we run the VRS-enabled and no-VRS ads on separate audience partitions. This step differs from our prior work where we ran ads for different opportunities concurrently [21, 23], so we could explicitly see which ad the auction algorithm favors. Here we run identical ads, varying only the status of VRS, so we run on different audiences to avoid self-competition. In order for ads not to waste recipients' time, each ad links to real websites where users can learn more about the product or opportunity we advertise. We use "Traffic" objective for all ads, which optimizes delivery for increasing traffic (i.e. clicks) for websites our ads link to.

4.2 Experiments

We next apply our methodology to real-world ads to support our two key claims: VRS does reduce variance with respect to the metrics agreed upon in the settlement (§4.2.1); and the impact of enabling VRS is higher per-ad costs for advertisers and therefore fewer opportunity ads shown to recipients for a given budget (§4.2.3).

Each experiment consists of a pair of a VRS-enabled and a no-VRS ad that are otherwise identical. To evaluate variance by both race and gender, we run two separate experiments for each case: one using an audience demographically balanced by gender and one balanced by race. To confirm the results are repeatable, we replicate each experiment on three difference audiences. Using the six ad creatives summarized in Table 2, two demographic attributes, and three replications, we run a total of $N = 36$ experiments of paired no-VRS ads and VRS-enable ads (achieved by self-declaring those ads as belonging to the special ad interest category of housing).

4.2.1 Does VRS Reduce Variance with Respect to Eligible Ratio? We next confirm that VRS does reduce variance with respect to our estimate of the eligible ratios, i.e. with respect to the compliance metrics agreed upon in the settlement for delivery of housing ads.

We first estimate what eligible ratio VRS uses using our methodology in §4.1.2. Figure 4 shows the delivery ratios across all ads by race and gender, indicated in each figure by $\bar{x}$ and the vertical dotted lines. From this result, we estimate VRS uses 0.42, 0.58, 0.45, and 0.55 as eligible ratios for Black, White, male and female groups, respectively. In Appendix C, we show the non-balanced ratio even though we target a 50:50 audience can be explained by differences in matching rates across groups in the Custom Audiences we use.

Using our estimate of the eligible ratio, Figure 5 compares the variance of all $N = 36$ pairs of VRS and no-VRS ads. The orange dots (vrs-housing) and the blue cross-marks (no-vrs) show the level of variance when VRS is enabled and disabled, respectively. The green arrows indicate that, compared to no-VRS, enabling VRS reduces variance. A red arrow to the right indicates VRS increased the variance instead.

In the left figure for gender, variance is less than 5% even without VRS in 15 out of 18 cases, but enabling VRS further reduces the variance in some cases. In the right figure for race, we see variance is more than 10% without VRS in all 18 cases. VRS effectively reduces variance to less than 10% in all cases, bringing it down to less than 5% in 15 of the 18 cases. This result is the first in-the-wild verification that VRS reduces variance with respect to the compliance metrics specified in the settlement.

4.2.2 Evaluating VRS's Performance on Employment and Credit Ads. We next show the reduced variance for ads declared as housing that we saw in the previous section do not extend to employment and credit, two special ad category domains to which Meta has voluntarily expanded its application of VRS. Meta is not legally required to meet the same variance threshold and coverage requirements that apply to housing, potentially allowing it to use a higher variance threshold or a lower coverage threshold.

For this evaluation, we run additional experiments where we enable VRS by declaring an ad as a credit and an employment ad, instead of housing. We then compare the outcome with the no-vrs and vrs-housing ads we ran in §4.2.1 where we enabled VRS by declaring an ad as a housing ad. We conduct this evaluation using the hair product ad which we expect to skew towards Black women when VRS is not enabled. We reproduce all ad campaigns on threes different audiences for reproducibility.

Figure 6a shows the outcome of VRS for all three domains: housing, credit and employment. In the left figure, we compare the fraction of Black users each type of VRS-enabled ad was shown to. The horizontal dotted line indicates 50% delivery to Black users. We see that an ad declared as housing is delivered to approximately 43-44% Black users in all three repetitions, consistent with the overall trend of delivery ratio we observed VRS achieves for race in §4.2.1. However, the ads declared as credit and employment are delivered to 55-60% Black users, an outcome that is closer to the 60-65% Black outcome we see for the no-vrs case.

We further illustrate the outcome in Figure 6b, where we compare the variance for all the ads using as a baseline the eligible ratio we estimated in §4.2.1. The horizontal dotted line represents the maximum 10% variance threshold that is allowed for housing ads by the settlement. This figure evidently shows VRS reduces variance to below 10% for the ad declared as housing, but not for the ads declared as credit or employment. The variance for those two categories remains above 10% and is comparable to the delivery outcome without VRS, suggesting the variance or coverage metrics Meta applies for those categories of ads are more lenient than those agreed upon in the settlement for housing.

These findings indicate that the voluntary expansion of VRS to employment and credit domains, while a commendable initiative, does not match the expectation that it performs in the same way as in the housing domain. More broadly, given the current settlement and compliance metrics are narrowly focused on housing, our result suggests more transparency is needed on what variance and coverage metrics are used when deploying VRS in new domains.

4.2.3 VRS Reduces Utility for Users and Advertisers. We next evaluate how VRS affects exposure to opportunity ads for users and costs for advertisers. We measure utility for users in terms of the number of unique people an opportunity ad is shown to across different demographic groups (i.e. in terms of the ad's reach). We also test whether VRS results in the leveling-down effect (hypothesized in §3.3), where lower variance is achieved by decreasing exposure to the ad for the advantaged group without benefiting the previously disadvantaged group. We measure utility for advertisers in terms of cost per reaching 1,000 unique ad recipients of a certain demographic (CPP or "Cost per Point"), which captures how advertisers are affected by VRS. We find that the cost of VRS is passed on to advertisers and the VRS implementation does not necessarily lead to a greater exposure of opportunities to users.

The scatterplots in Figure 7a comparing the reach and CPP for the same ad run with and without VRS enabled reflect the results of 36 experiments. Each point corresponds to a paired ad experiment, with its reach (resp. CPP) for the no-VRS version on the x-axis, and reach (resp. CPP) for the VRS version on the y-axis. In the top figure, the majority of the points in the scatterplot lie below the diagonal line ($y = x$), indicating that fewer people receive the ad when VRS is enabled. Across all the paired ads, enabling VRS reduces mean reach by 9.82%. In the bottom figure, the majority of the points lie above the diagonal line, indicating that the cost increases when

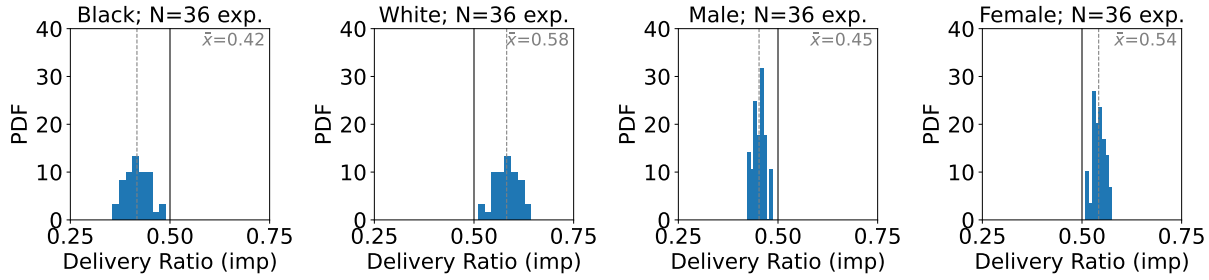


Figure 4: Delivery ratio by race and gender for all VRS-enabled ads. We use the mean delivery ratio for each demographic group (shown by $\bar{x}$ and vertical dotted line in the figures) as an estimate for what eligible ratio VRS uses to reduce variance.

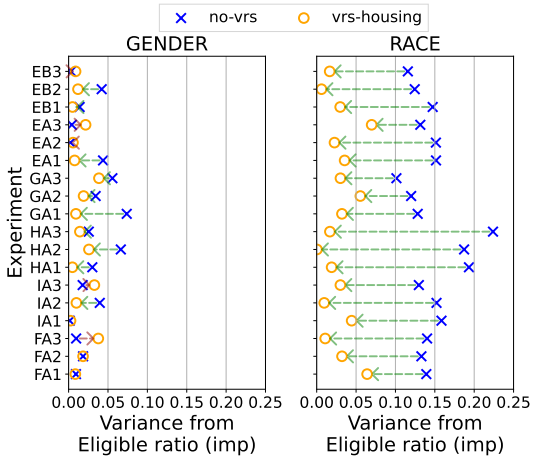


Figure 5: Comparison of variance with and without VRS using impressions and eligible ratio as a baseline. VRS generally reduces variance compared to the no-VRS case for both race and gender.

VRS is enabled. Enabling VRS increases CPP by 12.02% on average across all experiments. Taken together, these results demonstrate that enabling VRS results in fewer exposures to opportunities for individuals, with each ad recipient costing advertisers more.

We next consider how the decrease in exposure and increase in cost is distributed among different demographic groups. For this evaluation, we look at the hair-product ad whose delivery we expect to be skewed towards Black women when VRS is not enabled across 6 paired experiments on disjoint audiences. The bar chart in Figure 7b shows the effect of enabling VRS on the number of users reached by the ad from each demographic group, with the top figure giving the break down by race, and the bottom one by gender. As expected, VRS reduces the number of Black users and the number of women who see the ad, compared to its no-VRS option. We do see the leveling-down effect in 2/6 cases (HA2-r and HA2-g), where both groups see a decrease in exposure. In another case (HA3-r), one group experiences a substantial reduction in ad exposure while the other does not receive a corresponding increase.

Figure 7c presents the results of the same 6 paired experiments analyzed from the perspective of the cost to reach users from each group. CPP consistently and significantly rises with enabling VRS (except in HA1-r), showing Meta pushes the cost of achieving fairness to advertisers Overall, these results demonstrate that VRS achieves lower variance by making impressions scarcer and, therefore, more expensive for advertisers.

Recommendation: We suggest Meta should explore alternative strategies to ensure the cost of achieving fairness is not fully passed on to users and advertisers. One way is to add a constraint on the number of impressions the VRS should achieve – non-decreasing compared to the non-VRS version, as discussed in §3.3. Another strategy is for the platform to offer subsidies or discounts for advertisers' bids on VRS-enabled ads to offset advertiser higher cost per ad recipient or to otherwise modify the Total Value computation or its use in price-setting. Meta can introduce cost-sharing models between the platform and advertisers to distribute the cost of fairness more equitably.

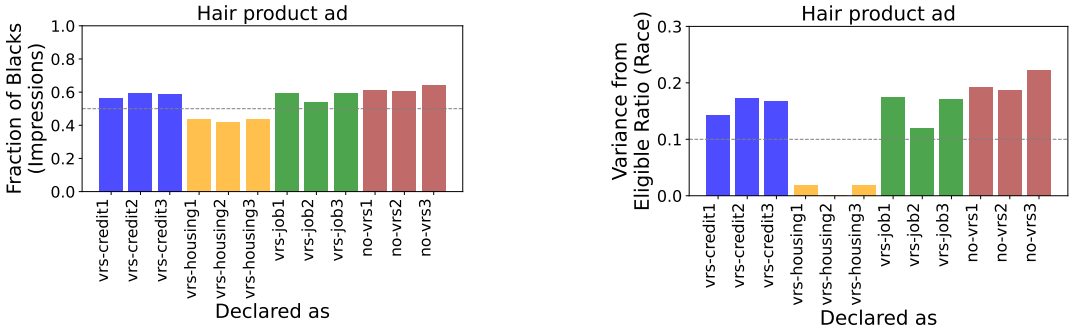
5 VRS is a Suboptimal Implementation of Settlement

We next demonstrate reduced utility for users and advertisers shown in the previous section is an artifact of the specific implementation for VRS chosen by Meta, and not an inherent consequence of the settlement goals. We show this by experimentally comparing the outcomes of VRS with a simple alternative approach aimed at reaching a demographically balanced audience: splitting an ad's total budget equally among all targeted demographic groups, and then running separate ad campaigns for each group. Based on the experiments in §5.2, we show this alternative approach outperforms VRS by increasing exposure to opportunities for all groups and reducing cost to advertisers, compared to the VRS-enabled run. We caution that splitting the budget equally does not necessarily guarantee equal impressions; however, since even this simple approach outperforms VRS, we conclude that VRS is suboptimal.

5.1 Methodology

We explore an approach where we manually split a campaign budget evenly among demographic groups instead of relying on VRS to automatically and implicitly determine how much is spent on each group.

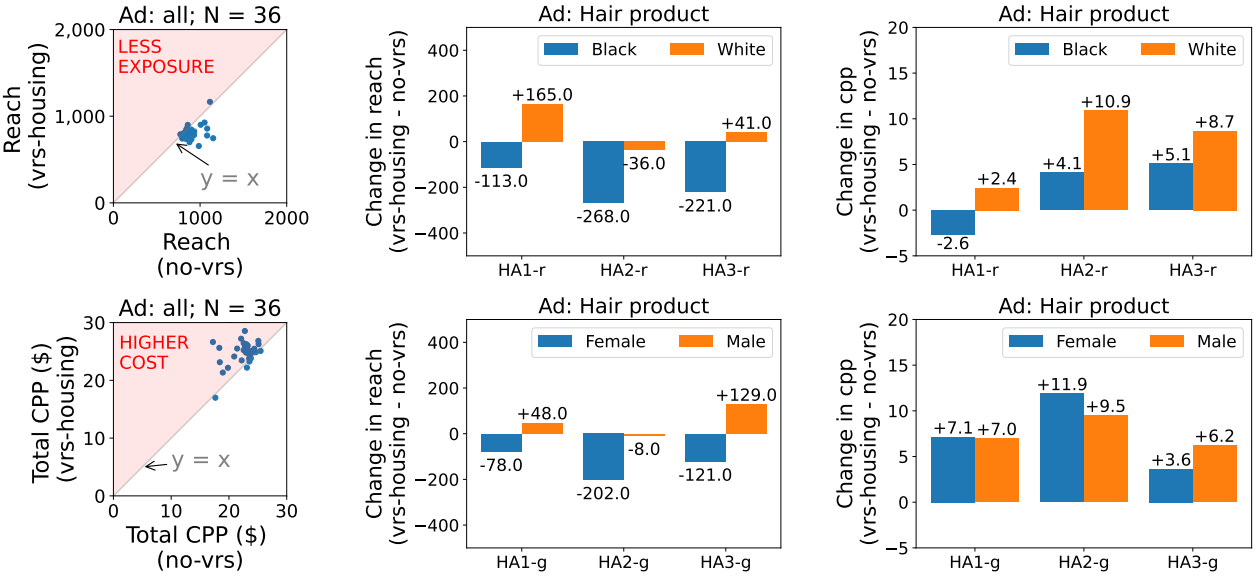
We run separate ad campaigns for each demographic group with an equal share of the total budget. For our experiments, we focus on both gender groups and the two largest racial groups



(a) Demographic breakdown of ad impressions

(b) Variance from estimated eligible ratio for race

Figure 6: Comparison of VRS's performance on housing ads, the only domain within the scope of the settlement, with employment and credit ads, two domains for which Meta voluntarily deployed VRS.



(a) Total reach and CPP

(b) Change in reach per group

(c) Change in CPP per group

Figure 7: Comparison of reach metric and cost per 1,000 reach (CPP) with and without VRS.

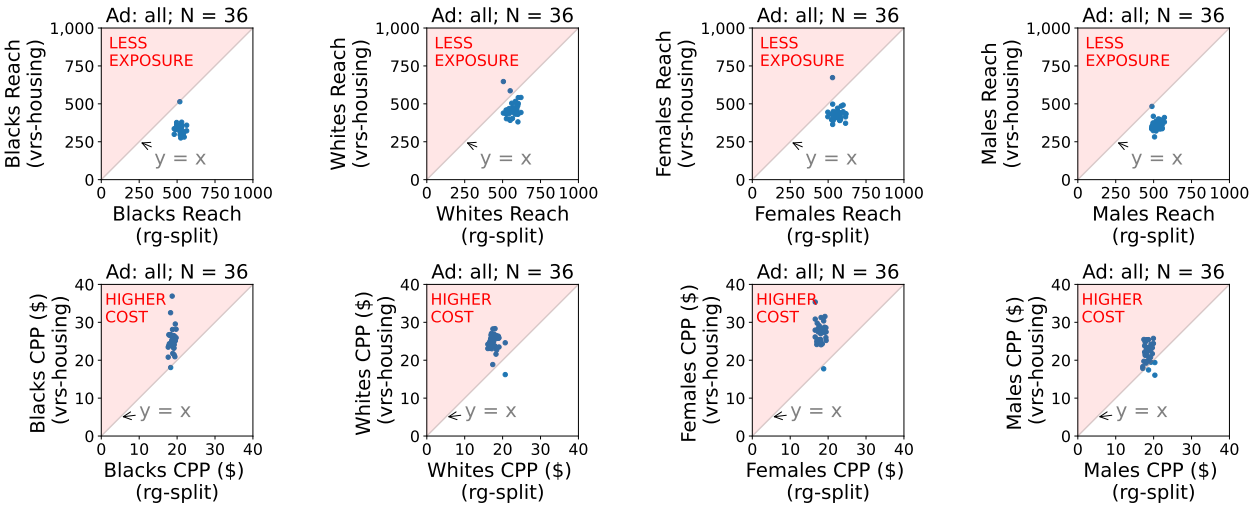
most well-represented in our audience dataset: Black and White. Consequently, we split the total budget among the following four subgroups: White males, White females, Black males, Black females. We then run four separate ad campaigns targeting each subgroup with a quarter of the total budget. We set the total budget to $20, so each subgroup receives a budget of $5. We additionally run a single VRS-enabled campaign with the full $20 budget, targeting the combined audiences. We aggregate the results across the two split campaigns and compare the outcome with that of the VRS-enabled campaign along the following metrics: variance, reach and cost for advertisers.

Other parts of the methodology, such as the ad creatives, the audience sources and campaign parameters follow our first methodology described in §4.1.

5.2 Experimental Results Comparing VRS and Budget-Splitting

Using the six different ad creatives in Table 2, two demographic attributes (race and gender), and replication on three different audience partitions, we run a total of $N = 36$ experiments comparing the outcome of ad delivery with VRS and with budget-splitting.

Across all demographic groups we consider, the budget-splitting approach outperforms VRS by increasing reach for all groups. Top row of Figure 8 illustrates this result, by presenting the scatterplots for the reach achieved by each of the 36 pairs of ads when using the budget-splitting strategy (x-axis) vs. the VRS strategy (y-axis). For Black, White, Female and Male users, the vast majority of the points are below the $y = x$ line, demonstrating that the budget-splitting method gives more exposure to that demographic group



(a) Black **(b) White** **(c) Female** **(d) Male**

Figure 8: Comparison of total number of reach and cost per 1,000 reach (CPP) per demographic group. The top four figures show, compared to VRS, splitting budget by race and gender increases the number of impressions for all groups. The bottom four figures mirror this result by showing VRS as a result has higher cost per impression than budget-splitting.

than the VRS method, while keeping the budget constant. The increase in reach across the board shows budget-splitting achieves a categorically better outcome than VRS in terms of increasing access for the economic opportunities advertised.

Given we use the same total budget for all campaigns, the reduced reach of VRS compared to budget-splitting implies VRS also has a larger cost per person reached for advertisers than budget-splitting. Bottom row of Figure 8 illustrates this result, where VRS results in higher CPP as shown by the points above the y = x line (shaded red).

Taken together, these results show that simply splitting a campaign's budget across demographic groups would be both more effective in terms of people reached and in terms of cost for mitigating bias in ad delivery than VRS.

6 Open Questions for Non-Discrimination in Ad delivery

Our findings raise several questions that should guide future discussion on fairness interventions in ad delivery: considering alternatives to VRS's approach that provide better trade-offs between fairness, utility and transparency, giving external reviewers more capabilities to conduct effective audits, and giving users more control over what high-stakes ads they are shown.

Are there better trade-offs between fairness, utility and transparency? The first is exploring approaches that provide better trade-offs between fairness and utility while also being transparent. Although we do not propose budget-splitting (from §5) as a concrete and final solution for ensuring fairness in ad delivery, it is a useful alternative to consider when comparing with the VRS implementation chosen by Meta. First, it demonstrates that the increased costs incurred by the advertisers and the decreased reach experienced by users under VRS are not an inherent limitation of all fairness interventions, and more efficient strategies than VRS should be explored. Recent studies explore the concept of "Less Discriminatory Algorithms" (LDAs) that can achieve the same business needs as standard algorithms while reducing disparate impact [9, 28]. Laufer et al. show that, even in complex systems, effective alternative policies can be discovered efficiently[28]. These studies supports the possibility of discovering practical alternatives that avoid the shortcomings of VRS.

Second, it demonstrates that a much more transparent and interpretable fairness intervention, that performs no worse than VRS, is feasible. The advertisers themselves may not be able to split the budget according to demographic groups, for example, because they may lack access to the demographic composition of their audiences and because splitting a budget between the intersection of numerous demographic attributes can be a challenge. However, Meta could employ attribute inference in the same way they do for VRS, and thus realize this approach on advertisers' behalf. Moreover, Meta already offers advertisers numerous tools where it allocates the budget on their behalf, through tools like campaign budget optimization [11] and Advantage+ [33]. The tools all involve Meta in automatically configuring advertiser campaigns and budgets in ways that leverage the platform's data and optimization capabilities. Thus, extending such efforts to fairness appears to be a path that does not require substantial engineering effort and would fit well into paradigms and tools advertisers are familiar with.

Finally, in scenarios where the advertisers may have better information about their audiences than the advertising platform, the budget-splitting approach may do away with the need for platform's inference. In some scenarios that may mean that non-discrimination could be ensured also along protected characteristics where inference methods do not exist or are not particularly accurate.

We hope our experimental results based on the budget-splitting approach (§5.2) are convincing in motivating the platforms to look for better, cheaper, more transparent and interpretable implementations for the desired fairness outcomes, and for regulators to push for them.

What capabilities should external reviewers have? Our work also raises questions about what capabilities and level of access external reviewers should be given to provide sufficient oversight over platforms. As discussed in §2.3, the current external reviewer mandated by the settlement only has access to aggregate ad delivery reports provided by Meta and does not have access to internal data and experimentation tools that would allow for independent verification of compliance. The limited access raises concerns about whether reviewers can robustly detect noncompliance or evaluate the broader impacts of algorithmic interventions like VRS. Future regulatory efforts need to consider these limitations and explore what additional levels of access are necessary for effective oversight. Potential approaches include giving auditors privacy-preserving access to internal data such as the output of personalization algorithms that platforms use to calculate the total value of ads used to determine winners of ad auctions [22], and providing means for auditors to perform socio-technical audits that study the impact of algorithmic interventions from the perspective of users [26].

What controls to offer users for high-stake ads? Finally, beyond addressing limitations of VRS or introducing alternative fairness interventions, a more fundamental question is whether platforms should give users greater control over how they receive high-stakes ads. Currently, platforms opaquely optimize ad delivery for "relevance" to users, but this approach has been shown to lead to discriminatory delivery of ads in high-stakes domains such as housing, employment and education [3, 21, 23]. One possible solution is to allow users to turn off relevance optimization for high-stake ads that offer economic opportunities. By selectively turning off relevance optimization, platforms can reduce bias in the delivery of opportunity ads while continuing to use their optimization algorithms in the delivery of other ads such as entertainment and product ads.

7 Conclusion

In this work, we evaluate the settlement and Meta's VRS implementation from the perspective of their ability to mitigate discrimination. We identify critical gaps in the settlement requirements that allow for an implementation that does not improve access to opportunities for individuals. We show that while VRS's implementation reduces variance as required by the settlement terms, it leads to fewer unique individuals being exposed to opportunity ads and increased costs for advertisers. We demonstrate that alternative strategies, such as budget-splitting, can achieve better outcomes, illustrating the sub-optimality of Meta's chosen approach and offering clues as to the possibility of improvement. We propose potential areas for improvement in the settlement terms and VRS's effectiveness, such as incorporating reach-, rather than impression-, focused metrics, having Meta share the cost of fairness intervention, and publishing the baseline used in the eligible ratios. Our work contributes to the overarching goal of increasing transparency, enabling independent evaluations of platform's efforts towards mitigating discrimination and opening up directions for future work.

Acknowledgments

This work was funded in part by the National Science Foundation grants CNS-1956435, CNS-2344925, and CNS-2319409, and by the Alfred P. Sloan Research Fellowship for A. Korolova.

References

[1] Rachad Alao, Miranda Bogen, Jingang Miao, Ilya Mironov, and Jonathan Tannen. 2021. How Meta is working to assess fairness in relation to race in the U.S. across its products and systems. Technical Report https://ai.facebook.com/research/publications/how-meta-is-working-to-assess-fairness-in-relation-to-race-in-the-us-across-its-products-and-systems.

[2] John Albert. 2023. Not a solution: Meta's new AI system to contain discriminatory ads. https://algorithmwatch.org/en/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. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, Vol. 3. ACM, 1–30. https://dl.acm.org/doi/10.1145/3359301

[4] Muhammad Ali, Piotr Sapiezynski, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2021. Ad Delivery Algorithms: The Hidden Arbiters of Political Messaging. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 13–21. doi:10.1145/3437963.3441801

[5] Julia Angwin and Terry Paris Jr. 2016. Facebook Lets Advertisers Exclude Users by Race – ProPublica. Retrieved February 18, 2021 from https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race

[6] Brian Asquith, Brad Hershbein, Tracy Kugler, Shane Reed, Steven Ruggles, Jonathan Schroeder, Steve Yesiltepe, and David Van Riper. 2022. Assessing the Impact of Differential Privacy on Measures of Population and Racial Residential Segregation. *Harvard Data Science Review* Special Issue 2 (jun 24 2022). https://hdsr.mitpress.mit.edu/pub/1rsg867y.

[7] Roy L. Austin. 2023. An Update on Our Ads Fairness Efforts. https://about.fb.com/news/2023/01/an-update-on-our-ads-fairness-efforts/.

[8] Joachim Baumann, Piotr Sapiezynski, Christoph Heitz, and Aniko Hannak. 2024. Fairness in Online Ad Delivery. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1418–1432. doi:10.1145/3630106.3658980

[9] Emily Black, John Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. 2024. Less Discriminatory Algorithms. *Georgetown Law Journal* 113, 1 (2024). doi:10.2139/ssrn.4590481 Washington University in St. Louis Legal Studies Research Paper Forthcoming.

[10] L. Elisa Celis, Anay Mehrotra, and Nisheeth K. Vishnoi. 2019. Toward Controlling Discrimination in Online Ad Auctions. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:162168628

[11] Meta Business Help Center. [n. d.]. About Advantage campaign budget. https://www.facebook.com/business/help/153514848493595?id=629338044106215

[12] Yashar Deldjoo, Dietmar Jannach, Alejandro Bellogín, Alessandro Difonzo, and Dario Zanzonelli. 2023. Fairness in Recommender Systems: Research Landscape and Future Directions. *User Modeling and User-Adapted Interaction* 33 (2023), 1–50. doi:10.1007/s11257-023-09364-z

[13] Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II* (Venice, Italy) *(ICALP'06)*. Springer-Verlag, Berlin, Heidelberg, 1–12. doi:10.1007/11787006_1

[14] Cynthia Dwork and Christina Ilvento. 2019. Fairness Under Composition. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPICS.ITCS.2019.33

[15] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. 2017. Exposed! A Survey of Attacks on Private Data. *Annual Review of Statistics and Its Application* 4, Volume 4, 2017 (2017), 61–84. doi:10.1146/annurev-statistics-060116-054123

[16] Marc N. Elliott, Peter A. Morrison, Allen Fremont, Daniel F. McCaffrey, Philip Pantoja, and Nicole Lurie. 2009. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. In *Health Services and Outcomes Research Methodology*. 69–83. https://doi.org/10.1007/s10742-009-0047-1

[17] 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* (Atlanta, GA, USA)

(FAT '19)*. Association for Computing Machinery, New York, NY, USA, 329–338. doi:10.1145/3287560.3287589

[18] Guidehouse. 2024. VRS Compliance Metrics Verification. https://www.justice.gov/crt/media/1362086/dl.

[19] Guidehouse. 2024. VRS Compliance Metrics Verification. https://www.justice.gov/crt/media/1385866/dl.

[20] Lily Hu and Yiling Chen. 2020. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 535–545. doi:10.1145/3351095.3372857

[21] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2021. Auditing for Discrimination in Algorithms Delivering Job Ads. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3767–3778. doi:10.1145/3442381.3450077

[22] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2023. Having your Privacy Cake and Eating it Too: Platform-supported Auditing of Social Media Algorithms for Public Interest. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 134 (April 2023), 33 pages. doi:10.1145/3579610

[23] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2024. Auditing for Racial Discrimination in the Delivery of Education Ads. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)* (Rio de Janeiro, Brazil). Association for Computing Machinery, New York, NY, USA, 14. doi:10.1145/3630106.3659041

[24] Basileal Imana, Zeyu Shen, John Heidemann, and Aleksandra Korolova. 2025. Dataset for experiments in "External Evaluation of Discrimination Mitigation Efforts in Meta's Ad Delivery". https://ant.isi.edu/datasets/addelivery-vrs/.

[25] Michael P. Kim, Aleksandra Korolova, Guy N. Rothblum, and Gal Yona. 2020. Preference-informed fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 546. doi:10.1145/3351095.3373155

[26] Michelle S. Lam, Ayush Pandit, Colin H. Kalicki, Rachit Gupta, Poonam Sahoo, and Danaë Metaxa. 2023. Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 360 (Oct. 2023), 37 pages. doi:10.1145/3610209

[27] 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 (2019), 2966–2981.

[28] Benjamin Laufer, Manish Raghavan, and Solon Barocas. 2025. What Constitutes a Less Discriminatory Algorithm?. In *Proceedings of the 2025 Symposium on Computer Science and Law* (Munich, Germany) *(CSLAW '25)*. Association for Computing Machinery, New York, NY, USA, 136–151. doi:10.1145/3709025.3712214

[29] Laura Murphy and Associates. 2020. Facebook's Civil Rights Audit – Final Report. https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf.

[30] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. https://nlp.stanford.edu/IR-book/

[31] Meta. 2024. Upcoming expansion of special ad categories to include financial products and services in early 2025. Retrieved December 11, 2024 from https://www.facebook.com/business/help/510724041294968

[32] Meta. 2025. Reach. Retrieved January 11, 2025 from https://www.facebook.com/business/help/710746785663278?helpref=faq_content

[33] Meta Business Help Center. [n. d.]. About Meta Advantage. https://www.facebook.com/business/help/733979527611858

[34] Meta Technical Report. 2023. Towards Fairness in Personalized Ads. Retrieved November 24, 2023 from https://about.fb.com/wp-content/uploads/2023/01/Toward_fairness_in_personalized_ads.pdf

[35] Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default. *Michigan Technology Law Review* (2023). https://arxiv.org/abs/2302.02404

[36] Varun Nagaraj Rao and Aleksandra Korolova. 2023. Discrimination through Image Selection by Job Advertisers on Facebook. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT 2023)*. 1772–1788.

[37] 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* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 337–347. doi:10.1145/3351095.3375783

[38] North Carolina State Board of Elections. [n. d.]. Voter History Data. https://dl.ncsbe.gov/index.html. Downloaded on March 12, 2024.

[39] Alexander Peysakhovich, Christian Kroer, and Nicolas Usunier. 2023. Implementing Fairness Constraints in Markets Using Taxes and Subsidies. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (2023). https://api.semanticscholar.org/CorpusID:257496315

[40] 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 (Oct. 2021), 896–904. doi:10.1038/s42256-021-00396-x

[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 *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 609–616. doi:10.1145/3514094.3534135

[42] Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, Sanne Vrijenhoek, Amy Zhang, McKane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, and Nina Vasan. 2024. Building Human Values into Recommender Systems: An Interdisciplinary Synthesis. *ACM Trans. Recomm. Syst.* 2, 3, Article 20 (June 2024), 57 pages. doi:10.1145/3632297

[43] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery: Google Ads, Black Names and White Names, Racial Discrimination, and Click Advertising. *Queue* 11, 3 (March 2013), 10–29. doi:10.1145/2460276.2460278

[44] Facebook's Civil Rights Team. 2021. Meta's Progress on Civil Rights Audit Commitments. https://about.fb.com/wp-content/uploads/2021/11/Metas-Progress-on-Civil-Rights-Audit-Commitments.pdf.

[45] The US Department of Justice. 2022. Justice Department Secures Groundbreaking Settlement Agreement with Meta Platforms, Formerly Known as Facebook, to Resolve Allegations of Discriminatory Advertising. Retrieved June 29, 2022 from https://www.justice.gov/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known

[46] The US Department of Justice. 2024. United States v. Meta Platforms, Inc., f/k/a Facebook, Inc. (S.D.N.Y.). Retrieved Jan 6, 2025 from https://www.justice.gov/crt/case/united-states-v-meta-platforms-inc-fka-facebook-inc-sdny

[47] 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)*. ACM, 4937–4947. doi:10.1145/3580305.3599916

[48] Ariana Tobin and Jeremy B. Merrill. 2018. Facebook Is Letting Job Advertisers Target Only Men – ProPublica. Retrieved February 18, 2021 from https://www.propublica.org/article/facebook-is-letting-job-advertisers-target-only-men

[49] Nils Wernerfelt, Anna Tuchman, Bradley Shapiro, and Robert Moakler. 2022. Estimating the Value of Offsite Data to Advertisers on Meta. *University of Chicago, Becker Friedman Institute for Economics Working Paper* 114 (2022).

[50] Wikipedia. [n. d.]. Power law. https://en.wikipedia.org/wiki/Power_law

A Privacy-preservation in VRS and its implications

In §2, we omitted details regarding steps that VRS's implementation takes when measuring variance to protect the privacy of its users. We now detail these steps and their implications for VRS's ability to mitigate discrimination and for auditors' ability to verify Meta's compliance with established metrics. Our key insights are that the privacy threats that necessitate taking the privacy-preserving steps are poorly articulated or justified, while the privacy-preserving steps such as noise addition make meaningful verification of compliance impossible. Our key insights that privacy may be used as a smoke-screen for partial compliance or non-compliance.

A.1 Use of Differential Privacy in VRS

Meta applies DP in two parts of VRS's implementation. Informally, DP is a technique that ensures individual's data privacy by adding carefully calibrated noise to statistical outputs computed on a dataset, making it harder to determine whether any single individual's data is included in the dataset [13]. First, Meta adds noise to the counts of impressions from each demographic group during each episode (i.e. after delivery of k impressions) of variance measurement [34, 47]. For each episode, after counting the number of impressions from each demographic group, noise is added to the counts before being passed onto the VRS module that calculates the variance and the controller that determines how to adjust the bids. The value of k is not publicly disclosed.

The second place Meta aims to ensure DP is in its BISG-based race classification tool [1]. To prevent access to individual-level BISG classification of race, this tool returns aggregate number of impressions at the group level for both eligible and actual audiences. The aggregated statistics for each racial group is made available to VRS's controller after applying noise using DP [34]. The details of how the DP mechanism is implemented are not publicly documented.

A.2 Lack of Clarity of Threat Model for Privacy

We next discuss what is known about the privacy threat model Meta considers when implementing VRS, and argue for why this threat model may be one that DP does not protect against. As summarized in §A.1, DP is used when applying BISG to infer race and when measuring variance in ad delivery. Meta's documentation mentions their motivation for using DP with BISG is to "prevent reidentification" of individuals' estimated race attributes [1]. Meta also vaguely states they use DP for variance measurement to "address various common issues such as privacy attacks discussed in [15]" [47].

While it is unclear what specific angles of attack Meta considered, we infer the likely threat model to be one that DP does not protect against. The documentation mentions noise is added to variance measurement to "to prevent the system learning and subsequently acting on individual-level demographic information" [34]. We gather this kind of attack to be analogous to the "smoking causes cancer" problem. This problem highlights how DP protects the privacy of an individual within a dataset, but that it does not address the underlying cause-and-effect relationship between a sensitive attribute and a measurements findings that would have been reached regardless of whether the individual was in the dataset or not. For example, consider an individual of a specific demographic group that receives an ad impression in episode t, after which VRS measure variance with DP and applies a bid adjustment for the ad. Then, say, a second individual from the same demographic group is browsing the platform in episode $t+1$ and has available ad slot. The individual's membership in that group will affect the bid for the ad slot regardless of whether the individual received impressions in previous episodes and whether computations were done with DP or not. This example demonstrates DP does not protect against this outcome because the individual's group membership affects the action VRS takes.

Beyond the VRS system itself, other potential threat actors are internal employees, external auditing or advertisers. Meta states that "human analysts reviewing VRS or its outputs" are potential actors that might violate the privacy of individuals [34]. If the intended adversaries include the advertisers or the external reviewer mandated by the settlement, they only observe the final aggregated delivery metrics, not the intermediate computations. In that case, adding noise just once at the end of the process would suffice, instead of every k impressions. If the intended adversaries are internal employees, who are granted deeper access to the system's inner workings, contractual and administrative controls could preclude their misuse of sensitive data without the need to continuously add noise.

A.3 Possible Harms of Using DP for Fairness and Auditability

Beyond the lack of clarity in VRS's threat model for privacy, the use of DP also introduces trade-offs with fairness and auditability. We next discuss how VRS balances privacy protections with its stated fairness and transparency objectives.

While DP can provide a rigorous privacy protection, it also adds noise that impact fairness evaluations. For example, prior work on DP in the U.S. Census has demonstrated that noise injection can disproportionately result in larger measurement error for smaller demographic groups [6]. It is not clear from VRS's documentation how Meta manages this trade-off to ensure its fairness goals are being meaningfully achieved.

The application of DP at multiple points in the system, without a clearly articulated threat model, also reduces the ability of independent researchers to audit VRS's outputs. Meta acknowledges that this approach increases the statistical variance of the system's measurements [47], allowing for a layer of plausible deniability where deviations observed by external researchers could be attributed to DP noise rather than underlying biases in ad delivery. This also affects the ability of Guidehouse, the independent reviewer mandated by the settlement, to verify Meta's reported numbers.

To partially address the information gap, the reviewer uses synthetic data to evaluate the impact DP on VRS's performance. The reviewer concludes based on synthetic data that the noise added for DP decreases coverage on average and, as a result, does not alter their conclusions that Meta meets the compliance requirements. However, they are unable to reproduce this finding on the real-world data that they use to evaluate compliance with the settlement because "disaggregated impression data…is not available" [18].

Meta also has not disclosed key DP parameters, such as ϵ, δ, and the episode length k, making it difficult for external researchers or auditors to evaluate the level of privacy protection or the trade-offs made in VRS's implementation. Without this information, it remains unclear whether the system is striking a reasonable balance between privacy and fairness.

B Implications of Use of BISG in VRS

As discussed in §2.1, VRS measures variance by race using BISG, which outputs probabilistic estimates that a person is of a given racial group based on their surname and zip code [1, 16]. Meta's implementation of BISG uses 50% probability threshold to assign estimated race based on the group to which BISG assigns the highest probability [18]. VRS does not get individual-level classification by race but rather gets aggregate number of impressions from each group with noise added using DP (see §A.1).

While the external reviewer mandated by the settlement uses synthetic data to study the effect of use BISG-estimated race on VRS's performance, open questions remain. The reviewer evaluates the impact of the 50% probability threshold that VRS uses. They compare it with a 60% threshold and conclude that the choice of threshold "may have an impact on Variance and Coverage", but that the 50% threshold is reasonable because it is considered best practice [18]. However, the extent to which the synthetic data accurately reflects the distribution of real-world users, whether the use of BISG leads to over- or underestimating variance and coverage, and the potential advantages of adopting larger BISG thresholds remain unexplored.

C Understanding Skew in VRS's Delivery Ratio

In §4.2.1, we saw the delivery ratio for VRS does not match the even demographic split in the audiences we target. In particular, our custom audiences contain equal number of individuals from all demographic groups (see §4.1.3), but we found the delivery ratios to be skewed by race (0.42 for Black; 0.58 for White) and gender (0.45 for male; 0.55 for female) even after VRS's intervention.

We next show this gap can be explained by differences in the matching rates across demographic groups in our Custom Audiences. For all audiences we used in our experiments in §4.2.1, we first get the post-matching audience size for each group by uploading the audience list for each group separately. Meta provides an API for querying how large a Custom Audience is after they match the information in the list we upload with real user accounts. The API provides the sizes as a range for privacy reasons, so we take the midpoint of the range as an estimate. We then uses the estimated post-matching audience sizes to calculate the fraction of people included from each group.

Figure 9 shows the post-matching fractions per group. The mean fractions we observe for each group is shown by $\bar{x}$ and vertical dotted line in the figures. The fractions are 0.42 for Black, 0.58 for White, 0.46 for male and 0.54 for Female, which closely match the delivery ratios we observed in Figure 4 in §4.2.1. Therefore, even though we upload an audience with an equal number of individuals from each group, the post-matching audience sizes can still be uneven. This imbalance in matching rates skews the eligible ratio that VRS uses as a baseline, and explains the skewed delivery ratio we observe when VRS is enabled.

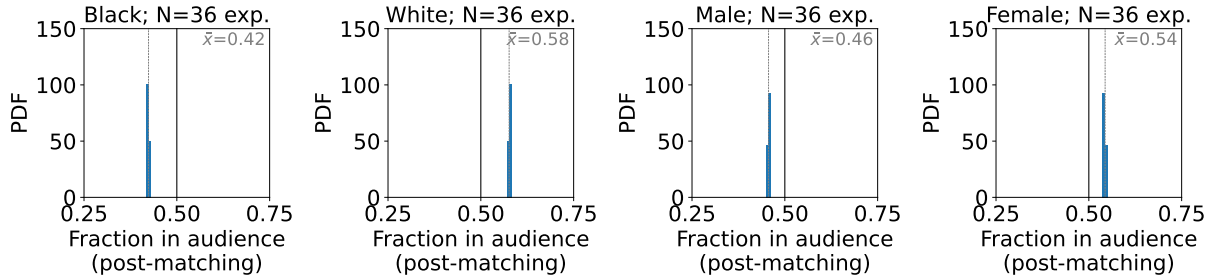


Figure 9: Fraction of individuals from each demographic group calculated after the Custom Audiences we upload are matched with real user accounts. The mean fraction for each demographic group is shown by $\bar{x}$ and vertical dotted line in the figures.