January 2023

Toward fairness in personalized ads

Meta

TABLE OF CONTENTS

Executive Summary

Fairness in personalized ads has emerged as an area of significant focus for policymakers, regulators, civil rights groups, industry and other stakeholders. Meta's early efforts to address related concerns focused on preventing potential discrimination by changing how ads (especially those offering housing, employment or credit) can be targeted by advertisers. Over time, concerns have shifted to the potential for discrimination in ads delivery–the machine learning-driven process that platforms like Meta uses to decide who within the target audience ultimately sees an ad. As a part of our June 2022 settlement with the Department of Justice, representing the US Department of Housing and Urban Development (HUD), we announced new steps we're taking to enhance equitable distribution of ads on Meta's platforms, and we're pleased to be able to more openly discuss work Meta has done over the past several years to explore and develop approaches to advance fairness across the breadth of our ads system. In this paper, we describe how our ads system and our approach to fairness have evolved; share technical details on the Variance Reductions System (VRS), a cutting-edge approach that uses new machine learning technology in ads delivery to help more closely align the demographics of an ad's eligible audience and the audience who sees that ads; and discuss critical, novel policy questions that come along with these efforts. Our hope is to catalyze discussion and consensus around how to use machine learning responsibly to fairly deliver personalized ads so they can continue to unlock economic opportunity for people and businesses alike.

Authors
Miranda Bogen, Pushkar Tripathi, Aditya Srinivas Timmaraju, Mehdi Mashayekhi, Qi Zeng, Rabyd (Rob) Roudani, Sean Gahagan, Andrew Howard, Isabella Leone

Introduction

In 2019, as part of a historic settlement with prominent civil rights organizations, Meta announced changes to help prevent discrimination in personalized ads that relate to offerings of housing, employment, and credit. While the company already prohibited advertisers from using our ad products to discriminate, as part of the settlement we took additional steps to prevent potential advertiser misuse of the platform's targeting capabilities, which substantially changed the way advertisers can target these ads. We have also worked to enhance people's access to these ads through our Ad Library so they can easily search for and view the housing, job, or finance-related information the ads might contain—regardless of whether they are in an advertiser's intended audience.

Yet around that time, external researchers began identifying additional concerns that Meta's ad delivery system, which affects who among an advertiser's target audience ultimately sees a given ad, may nudge ads to certain groups of people even when the advertisers select extremely broad targeting options.[1] Meta's civil rights auditors also noted that the civil rights community shared concerns that the algorithmic systems used for delivering ads had the potential to reflect unfair bias, even absent advertisers targeting people in an inappropriate or discriminatory manner.[2]

Within Meta and across both the research community and industry, approaches to fairness and inclusivity in the use of AI are still evolving, particularly in the realm of personalized, auction-based advertising systems. Even standards organizations have struggled to articulate clear expectations for what bias or fairness ought to mean in particular contexts, especially when faced with contradictory definitions or expectations. But we know we cannot wait for consensus to make progress in addressing important concerns about the potential for discrimination in ad delivery systems such as ours — especially when it comes to questions of housing, employment, and credit, where the enduring effects of historically unequal treatment still have the tendency to shape the economic opportunities of too many.

For that reason, a team of several dozen interdisciplinary experts at Meta, including product managers, engineers, data scientists, user researchers, policy and legal experts, and the Civil Rights Team has devoted much of the last several years to exploring and developing approaches to advance fairness across our advertising system. And, with the announcement of

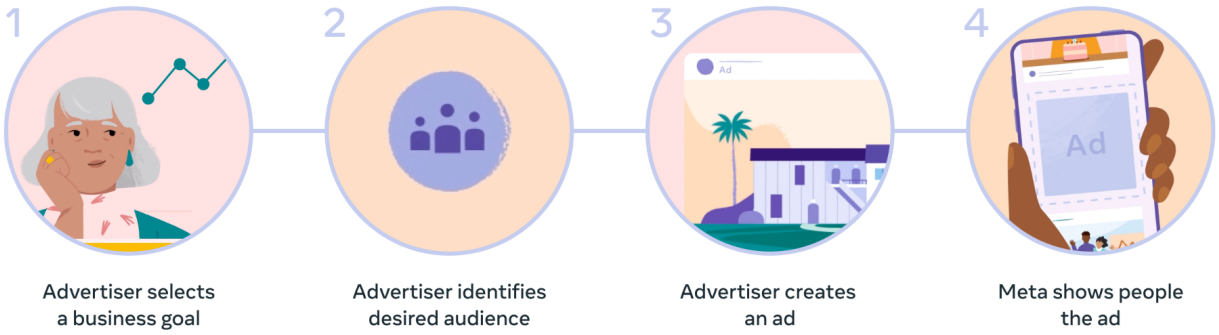
our recent settlement with the US Department of Justice in June 2022, we are excited to more openly discuss and engage with interested communities on this work.

In the following sections, we describe early changes we made to our advertising tools and chart the evolution of our approach as both the problem and solutions spaces have evolved and matured. We discuss in detail the significant changes we are making in how our system will deliver housing, employment, and credit ads in certain markets with a focus on the outcomes of that system. We then discuss challenges and questions we have been navigating, which we hope can serve as useful signposts in the critical conversations ahead about ads fairness and about machine learning fairness more broadly.

The goal of this paper is to reset the foundation for conversations about fairness in Meta's personalized advertising systems, describe some of the changes we have made and are in the process of implementing, and lay out open questions that we hope can inform active conversations among advocates, researchers, and policymakers around how to chart a path forward — for us at Meta, for our industry peers, and for the broader field.

Ads 101: From Design to delivery

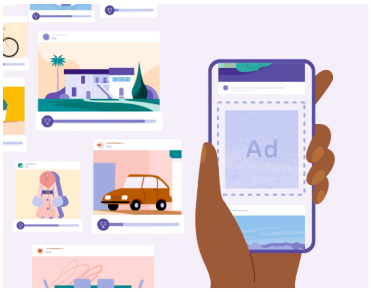
Assessing potential fairness risks in any algorithmic system starts with a baseline understanding of that system, since that helps to spot potential harms and build context-specific solutions. As such, we begin with an overview of how Meta's personalized ad system works, and how we use machine learning to deliver ads.[3]



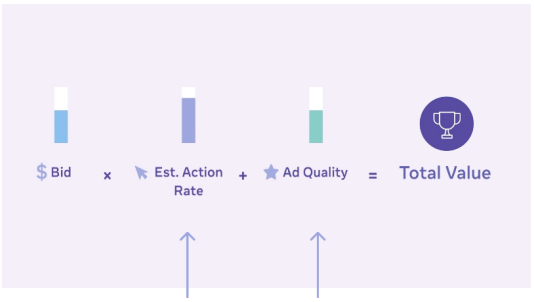
When an advertiser decides to run an ad on one of our platforms, they start by choosing their business objective, desired audience, and content of the ads.

Advertisers can choose from a set of business objectives, such as increasing visitors to their website or driving downloads of an app.[4] Advertisers may then choose the desired audience for their ad based on broad options like age, gender, location and language as well as detailed options like interests, demographic and behaviors. Advertisers can also choose to use information they already have about their desired audience, for example people who have visited their website. This stage of the advertising process is often called **targeting**.

Then, we determine which ads to show people, based on the input advertisers provided when creating the ad and the results of our ad auction. Once the advertiser has defined its business objective, selected its desired audience, and uploaded their ad content, those ads are ready to compete in ad auctions. The outcome of these auctions is sometimes referred to as **delivery.**



Gathers all the ads that include the people in the advertiser's chosen audience



The likelihood that an impression shown to this person will lead to the desired result

How interesting do we think this individual is going to find this ad? Is this a high-quality ad?

Ad with the highest total value

When an ad has a chance to be shown—such as when someone is scrolling through a Meta app and is about to hit a spot where an advertisement will be displayed—our system gathers ads that would be applicable to show to that user based on advertisers' audience choices, and moves these ads to the auction stage. The goal of the auction is that the winning ad is the one that delivers the most **total value** for both people and the advertisers.

There are three key components of total value: the advertiser bid (how much the advertiser is willing to pay for somebody to take their desired action, as specified through their business

objective), the estimated action rate, and ad quality. We use machine learning models to predict estimated action rate, or a person's likelihood of taking the advertiser's desired action — again, this action might be visiting a website, watching a video, or completing a purchase. This prediction is based on **ad delivery *inputs*** such as clicking an ad, engaging with a Page, or installing an app.

Similarly, we use machine learning models to understand more about the ad itself to gauge the ad quality and generate a quality score. This includes inputs such as the feedback from people viewing or hiding the ad, as well as assessments of low-quality attributes (for example, if it appears to have sensationalized language) that help us ensure that the ad experience is positive for people.

The auction multiples the advertiser's bid with the estimated action rate, and takes into account ad quality to calculate the total value, and the ad with the greatest total value is displayed. The impression of the ad shown is the **ad delivery *outcome.***

Mapping ads fairness concerns to our system

Advocates have long been concerned about the relationship between advertising and economic mobility, and that apprehension has extended to worries that digital advertising could usher in a new era of discrimination. The concerns can be described in the context of each part of the ads system:

- **Advertiser targeting choices:** Targeting options available on platforms that might enable advertisers to unfairly exclude people from seeing ads by targeting an audience that does not include certain communities, or by explicitly excluding certain communities from their audience — especially in the protected areas of housing, employment, and credit opportunities.

- **Ad delivery inputs:** Data used as features in machine learning models that expand ad audiences or inform how ads are delivered, and whether those signals include protected characteristics or data that could identify protected characteristics. For example, including gender in a machine learning model to predict whether an ad is likely to be relevant might improve the accuracy of that model and provide a better user experience, but might be more sensitive to use as a signal for ad delivery models in areas like housing, employment, and credit.
- **Ad delivery outcomes:** Even with seemingly neutral targeting options and model features, competitive dynamics or differences in people's interests or activity on the platform could affect how ads are ultimately distributed. These differences may be reasonable at first glance (for example, if younger people are more interested in certain types of professional opportunities, they might click on ads for those types of opportunities at a higher rate and thus see them more frequently) — but if those offline differences reflect or amplify existing disparities in society (like women having fewer role models in physics and so expressing less interest in careers in physics), some observers worry that these differences could be cause for concern.

As the landscape of potential implications of personalized advertising on civil rights has continued to evolve and come into focus, so have the actions we've taken to navigate novel policy questions and mitigate issues as they crystallize. Our work continues to be informed by ongoing conversations with internal and external experts.

Below, we outline the steps we've taken to address concerns related to each of these conceptual buckets.

Ad targeting: Working to prevent potential discrimination by advertisers

Early on, advocates surfaced concerns about how the tools we made available to advertisers might facilitate potential misuse or discrimination, such as an employer purposely excluding

people of a certain age or from a certain place from seeing their ad for an open position.

Over a number of years, we've evolved the tools we offer advertisers with the feedback from these and other stakeholders in our efforts to help prevent this sort of prohibited conduct. We've changed the way advertisers can run ads on our platform, and put additional restrictions in place across the areas of housing, employment, and credit.

For all advertisers, we;

- **Require any advertiser running ads across our platforms to read and certify their understanding of our non-discrimination policy**[5] through an educational module before they're even able to create an ad.
- **Removed ad targeting options people may find sensitive to protect against the potential abuse of our tools** across all ads, not just those about housing, employment or credit.[6] These include targeting options referencing causes, organizations, or public figures that relate to health, race or ethnicity, political affiliation, religion, or sexual orientation.
- **Maintain all active ads in our Ad Library,** so that everyone - even if they don't have a Facebook account - can see the ads being run across our services.
- **Continuously review and update the tools we make available for advertisers.** We continually review our tools and products to make them simpler, easier to use and more effective for advertisers and ensure they are consistent with our ad principles.
- **Give people the tools to hide ads** from any advertiser they would prefer not to engage with and, anytime someone sees an ad on our services, report the ad to us if they believe there's a problem.

For advertisers who run housing, employment, and credit advertisements, we;

- **Restrict how housing, employment and credit advertisers can create their target audiences.** When an advertiser identifies their ad as offering housing, employment or credit, they are not permitted to target based on gender, age, or interests that appear to describe people of a certain race, religion, ethnicity, sexual orientation, disability status, or other protected class. If they opt to target by location, that location targeting must have

a minimum 15-mile radius. The categories that remain available to these advertisers were the result of in-depth conversations with civil rights stakeholders.

- **Make Lookalike targeting unavailable** to advertisers running housing, employment, and credit ads. Initially, when we made Lookalike targeting unavailable to advertisers, we introduced an alternative called Special Ads Audiences, which were audiences selected, using machine learning, based on similarities in online behavior and activity to those on a customer list — but without considering age, gender, or ZIP code. The field of fairness in machine learning is a dynamic and evolving one, and Special Ad Audiences was an early way to address concerns by removing input features from being considered when determining similarity. However, informed by concerns that these audiences might still be misused by advertisers attempting to circumvent our protections against discrimination, we discontinued this tool for housing, employment and credit ads in all markets where it was available to shift our focus toward new approaches to improving fairness.
- **Maintain all active ads related to housing, employment or credit opportunities in our Ad Library**, giving everyone a chance to see these ads regardless of whether they were in an advertiser's intended audience.[7]

While these changes initially applied in the United States, we have expanded these updates to Canada and the EU as part of our global approach to help prevent discrimination on our platform. And just as we have for the past several years, we will continue to evaluate and evolve our ad system.

Ad delivery *inputs:* Removing Sensitive Model Features

In addition to considering what categories and model inputs are appropriate to use at the ad targeting stage, we reviewed our machine learning models that inform how housing, employment, and credit ads compete in ad auctions and the inputs into those models.

Specifically, for the machine learning algorithms that we use to deliver housing, employment, or credit opportunity ads to people residing in the US and Canada, we have undertaken efforts to

identify and remove the following types of features from the delivery of such ads when we find them:

- features that use protected characteristics reported by people, for example, in the “About You” section of the user’s profile on Facebook and/or Instagram.
- features with names that suggest they may identify protected characteristics.
- features that use interest segments that we removed as part of the 2019 settlement related to targeting in housing, employment, and credit opportunity ads.

For the purposes of these efforts, protected characteristics are defined based on major US federal and state laws and include the following: race, color, religion, national origin, disability, age, marital status, income from public assistance, exercise of any right under consumer credit protection act, familial status, pregnancy, childbirth and related medical conditions, sex, sexual orientation, gender identity, gender expression, ancestry, creed, genetic information, military/veteran status, source of income, victim of abuse, sexual assault or stalking, medical condition, citizenship, primary language, immigration status, arrest/court record, political affiliation, personal appearance, civil union or domestic partnership status, use or nonuse of lawful products off the employer’s premises during non-working hours, breastfeeding, family responsibilities, matriculation, legal marijuana use, credit history, smoking, membership or activity in a local commission, GED vs. high school diploma, declining to participate in a religious or political meeting, lawful occupation, wealth-, income- and economic status-related variables (e.g. property value, assets), length of time at address, refusal of another insurer to write policy, or cancellation/refusal to renew policy, claim history, firearm ownership, prior purchase of insurance through the associated auto insurers plan, years of driving experience, and source of payment.

We'll use existing infrastructure to identify and remove such features from the machine learning algorithms that our system uses to deliver housing, employment and credit opportunity ads to people residing in the US and Canada. We’ll make tool enhancements to support ongoing governance, and we’ll keep working to improve this functionality over time.

Ad delivery outcomes: Reducing outcome variance across demographic groups

Procedural protections are generally important steps to help curtail the potential impact of systemic inequity reflected in automated systems: in certain cases, errors in the tools and models that drive personalization systems could lead to unfair outcomes, such as if the system has not learned from sufficiently diverse data. Issues could arise, for instance, if models in such systems over- or under-predicting the likelihood people from a certain demographic group are interested in a particular type of ads, those models might thus under- or over-deliver those types of ads to different audiences based on those erroneous predictions. Using tools like Fairness Flow[8], we are working to understand if there are implementation issues in our ads machine learning systems, and we're exploring ways to address issues should we find them, such as adjusting our upstream model training approaches or parameters.

But sometimes—like in contexts where long-term historical injustice has shaped people's access to economic opportunity, like the housing market in the United States—people's own preferences and behavior may reflect circumstances shaped by these injustices, and personalization systems have the potential to reflect that legacy. In these cases, procedural improvements won't necessarily have the effect of eliminating outcome differences of concern.

Academic researchers have for some years explored different dimensions of fairness, bias, and discrimination both broadly in the context of machine learning and more specifically and in the context of online advertising.[9] However, much of the general research focuses on simple models such as binary classifiers and organic recommendation systems, characteristics of which do not map cleanly onto auction-driven recommendation systems made up of hundreds of interacting models.

This complex and evolving research and policy context have rarely resulted in off-the-shelf solutions, and further exposed the need for significant deliberation and iteration to navigate this important space. *See Appendix A for a detailed discussion of the academic literature on this topic.*

Despite these challenges, we are committed to building on protections we've already put

in place and doing our part in helping to broaden opportunities for marginalized communities in these spaces and others by striving toward a more equitable distribution of housing, employment, and credit ads through our ad delivery process. So, as part of conversations with the US Department of Justice, we designed a system we believe can help address the most acute concerns that machine learning systems might amplify concerning disparities.

Our new method, which we refer to as the Variance Reduction System, is designed to help ensure the audience that ends up seeing a housing, employment, or credit ad more closely reflects the eligible targeted audience for that ad. In other words, the system is being designed to help ensure the age, gender and estimated race or ethnicity of a housing ad's overall audience matches the age, gender, and estimated race or ethnicity mix of the population eligible to see that ad.

To facilitate this goal, the system will start by measuring the demographic distribution of the **baseline** or **eligible ratio** of the age, gender, and estimated race or ethnicity distribution of the population of users to whom an advertiser has indicated they would like their ad to be displayed.

Recall that our policies already prohibit advertisers from using our ad products to discriminate against individuals or groups of people, and that we've disallowed the use of gender, age, or ZIP code targeting, removed detailed targeting options that describe or appear to relate to protected classes, and require location targeting to have a minimum 15-mile radius. Given these restrictions, the demographic distribution of this baseline represents a reasonable foundation against which to measure the demographic distribution of the delivered impressions.

$$Eligible\ ratio_{ad,subgroup} = \frac{\sum_{ad \in\ adtargetaudience \cap subgroup} Impression_{user}}{\sum_{ad \in\ adtargetaudience} Impression_{user}}$$

Note that because we cannot predict precisely who, despite being theoretically eligible to see an ad (due to being included in the audience selected by an advertiser), will actually use a Meta product like Facebook or Instagram during the period the ad is being delivered, the baseline measurement relies on estimates based on recent ad impressions.

Then, as the ad is being delivered, we periodically measure the **delivery ratio,** or the demographic distribution of the impressions of an ad as it begins to be delivered.

$$Delivery\ ratio_{ad,\ subgroup} = \frac{Impressions_{ad,\ subgroup}}{\sum_{subgroups} Impressions_{ad}}$$

Using these two measurements, we can determine the difference between these distributions, and detect whether the delivery ratio has diverged from the eligible ratio, thus necessitating action to reduce the computational distance between these two measures.

$$Variance_{ad,subgroup} = Delivery\ ratio_{ad,subgroup} - Eligible\ ratio_{ad,subgroup}$$

How it works

Specifically, we formulate the Variance Reduction System as an offline reinforcement learning framework with the explicit goal of minimizing an ad's impression variance across the demographic subgroups. Reinforcement learning is a type of machine learning that learns from trial and error to optimize toward a predefined outcome. In this case, the outcome that the system is instructed to seek is to **minimize ad impression variance** across demographic subgroups, no matter the cause of that variance (variance has as much chance of being caused as easily by factors like competitive dynamics of ad auctions, traffic spikes, and randomness noise as it does by demographic patterns, but due to the difficulty of attributing these potential sources of variance, the VRS is instructed to work agnostic to source of variance).

Specifically, the Variance Reduction System will rely on a **controller** that has the ability to change one of the values used to calculate an **ad's total value** in our ad auction, which will have the effect of changing the likelihood that a given ad will win an auction and be shown to a user. The value that the controller will be able to adjust is called the **pacing multiplier**, which is adjusted through a **boost multiplier.**

How the Ad Auction and Pacing Affect Ad Delivery

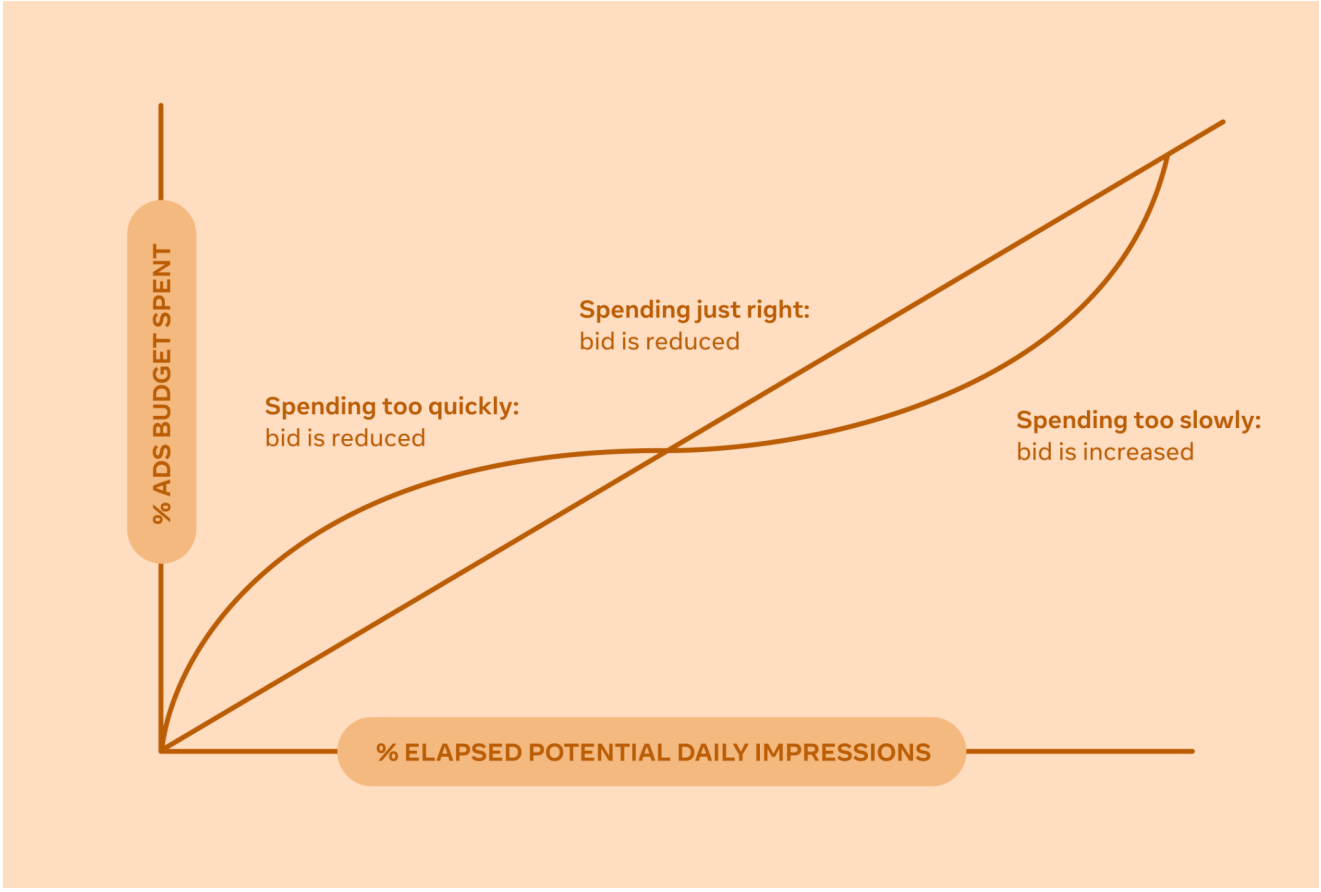
We determine which ads to show people based on two main factors: audience targeting options selected by advertisers and the results of our ad auction. Once an advertiser selects their audience, business objective (e.g. interacting with the ad, downloading an app, etc), and bidding strategy, and an ad has an opportunity to be shown, our system gathers ads that include people who fit the audience criteria in the advertiser's chosen audience and those ads move to the auction stage.

For ads that enter the auction, Meta selects the top ads to show to a person based on which ads have the highest total value score — a combination of advertiser value and ad quality. We find advertiser value by multiplying an ad's bid by the estimated action rate. This is an estimate of how likely that particular person is to take the advertiser's desired action, like visiting the advertiser's website or installing their app. We then add the ad quality score, which is a determination of the overall quality of an ad, such as whether the ad is likely to be engagement bait or whether people are likely to provide negative feedback about the ad, such as repeatedly hiding or reporting it.

We use machine learning to generate the estimated action rate and the ad quality score used in the total value equation. To find the estimated action rate, machine learning models predict a particular person's likelihood of taking the advertiser's desired action, based on the business objective the advertiser selects for their ad, like increasing visits to their website or driving purchases. To do this, our models consider that person's behavior on and off Facebook (in

accordance with the user's ad preferences and settings), as well as other factors, such as the content of the ad, the time of day, and interactions between people and ads. To generate an ad's quality score, our machine learning models consider the feedback of people viewing or hiding the ad, as well as assessments of low-quality attributes (like too much text in the ad's image, sensationalized language, or engagement bait).[10]

In addition to these signals, the auction takes into account advertisers' bidding strategy to ensure that budgets are not exhausted too quickly or slowly. For example, advertisers may request a strategy that delivers ads consistently over a defined time period, or may prefer accelerated delivery when their ad may relate to time-limited circumstances such as a live event or a flash sale. Pacing multipliers are used to adjust the auction calculation to help increase or reduce the pace at which ads are delivered, which implicates the pace at which an advertiser's budget is spent. Importantly, pacing multipliers do not necessarily change how much advertisers pay for a given ad, since final costs are dependent on other ads competing in that particular auction, ad quality, and other factors.



The Variance Reduction System will deploy boosts to this pacing multiplier function as a lever to help adjust how housing, employment, and credit ads are delivered, with the goal of reducing variance across demographic groups.

For a given ad, the controller experiments with different ways to apply multipliers that most effectively reduce impression variance. The controller is periodically provided with updated aggregated impression variance measurements that signal whether the strategy used has been effective in reducing impression variance or not, and inform whether a new strategy should be deployed.

For example, say an advertiser posts an ad for a nonprofit fundraising internship and selects as their target audience as everyone interested in charitable organizations. If the population of the target audience who regularly log onto Facebook with that interest is split equally between men and women, the eligible ratio for that ad would be 50%. As the ad begins being delivered, the Variance Reduction System will measure the proportion of impressions of that ad that are being shown across men and women. If the delivery ratio measurement indicates that ad impressions are being shown more frequently to men than to women, the controller will be able to adopt a new strategy for adjusting multipliers with the aim of shifting the distribution of impressions back toward the eligible ratio.

The controller is analogous to how an organization might work to increase diversity in its workforce. Imagine that a team starts its efforts by measuring the demographic distribution of their workforce and notices that it has not yet reached its goal to increase diversity on the team. The organization might go through another round of hiring, and then measure the aggregate demographic distribution of their workforce again to see whether they made progress. If they did, they might lean into the recruitment strategies that appeared to be successful at attracting candidates from diverse backgrounds. If they still have progress to make, they might shift their recruitment strategy and try again. For example, they might do concerted outreach to different communities, change the questions that are asked during interviews, or advise hiring managers about how to interpret nontraditional academic or professional backgrounds. After deploying a new recruitment strategy and going through another cycle of hiring, the organization might measure again, and so on, until they reach their goal.

Similarly, our controller operates in discrete episodes, or time periods between measurements. After each time period t (represented by $\mathbf{k}$ delivered impressions of the ad in question), the controller will adopt a new set of actions based on a menu of approaches it has learned have the potential to successfully reduce variance from prior, offline training.

The actions available to the controller at auction include **apply an adjustment to the multiplier**, or **do not apply an adjustment to the multiplier**. Ideally, when the controller outputs an adjustment factor for an HEC ad, the ad would be boosted to the top of the auction among all candidate ads in the auction stage, so that it has a higher chance of becoming a realized impression. Similarly, when the controller outputs no adjustment factor for an HEC ad, its goal is that the ad lands below other candidate ads competing in the auction with higher total value, and thus reduces the likelihood it becomes a realized impression. Both types of action can play a role in reducing variance. The controller is rewarded if at the end of episode t it receives signal (in the form of reduced variance) that it has more frequently selected the correct adjustment option, which translates into reduced delivery variance.

Importantly, **the Variance Reduction System will not be provided with individual-level age, gender, or estimated race/ethnicity** to make these determinations. The system will instead receive aggregate measurements across these demographics.

So, how does the system determine whether adjustments will be effective in maximizing the system's reward without awareness of this individual demographic information? To function, the system relies on the following information:

- **User features:** Information about content and ads that users have interacted with is abstracted and summarized by mapping these variables into lower-dimensional continuous values that the VRS can more effectively process. Specifically, the system will rely on a subset of the features, which are inputs into the machine learning algorithms, that are already used for the delivery of housing, employment, and credit ads. (As described in the previous section, as part of our ongoing work to respond to concerns that have been raised about ads fairness, our systems are designed to remove features that use protected characteristics reported by users, such as age and gender, or have descriptions that suggest they may identify protected characteristics for the delivery of HEC ads — the same features are made unavailable to the VRS.)
- **Ad features:** data about the ads that have been delivered to different users, such as ad text and images, user actions on ads (e.g. reporting or x-ing out of ads), and advertiser page information.

- **Impression variances:** aggregate impression variance of different ads that have already been delivered.

First, using user features and ad features, we generate user summaries based on past user-ad interactions. Then, we compare these user summaries to historical patterns of impression variances of ads that different users have seen and interacted with. We then train the VRS's core model by randomly applying boosts to hypothetical pacing multipliers in an offline simulation environment, and calculating the reward function — that is, the degree to which boosts in the simulation environment had the effect of reducing impression variance, given the user summaries associated with the relevant ad impressions. These simulations help the VRS learn a variety of multiplier adjustment strategies that may be effective to reduce variance once the system is deployed in the real-world ad delivery environment. The VRS is then equipped to deploy these strategies, and shift among them as it receives aggregate measurements of variance as a given ad is being delivered.

While user summaries are likely to have some correlation with the demographic dimensions of interest, it's important to note that signals like users' behavior can be influenced by many factors that are orthogonal to demographic characteristics — for example, time of day, current events, or pop culture moments. Moreover, there is substantial heterogeneity in user behavior across platform surfaces (e.g. Facebook Feed compared to Instagram Reels), content types, and account activity levels. Just as impression variance can be created in a system with no explicit dependence on demographic characteristics, it can be helpful to conceptualize the user summaries leveraged by the VRS as abstractions of the same broad array of features that may relate to the underlying impression variances these summaries are being leveraged to prevent. (Recall that in the case of HEC ads, certain features have already been removed from this set). We discuss privacy protections applied to help prevent these summaries from learning, and then inadvertently disclosing, data about sensitive characteristics in the Privacy section below.

Multi-objective Modeling

Ultimately, wherever it operates, the Variance Reduction System will aim to minimize delivery variance for not just one demographic dimension, but multiple dimensions simultaneously: age, gender, and estimated race or ethnicity.[11] This requires instructing how, within a particular ad auction, the system should consider each of those dimensions. This is especially important if the system is observing variance in opposite directions across multiple dimensions as compared to the baseline. For example, early on in an ad's delivery period, an ad impression variance might

appear to be improving for estimated race, but appear to be worsening for gender. If the system predicts that a given future ad impression will be helpful in reducing one dimension of variance while amplifying another, it will still need to determine what boosting action will best serve the system's goals of reducing variance across all dimensions simultaneously.

The Process

Step 1: Offline model training — Prior to the Variance Reduction System launch, user summaries are generated, and a new variance reduction model is trained in a simulation of the ads delivery production environment to learn different approaches to adjusting multipliers.

Step 2: Episode(1) — Ad delivery commences. Once an advertiser publishes an HEC ad and defines the audience eligible to see that ad, the ad is available to begin competing in auctions. For the first k auctions the HEC ad wins (resulting in an impression), the Variance Reduction System does not apply any adjustments but simply logs the impressions to facilitate measurement. Every k impressions constitutes an episode.

Step 3: Snapshot Measurement(1) — At the end of the first episode (after k impressions), the age, gender, and estimated race or ethnicity of the viewers of those impressions are counted in aggregate to measure the initial variance for that campaign. To prioritize privacy within the system and its resulting measurements, differential privacy noise is added to these counts (see the Privacy section below for additional detail).

Step 4: Episode(2,3,4,...n) — The multiplier adjustment strategy is applied. Based on the aggregate measurements and whether they reveal an apparent variance for any of the included dimensions, the Variance Reduction System adopts a multiplier adjustment strategy and begins deploying it during the auction. At auction time, the Variance Reduction System receives summarized data about the potential ad viewer (recall the system does not receive individuals' demographic data like their gender, age, and estimated race) and applies the adjustment that the multiplier adjustment strategy selected for this episode predicts is most likely to contribute toward reduced variance. Like in the initial episode, if the HEC ad wins the auction, the system logs the impression to facilitate aggregate measurement at the end of the episode.

Step 5: Snapshot Measurement(2,3,4,...n) — At the end of each episode, the system measures the impression variance across the included dimensions and informs how the system will adjust its strategy in the next episode.

Step 6. Performance Evaluation — Once the ad's campaign ends (whether because the campaign end date passed, the ad budget has been exhausted, or the advertiser manually terminates or pauses ad delivery), the system performs a final measurement to determine how much variance was reduced across each dimension and whether the VRS was adequately performant for the given ad. These measurements can then be made available for purposes like compliance reporting, facilitating third-party review[12], and periodic online training to help improve the system's performance over time.

Privacy

One of the key priorities of the Variance Reduction System is to reduce variance for ads delivery in a privacy-preserving way. In particular, we aim to avoid demographic information making its way into the Variance Reduction System or being discernable to human analysts reviewing the VRS or its outputs. To implement the Variance Reduction System, we plan to use the following privacy-preserving approaches:

- **The Variance Reduction System will not have access to individuals' age, gender, or estimated race or ethnicity.** The system will only receive user summary vectors (which do not include as inputs age, gender, or estimated race), alongside aggregate demographic measurements.
- **Race and ethnicity will be measured using Meta's privacy-enhanced implementation of Bayesian Improved Surname Geocoding (BISG),** which as we have previously described incorporates numerous privacy adaptations designed to ensure BISG classifications are aggregated and that tools avoid generating individual-level race inferences.[13] Additionally, we have further adapted the method to incorporate differential privacy at calculation time to help prevent individual predictions from becoming discernable in the event that multiple, similar queries are compared.
- **Aggregate demographic measurements that are generated and used by the Variance Reduction System will include differential privacy noise** to help prevent the system learning and subsequently acting on individual-level demographic information with high

fidelity. We note that while these privacy protections may slightly degrade the maximum performance of the system in reducing variance, we aimed to strike a reasonable balance between advancing fairness in an important context and honoring people's privacy.

Measurement

Since the Variance Reduction System will be an iterative, measurement-based system, we discuss here some details about how the system measures variance both during and after an ad's run. As described above, the system takes snapshot measurements after every k ad impressions, and performs performance evaluation measurements once an ad stops running. While variance for demographics that are implemented as binary variables (e.g. gender and age[14]) can be measured using simple ratios, race is a multi-category variable[15] and so requires a distinct approach. To accommodate this difference, we will measure the variance for this dimension using **shuffle distance.** Shuffle distance is the minimum fraction that needs to be moved (or shuffled, hence the name) from an actual distribution p (in our use case this translates to actual impressions delivered), to match a reference distribution π (in our use case this translates to baseline). Mathematically, shuffle distance can be calculated as half of L1 distance between distribution p and π:

$$Shuffle\ Distance\ = \frac{||p-\pi||}{2}$$

Intuitively, a shuffle distance of 10% means that if 10% of impressions had been moved from group A to group B, we would achieve a variance of 0.

Both the snapshot measurements and performance evaluation measurements compare the eligible ratio and the delivery ratio to measure the ad's variance across the relevant demographic groups — and both measurements incorporate differential privacy — but the measurements differ by necessity in several key ways:

First, the snapshot measurements calculate the eligible ratio by using an estimate of who could have been eligible to see the ad, given the advertiser's audience selection. An estimate is required because while Meta can observe the distribution of users who have logged on and saw an ad impression over a given prior time period, it cannot perfectly predict who will log on and have the opportunity to see an ad in a future time period. For this reason, the snapshot

measurements must compare the estimated eligible ratio with the realized distribution ratio. The VRS can then act based on the comparison between these two ratios.

The performance evaluation, meanwhile, has the benefit of being able to directly observe the realized distribution of people who had the opportunity to see an ad during the period when the HEC ad being measured was running, and can therefore use that information to calculate both the realized eligible ratio and the realized distribution ratio.

The inherent uncertainty of estimated measurements compared to realized measurement unavoidably complicates the functioning of the Variance Reduction System, because the system only learns at the end whether the realized distribution was aligned with the estimated one, and to what degree. If the distribution differed — such as if an unexpected influx of people unexpectedly visited a Meta platform for a major social or cultural event — even the best attempts of the VRS to address the variance that it observed as the ad was being delivered might leave some residual variance that it could not have addressed due to this intrinsic measurement gap.

Second, snapshot measurements consider many fewer instances (k) than performance evaluation measurements ($\geq k*n$, which represents the total number of impressions of the ad that were delivered). While differential privacy is applied in both circumstances, the degree of differential privacy noise that is needed to achieve the desired ε may differ. This will naturally add noise to the comparison between the two calculated ratios, which may constrain the maximal efficacy of the VRS.

Despite these constraints, we are confident that the Variance Reduction System will still lead to a substantial, albeit imperfect, reduction in variance.

Technical challenges and open policy questions

In developing this new method, we encounter a number of technical challenges, as well as open policy questions. We hope that by articulating the questions and tensions we have navigated and are still working to address, we can help inform conversations across industry and with experts

in civil society, academia, and policymaking so we can chart a responsible path forward together.

Technical Challenges

Availability of demographic data. Gender and age are data more readily available to tech platforms since they are commonly collected at account creation. But research has shown that for many companies, the absence of labeled demographic data, particularly race and ethnicity, has raised significant barriers to systematically investigating the potential for differences across those protected characteristics, or to monitoring the progress of fairness efforts over time. Our experiences bear this out: without labeled data regarding certain protected characteristics, we initially found it challenging to assess concerns that our products may replicate or amplify societal biases, and solving this challenge was a prerequisite to developing a system like the Variance Reduction System. Once we introduced a framework for studying our platforms and identifying opportunities to increase fairness when it comes to race, we were able to think creatively about how we might address key concerns about certain products like ads. However, following the recommendations of external privacy and civil rights experts, that framework specifically prioritized approaches designed to avoid Meta directly collecting, holding, or inferring individually identifiable race/ethnicity, which has constrained to some extent the technical solutions that would be available to address concerns. For that reason, we're designing the VRS in a way that relies on aggregated demographic measurements. We expect that similar questions will emerge if there are calls for the VRS or similar remedies to be applied to other protected categories where widely-used measurement methods do not yet exist, and where the direct collection of such data would raise important privacy questions. This will be a fundamental challenge for the advancement of algorithmic fairness across additional underserved communities, and even more so for considering intersectionality in such work.

Multi-objective learning. The Variance Reduction System will simultaneously prioritize multiple objectives: reducing variance for gender, for age, and for estimated race or ethnicity — and to do so for all effectively, must not consistently prioritize any one of those goals over the other. Like most societal challenges, navigating multiple objectives and potential tradeoffs will likely mean that multi-objective systems are unlikely to be able to achieve every component goal perfectly. In the case of the Variance Reduction System, this means it will likely not be possible to entirely remove variance in all instances, because the system can't perfectly predict how many opportunities to deliver an ad impression to any given demographic group will manifest before

the ad completes its run (and this gap is compounded by the measurement noise described in the previous section). It must work with the information it receives in real-time, and adapt as well as it can amid that uncertainty to most effectively reduce all the variances it is tasked with reducing.

Low volume ads. Because the Variance Reduction System will work in episodes—that is, it delivers k ads, measures variance, then shifts strategy for the next set of k ads with the aim of reducing variance before the following measurement—the number of ad impressions in a given ad's run relate directly to how many opportunities the system has to find successful strategies to reduce variance. For example, with a k of 15 and an ad with 15,000 impressions, the VRS will have 1000 opportunities (that is, 1000 episodes) to find and deploy effective strategies to reduce variance, while an ad with 150 impressions will only have 10 such opportunities (10 episodes). So few chances mean the system may not be able to reduce variance as much as might be expected, but reducing the size of k to increase the number of episodes has direct implications for the degree to which the system can uphold privacy protections. Ultimately, we seek to identify a value of k that maximizes the number of available episodes per ad run (and thus the effectiveness of the VRS) while maintaining a reasonable degree of aggregation, augmented by differential privacy. Because of this, it is possible that the VRS will achieve more significant reduction of variance in ads with a high volume of impressions, and face some challenges for ads with a lower volume of impressions. We will monitor the performance of the system across ad campaigns of various sizes, and continue to explore techniques to ensure the system meets its variance reduction expectations.

System latency. Because the Variance Reduction System works based on iterative measurements that rely on several pieces of technical infrastructure, there will be latency between when measurements are taken and when those measurements are shared back to the VRS. This challenge may be exacerbated when there are sudden spikes in an ad's delivery. We intend to seek ways to reduce latency time through infrastructure improvements, which we anticipate helping system performance improve over time.

Open Policy Questions

How should platforms navigate tensions between privacy and fairness? Similar to the questions raised by race and ethnicity, we expect that similar questions will emerge if there are calls for approaches like the Variance Reduction System to be applied to other protected categories.

Unlike race in the US, though, where widely-used measurement methods have already been established, other demographic dimensions do not have such established measurement methods and in fact, may raise even more acute privacy questions. In jurisdictions with strict data protection requirements that lack an articulation of what legal basis would permit data processing for the purposes of fairness measurement or interventions (both in the context of ads fairness as well as more broadly), the hurdles loom even larger. Meanwhile, prominent researchers in the field of machine learning fairness have noted that there may be compelling reasons to include relevant demographic characteristics in order to more equitably address variance in outcomes, but they note that legal constraints may preclude such approaches.[16] These tensions will be fundamental challenges for the advancement of algorithmic fairness in the context of ads and beyond, as well as across additional underserved communities — and even more so for considering intersectionality in such work.

Is there a role for affirmative outreach? In certain contexts, such as with US federal contractors obligated to take specific action to diversify their recruitment efforts, affirmative outreach to underserved populations is not only expected but required. However, options to enable good faith outreach to sensitive populations about economic opportunities have been increasingly limited and many are no longer available whatsoever, despite statements of caution from public interest organizations and employers engaged in that affirmative outreach. Are there any cases where such outreach ought to be allowed, especially as a way to remedy societal gaps, whether longstanding or newly identified? Similar questions are under active consideration in judicial and policy contexts, and will have direct bearing on the availability of certain approaches to addressing ads fairness concerns.

What is the right granularity of "opportunity" to consider? Conversations about ads fairness often focus on the allocation of opportunity via advertisers and advertising platforms. Ad impressions, while a rational level to consider, are functionally distinct from the underlying opportunities they may be describing. Since many housing, employment, and credit transactions happen off-platform, online platforms may not know which ads and with what frequency those messages truly enhance or impact people's economic security and wellbeing. Different people who see these ads may value this information differently and may choose differently whether or not to seek out the job, apartment, or financial resource in question, or may prefer different information than what advertisers choose to share before doing so. Moreover, people may seek out information about these opportunities across different online and offline platforms, and so

measuring impression variances in a siloed manner might obscure disparities that still exist in the broader marketplace.

What is the right baseline against which to compare? When considering whether there is a difference in outcomes that ought to cause concern, a critical factor is the baseline against which outcomes are compared. In the case of ad delivery, the question is: what is the appropriate base population against which to compare whether there is an improper variance? Even the most neutral of audience selection options will reveal differences in audience distribution — the population of a state in the southwest United States will vary significantly from that of a large region in the northeast, so even that bluntly targeted audience might lead to some observation of outcome differences. Given the significant limitations that have been placed on the targeting phase of HEC ads that will be subject to the Variance Reduction System, we agreed with the US government that the initial targeted audience for HEC ads—that is, people who would be eligible to see an ad, is the right baseline. We note that this baseline does not take into account whether, even among that population of people who are eligible to see the ad, there may be substantially different levels of interest in the ad in question. In the context of employment, determinations of outcome differences in employment selection procedures take into account qualifications for the role in question, while in the context of credit, financial circumstances are acceptable factors to control for when measuring outcomes. Long-term questions in the context of ads might include: are there reasons, such as different underlying interests people have, that might inform the measurement of variances? If one population is substantially less interested in particular vocational opportunities than another, to what extent is the expectation that information about that opportunity is nonetheless distributed evenly, regardless of such factors — especially if that means other vocational opportunities that may be of greater interest could be displaced in the interest of more balanced ad delivery across populations? Do the answers to these questions change if people's preferences to what information they would like to see are stated explicitly rather than being predicted? Policy conversations on this topic remain nascent.

How should the wider industry address this challenge? In the long run, it may be important to consider how economic value itself is allocated not only within but across platforms, and identify steps that can be taken to more equitably distribute information that people find valuable and actionable to advance their economic circumstances in that broader context while still delivering value through personalized experiences. This is all the more important because algorithmic

fairness is an industry-wide issue – if digital advertising becomes more constrained on one platform, advertisers will move to another that has not implemented such safeguards. How can we level the playing field so the problem is not simply displaced?

Conclusion

The field of fairness in machine learning is a dynamic and evolving one, and the changes described in this paper represent several years of progress in consultation with a broad array of stakeholders. Much of this work is unprecedented in the advertising industry and represents a significant technological advancement for how machine learning is responsibly used to deliver personalized ads. We are excited to pioneer this effort, and we hope that by sharing key context and details about how we are tackling this multidimensional challenge that other AI and digital advertising practitioners can more easily adopt and take similar steps to help prevent discrimination and avoid amplifying societal biases whose impact extends far beyond any one platform.

Beyond our advertising system we continue to pursue work to embed both civil rights considerations and responsible AI into our product development process, some of which was shared in Meta's civil rights audit progress report released in late 2021. We know that our ongoing progress — both in ads fairness and broader civil rights initiatives — will be determined not just by our commitment to this work, but by concrete changes we make in our products. We look forward to not only building solutions, but also participating in and supporting the critical, industry-wide conversations that lie ahead.

Appendix: Ads Fairness Research Landscape

Academic researchers have for some years explored different dimensions of fairness, bias, and discrimination both broadly in the context of machine learning and more specifically and in the context of online advertising.[17] However, much of the general research focuses on simple models such as binary classifiers and organic recommendation systems, characteristics of which do not map cleanly onto auction-driven recommendation systems made up of hundreds of interacting models.

Fairness in binary classification

In the context of binary classifiers, researchers have grappled with contradictions, impossibilities, and unintended consequences of fairness metrics. For example, the popular notions of statistical or demographic parity (equal positive classification rates) or equality of opportunity (equal false positive/negative rates) have the potential to inadvertently harm marginalized groups.[18] Some have noted that demographic parity (often analogized to the legal notion of disparate impact), when applied in contexts where underlying behavioral base rates differ across groups, necessitates setting different thresholds for making the decision in question — that is, applying unequal treatment for the purpose of ensuring similar outcomes.

Conditional demographic (dis)parity, meanwhile, focuses on the comparative proportion of relevant populations who experience the positive or negative outcome of interest[19], but requires consensus on the set of factors that can legitimately be considered conditional with respect to that outcome. For example, in the case of measuring the proportion of a university's applications who are accepted across demographic groups conditioned on academic qualification, the challenge rests in the definition of relevant academic qualifications, especially when upstream inequities might have affected which communities have more ready access to advantageous academic resources to prepare for the university admissions process.

Practitioners at Meta have previously described a preference for calibration-based approaches to fairness in the context of binary classifiers (considering whether a model has the propensity to over- or under-predict the outcome in question at greater rates for given populations) since it avoids the potential unintended consequences exhibited by other fairness metrics.[20]

Notwithstanding their relative merits and limitations, these metrics are ill-suited to the context of auction-based personalized ads systems, which are substantially dissimilar from binary classifications in that there is significantly less clarity in notions of “positive” or “negative” outcomes. The delivery of an employment ad could be a positive outcome, for example, if the substance of the ad is of sufficient interest either to the individual who sees it, or to the extent it serves a societal goal of increasing broad access to information about an opportunity even if that means some individuals who ultimately see that ad are agnostic to or disinterested in the ad. But the delivery of that ad could also be a neutral to negative outcome if the substance of the ad directs the person to low-value information, or if that ad displaces another employment or economic opportunity-related ad that is of more direct interest or utility to the person. As such, many established fairness tools that are tuned for binary classification problems have not been readily deployable to respond to concerns around unfairness in personalized ads systems, and additional foundational research has been needed to identify paths forward that are most appropriate to this specific context.

Fairness in ranking and recommendation systems

Research on fairness in recommender systems has included exploration of the comparative exposure of items in ranked lists such as businesses or job candidates surfaced in search results[21] with groupwise differences commonly measured against the demographic or other relevant subgroup information related to the ranked items. Related concepts include the diversity and/or proportional representation of the items ranked[22], and often explore how to impose such representational fairness constraints while preserving utility of the original ranking quality metrics to consumers of the ranked information[23]. Researchers at Meta have explored the notion of envy-free recommendation, under which a system would be fair if each user prefers their recommendations to those of all other users.[24] And some have explored fairness in reciprocal recommendation systems such as dating websites and marketplaces.[25] Online advertising most closely resembles such a two-sided system where both producers and consumers have interests in the fairness of ranked items[26], with further confounders deriving from auction dynamics.

However, the most prominent concerns about algorithmic bias in online ads relate most substantially to audience-side distribution — that is, the demographic distribution of the people exposed to each ranked item. This lens presents substantially different implications than the demographic diversity or representation of the ranked items themselves. To measure this concern, ranked items are measured against the aggregate distribution of people who observed or interacted with an impression of that advertisement. Because measurements here rely on characteristics of entities other than the ranked item itself, opportunities for item exposure are differently constrained than who sees the item and are thus not pertinent to the fairness concern in question. For example, imagine that more women than men log on to a recommendation platform during the period of time when an item is eligible to be ranked — during that period of time, there will naturally be more opportunities to expose the item to women, regardless of how the item is scored by the ranking model. For this and similar reasons, concepts from the literature around fairness in ranking are informative, but not directly applicable, to fairness in personalized ad systems.

Fairness in personalized advertising

Early research around fairness in the context of online advertising focused on advertiser targeting options as a vector for intentional discrimination via explicit inclusion or exclusion of protected characteristics[27], attributes that may be close proxies for those characteristics[28], or both.[29] Research regarding implications of algorithms in ad delivery initially focused on the potential for offensive associations between search terms and the content of keyword ads in search engine advertising[30] and the impact of competitive auction dynamics on the demographic distribution of online ad delivery.[31]

These foundational works inspired further research noting how personalization algorithms could lead to deviations between the demographic distribution of an advertiser's targeted audience and the demographic distribution of the audience who ultimately saw the advertiser's message[32], positing that such variances between potential and actual audience composition must indicate that delivery algorithms improperly rely on stereotypes to predict ad relevance. This stream of research and related media coverage argued that any substantial variance between targeted and actual audience — and in some cases, any demographically skewed outcomes regardless of an advertisers' targeting selections[33] — should be viewed with suspicion. Some have proposed interventions designed to ensure demographic parity of

outcomes, to be accomplished by subdividing advertising campaigns into demographic-specific instances to enforce proportionality in spend across demographics.[34] While theoretically responsive to prior research about outcome variances, such a proposal may not be realizable in circumstances where the direct use of demographic characteristics to affect system decisions or behavior is disallowed.

Other research, in the context of employment ads, aims to differentiate between skews in ad delivery that might be caused by differences in professional qualification from those that appear to be driven by protected category membership.[35] Here, the researchers compared the distribution of ads on LinkedIn (a platform with explicit knowledge of professional qualifications) with Meta (which lacks such structured data); we posit that the differences in results the researchers found could be confounded by the underlying data about professional qualifications that are available to the advertising systems in question, but more research is needed to determine the cause of differences in these systems' outcomes. Moreover, this methodology suggests it may be reasonable to control for certain factors when comparing how certain segments of personalized ads are delivered, raising questions about what underlying factors may be appropriate for segmenting analysis in other domains where qualification information is less readily available or more obscure.

Recent research has offered a possible response to this quandary, exploring definitions of fairness such as envy-freeness which proposes that systems might be considered fair if no individual prefers another individual's outcome over their own. Such definitions are thus contingent on parameters such as people's individual preferences, suggesting that accounting for individual preferences when considering the fairness in a personalized ad system may be acceptable.[36] We note that such an approach, while intuitive, might still result in observed outcome differences across demographic groups if individual preferences tend to differ across those communities, which may be less palatable from a collective or societal standpoint. (This tension in particular is one that is unlikely to be resolved by the development of new fairness methodologies or techniques, but rather by deliberation among stakeholders and policymakers to articulate the preferred path forward).

Meanwhile, related research has explored to what extent targeting tools commonly used across online advertising platforms such as lookalike audiences may reflect or amplify the underlying demographic distribution of source audiences[37], and whether removing features as inputs to

such models has an observable effect on audience composition.[38] The majority of such research about ad delivery outcomes, and machine learning fairness more broadly, presumes ready access to relevant protected characteristics for the purposes of evaluation or mitigation, which are often not available (or their use substantially constrained) in practice.[39]

In sum, the increasingly lively field of research and commentary in the field of machine learning fairness simultaneously adds helpful nuance to the conversation and proposes divergent—and sometimes contradictory—fairness metrics and related interventions for online advertising systems. The complex and evolving research and policy context has meant off-the-shelf solutions are rare and significant deliberation and iteration have been required to navigate this important space.

References

1. Ali, Muhammad, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. "Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes." *Proceedings of the ACM on human-computer interaction* 3, no. CSCW (2019): 1-30. https://arxiv.org/abs/1904.02095
2. Murphy, Laura, and Megan Cacace. "Facebook's civil rights audit–Final report." *Facebook* (2020). https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf
3. Good Questions, Real Answers: How Does Facebook Use Machine Learning to Deliver Ads? https://www.facebook.com/business/news/good-questions-real-answers-how-does-facebook-use-machine-learning-to-deliver-ads
4. How to choose the right Meta Ads Manager objective, https://www.facebook.com/business/help/1438417719786914
5. Meta's Nondiscrimination Policy, https://www.facebook.com/certification/nondiscrimination/
6. Removing Certain Ad Targeting Options and Expanding Our Ad Controls, November 9, 2021, https://www.facebook.com/business/news/removing-certain-ad-targeting-options-and-expanding-our-ad-controls
7. Ad Library, https://www.facebook.com/ads/library
8. Fairness Flow is a technical toolkit developed by Meta researchers in consultation with external experts to enable our teams to analyze how some types of AI models and labels perform across different groups, which can help address fairness concerns in our products and services. https://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone/
9. E.g. Barocas, Solon, Moritz Hardt, and Arvind Narayanan. "Fairness in machine learning." *NIPS Tutorial* 1 (2017): 2. https://fairmlbook.org/
10. Most of this comes from https://www.facebook.com/business/help/430291176997542?id=561906377587030
11. Our initial launch will focus on gender and estimated race. Note that Meta currently only has methods to measure estimated race or ethnicity in the United States, inspired by methods that are broadly used by other industries and regulatory agencies; we cannot guarantee similar methods that can provide sufficient privacy protections yet exist in other markets.
12. Pursuant to [the settlement] an independent third-party Reviewer will review each Compliance Report and verify compliance with the VRS Compliance Metrics. https://www.justice.gov/opa/press-release/file/1514031/download
13. Alao, Rachad, Miranda Bogen, Jingang Miao, Ilya Mironov, and Jonathan Tannen. *How Meta is working to assess fairness in relation to race in the US across its products and systems*. Technical Report. Meta AI, 2021.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
14. While we recognize gender is not a binary identity, we are constrained by the data that is internally available. For the purposes of the Variance Reduction System, age is categorized as under 40 and over 40.
15. We rely on the US Census categories as incorporated into the Bayesian Improved Surname Geocoding methodology. In general, the method outputs a set of probabilities that an individual belongs to one of six groups: Hispanic; Non-Hispanic white; Non-Hispanic/Latinx Black; Native American/Alaskan Native; Asian, Native Hawaiian, or Pacific Islander; and Multiracial. For the Variance Reduction System to operate with sufficiently small episodes (which increases the proportion of ads for which the system can effectively reduce variance) while still upholding Differential Privacy guarantees, we faced mathematical limitations which imposed constraints for the smallest demographic groups. As such, the Variance Reduction System will attempt to reduce variance across the largest four categories: Hispanic/Latinx, white, Black, and a combined category of remaining groups.
16. Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. "Algorithmic fairness." In *Aea papers and proceedings*, vol. 108, pp. 22-27. 2018. https://www.cs.cornell.edu/home/kleinber/aer18-fairness.pdf; Lipton, Zachary, Julian McAuley, and Alexandra Chouldechova. "Does mitigating ML's impact disparity require treatment disparity?." *Advances in neural information processing systems* 31 (2018). https://arxiv.org/pdf/1711.07076.pdf
17. E.g. Barocas, Solon, Moritz Hardt, and Arvind Narayanan. "Fairness in machine learning." *NIPS Tutorial* 1 (2017): 2 https://fairmlbook.org/
18. Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. "Delayed impact of fair machine learning." In *International Conference on Machine Learning*, pp. 3150-3158. PMLR, 2018. https://arxiv.org/abs/1803.04383; Hu, Lily, and Yiling Chen. "Fair classification and social welfare." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 535-545. 2020. https://dl.acm.org/doi/abs/10.1145/3351095.3372857
19. Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic decision making and the cost of fairness." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797-806. 2017.https://arxiv.org/pdf/1701.08230.pdf; Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI." *Computer Law & Security Review* 41 (2021): 105567. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3547922
20. Bakalar, Chloé, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall et al. "Fairness on the ground: Applying algorithmic fairness approaches to production systems." *arXiv preprint arXiv:2103.06172* (2021). https://ai.facebook.com/research/publications/applying-algorithmic-fairness-approaches-to-production-systems/
21. Singh, Ashudeep, and Thorsten Joachims. "Fairness of exposure in rankings." In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2219-2228. 2018. https://arxiv.org/pdf/1802.07281.pdf

22. Singh, Ashudeep, and Thorsten Joachims. "Fairness of exposure in rankings." In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2219-2228. 2018. https://arxiv.org/abs/1802.07281; Geyik, Sahin Cem, Stuart Ambler, and Krishnaram Kenthapadi. "Fairness-aware ranking in search & recommendation systems with application to linkedin talent search." In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2221-2231. 2019.https://dl.acm.org/doi/10.1145/3292500.3330691
23. Celis, L. Elisa, Damian Straszak, and Nisheeth K. Vishnoi. "Ranking with fairness constraints." *arXiv preprint arXiv:1704.06840* (2017). https://arxiv.org/abs/1704.06840
24. Do, Virginie, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. "Online certification of preference-based fairness for personalized recommender systems." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, pp. 6532-6540. 2022.https://ojs.aaai.org/index.php/AAAI/article/view/20606
25. Do, Virginie, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. "Two-sided fairness in rankings via Lorenz dominance." *Advances in Neural Information Processing Systems* 34 (2021): 8596-8608. https://proceedings.neurips.cc/paper/2021/file/48259990138bc03361556fb3f94c5d45-Paper.pdf; Zhou, Quan, Jakub Marecek, and Robert N. Shorten. "Subgroup Fairness in Two-Sided Markets." *arXiv preprint arXiv:2106.02702* (2021).https://arxiv.org/abs/2106.02702
26. E.g. Do, Virginie, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. "Two-sided fairness in rankings via Lorenz dominance." *Advances in Neural Information Processing Systems* 34 (2021): 8596-8608.
27. Datta, Amit, Anupam Datta, Jael Makagon, Deirdre K. Mulligan, and Michael Carl Tschantz. "Discrimination in online advertising: A multidisciplinary inquiry." In *Conference on Fairness, Accountability and Transparency*, pp. 20-34. PMLR, 2018. http://proceedings.mlr.press/v81/datta18a/datta18a.pdf
28. Speicher, Till, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. "Potential for discrimination in online targeted advertising." In *Conference on Fairness, Accountability and Transparency*, pp. 5-19. PMLR, 2018. http://proceedings.mlr.press/v81/speicher18a.html
29. Dalenberg, David Jacobus. "Preventing discrimination in the automated targeting of job advertisements." *Computer law & security review* 34, no. 3 (2018): 615-627. https://www.sciencedirect.com/science/article/pii/S0267364917303758
30. Sweeney, Latanya. "Discrimination in online ad delivery." *Communications of the ACM* 56, no. 5 (2013): 44-54. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2208240
31. Lambrecht, Anja, and Catherine Tucker. "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads." *Management science* 65, no. 7 (2019): 2966-2981. https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2018.3093
32. Ali, Muhammad, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. "Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes." *Proceedings of the ACM on human-computer interaction* 3, no. CSCW (2019): 1-30. https://dl.acm.org/doi/abs/10.1145/3359301; Kayser-Bril, Nicolas. "Automated discrimination: Facebook uses gross stereotypes to optimize ad delivery." *Algorithm Watch,* October 18 2020. https://algorithmwatch.org/en/automated-discrimination-facebook-google/
33. Kingsley, Sara, Clara Wang, Alex Mikhalenko, Proteeti Sinha, and Chinmay Kulkarni. "Auditing digital platforms for discrimination in economic opportunity advertising." *arXiv preprint arXiv:2008.09656* (2020).https:///arxiv.org/pdf/2008.09656.pdf
34. Gelauff, Lodewijk, Ashish Goel, Kamesh Munagala, and Sravya Yandamuri. "Advertising for demographically fair outcomes." *arXiv preprint arXiv:2006.03983* (2020).https://arxiv.org/abs/2006.03983
35. Imana, Basileal, Aleksandra Korolova, and John Heidemann. "Auditing for discrimination in algorithms delivering job ads." In *Proceedings of the Web Conference 2021*, pp. 3767-3778. 2021.https://www.ftc.gov/system/files/documents/public_events/1582978/auditing_for_discrimination_in_algorithms_delivering_job_ads.pdf
36. Kim, Michael P., Aleksandra Korolova, Guy N. Rothblum, and Gal Yona. "Preference-informed fairness." *arXiv preprint arXiv:1904.01793* (2019). https://arxiv.org/abs/1904.01793
37. Zang, Jinyan. "How Facebook's Advertising Algorithms Can Discriminate By Race and Ethnicity." *Case Studies in Public Interest Technology* (2021): 42.https://techscience.org/a/2021101901/
38. Sapiezynski, Piotr, Avijit Ghosh, Levi Kaplan, Aaron Rieke, and Alan Mislove. "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*, pp. 609-616. 2022. https://www.upturn.org/work/algorithms-that-dont-see-color/
39. Bogen, Miranda, Aaron Rieke, and Shazeda Ahmed. "Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination." In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 492-500. 2020. http://arxiv.org/pdf/1912.06171.pdf; Andrus, McKane, Elena Spitzer, Jeffrey Brown, and Alice Xiang. "What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 249-260. 2021. https://arxiv.org/abs/2011.02282