Measuring the Harms of Personalization through Advertising

by

Muhammad Ali

A dissertation submitted in satisfaction of the requirements

for the degree of Doctor of Philosophy in Computer Science

in the Khoury College of Computer Sciences of Northeastern University

in Boston, Massachusetts, United States of America

Committee in charge:

Alan Mislove, Northeastern University, Chair
Piotr Sapieżyński, Northeastern University
Saiph Savage, Northeastern University
Christo Wilson, Northeastern University
Elissa M. Redmiles, Georgetown University

Fall 2023

Abstract

Modern online services like social media and content sharing platforms almost always have more content than what a single user can consume. Therefore, they often have to select a subset of content that is shown to each user. To perform this selection, personalization has emerged as a nearly ubiquitous solution, where users are presented only the content that the platform considers most *relevant* to them. This mechanism can lead to users having varied and unique experiences, focused on content that they find interesting. Personalization is also a useful tool to demote content that might be low quality, or to introduce users to novel content. However, despite increasing engagement, there are concerns that it can lead to adverse user impacts for certain domains of content. Within these domains, judging relevance can harm users by excluding them from potentially life-changing opportunities, such as those for jobs and housing. It can also lead to inclusion of problematic content into users' experiences, such as misinformation and deceptive offers.

In this dissertation, I focus on the harms of personalization. I use targeted advertising as a case study of personalization and measure the adverse effects it could have on users. I focus specifically on Facebook's advertising system, the second largest, and one of the most feature-rich advertising platforms in the world. First, I build a series of algorithm auditing methods to isolate the influence of personalization with only black box access to the advertising platform. Second, I employ these methods to measure adverse user outcomes in multiple ad domains: jobs, housing, political campaigns, and problematic ads, as judged by users.

Advertising relies heavily on personalization to identify the users who an ad is relevant to. However, unlike other systems, it allows the advertisers to specify an explicit target audience. I leverage this targeting mechanism to design experimental methods that isolate the role of personalization specifically, and control for variance in audience and market competition. Advertising platforms also frequently provide explanations to the users about why an ad was shown to them. I exploit these explanations to measure the impact of personalization in observational data collected from real users.

In a series of large scale measurements, I first identify gender and racial disparities in job and housing ads due to personalization, even when the advertiser targets broadly. Second, for election campaign ads, I find evidence of personalization along users' political leaning as well as price differentiation, which potentially limits a candidate's ability to reach non-aligned voters. Third, I investigate individual user experiences of problematic ads, finding that ads considered clickbait and scam by users deliver primarily to a minority of users, who have disparately high exposure to such content. Together, these measurements quantify the extent to which personalization—independent of the advertiser's intent—produces harmful outcomes for users.

Through a detailed study of advertising, this dissertation contributes methods to measure the black box of personalization, and provides insights into its adverse effects.

To my daughter, Zeenat. May your paths be easier than mine.

Contents

Acknowledgments

When I started graduate school, I never imagined myself to do as well as I did. I am the first in my family to get a college education, much less complete a body of doctoral work. In grappling with the unlikeliness of what I have accomplished, I have found no explanation other than fate. I've been blessed on numerous occasions by the gracious yet firm hand of fortune, pushing me in the right direction and blocking paths that weren't meant for me, and for that I can only say *Alhamdulillah*.

Fortune usually manifests itself through the people that we run into, and I've been very fortunate to be surrounded by exceptional people at Northeastern. Foremost, I owe a debt of gratitude to my advisor, Alan Mislove, whose infectious enthusiasm has carried me through five years of long, hard work in graduate school. I have always walked out of my meetings with Alan with more motivation than when I entered the room, and with the most useful feedback on ongoing work. He has asked me hundreds of questions that made this work better, suggested dozens of experiment designs that worked out wonderfully, and refined so many sentences that I wrote. But perhaps most importantly, Alan has repeatedly told me that I was doing well, and that I didn't need to worry as much as I did. I've also had the privilege of having Piotr Sapieżyński as my closest collaborator and an informal advisor throughout graduate school. Nearly every analysis in this dissertation, and every talk I've given in the past years, has benefited from Piotr's feedback and polishing. Beyond his feedback as a mentor, as a friend, Piotr has repeatedly told me that I should rightfully be proud of the work I've done, and not to undersell it. These lessons in self-confidence were the most valuable thing my mentors could have done for me at my career stage, and I will always be grateful to Alan and Piotr for them.

I am thankful to my committee for helping me polish this thesis. Thanks to Christo Wilson for being the resident expert on algorithm auditing, and for his valuable advice on work and career whenever I asked for it. Thanks to Saiph Savage for her assuring encouragement, and for her care and attention towards this thesis. Thanks to Elissa M. Redmiles for being a collaborator and a friend, and for all her patience and guidance as the last study of this thesis took shape. Many thanks also to all my collaborators—Angelica Goetzen, Aleksandra Korolova, Aaron Rieke and Miranda Bogen—for their contributions to this work.

I've been fortunate to have friends, both new and old, whose company made these last few years easier. The sixth floor of the Interdisciplinary Science & Engineering Complex has been a wonderful place to find camaraderie over espresso in the middle of the work day. Thanks to Shuwen, Giorgio, Domien, Harshad, Amogh, Marinos, Lydia, Norbert, Chaima, Lulu, Jeffrey, Johanna, Aziz, Eysa, Cliff, Giri, Fan, Ahmad, Girik, Rashika, Jagat, and so many others at Khoury College, for their companionship. Thanks to Jack for baking pies that brought everyone at ISEC together. I am grateful for my Pakistani friends in grad school at Northeastern and elsewhere, with whom I've shared the predicament of trying to be passionate about work while navigating precarious visas and borderline poverty—thanks to Talha, Waqar, Batool and Usman for their friendship and solidarity.

I'm grateful to my family for their support through the hard times. Thanks to my parents, Nadia and Imran, for everything they've done for me. Thanks to my brothers, Umer and Shehroz, for their brotherhood and love. Finally, I am perpetually indebted to my wife Tazeen for her sacrifices, her proofreading, her feedback on my practice talks, her wisdom and foresight, and for building a home and family with me. It brings me no joy to admit this, but you are somehow always right.

Chapter 1

Introduction

Large online platforms today sift through billions of pieces of content, serving it to millions of online users every minute. On YouTube, the world's leading video sharing platform, 500 hours worth of videos are uploaded every minute [164]. Similarly, Facebook, the largest social media platform, sees 1.7 million pieces of new content every minute [164]. To filter through the deluge of content, and to keep audiences engaged, platforms frequently rely on *personalization* to surface content that each user would find most relevant. For example, YouTube personalizes videos for its users in the "Up Next" panel—which is designed to recommend additional videos that the user might be interested in, based on what they are currently watching [105]. On Facebook, users are presented with a "News Feed" of content, which is ranked to specifically surface the content that the user would find relevant, according to the platform's metrics [125].

Beyond these massive platforms as well, personalization has essentially become an integral part of consuming the modern Internet. Search results, job recommendations, and targeted advertisements are all personalized, and are typically ranked by some notion of relevance of content to the user. This can often result in each user having a unique, catered experience, possibly vastly different from what other users are experiencing. Personalization can also be helpful in finding content that the users would actively engage with [50, 54], avoid content that is overly promotional [125] or low quality [105], and even improve diversity of users' experiences [262].

However, the ubiquity of personalization also means it mediates our access to sensitive domains of content. The same algorithms that determine whether we are interested in a pair of boots can decide if we should see a job opportunity, which political messages we should be seeing, or whether we would engage with clickbait. Therefore, alongside the utility of personalization (formally called *recommender systems* [26]), there are increasing concerns about the adverse user effects of these systems. For example, in video recommendation, YouTube's system has been been critiqued by the public [96], media [233], and scholars [233] for causing amplification of videos containing radical content. Research has found on multiple occasions that while a majority of users do not experience radical content, personalization does cater to the small minority of far-right users who engage with extremist videos [118, 121, 204]. In advertising—the topic of this dissertation—personalization is believed to be responsible for limiting access to high-paying job opportunities for women [52, 147], and exposing vulnerable users to ads that they find distressing [97, 100].

The adverse outcomes of personalization closely relate to broader issues of algorithmic bias and fairness in machine learning systems [22, 67, 111], where individuals or certain groups receive worse outcomes due to correlations learnt from data [30, 32, 133]. This is partly because recommending content, at its core, is about drawing inferences about users' preferences, which opens up the potential for inferring stereotypes and undesirable biases [55, 101, 111, 130]. Optimizing for a user's immediate engagement might not result in their long-term satisfaction or well-being [179, 193]; it could also be antithetical to social values and legal constraints in specific domains of content. Within such sensitive content domains, personalization could amplify pre-existing biases by unfairly distributing content across users. This unfairness could manifest either by excluding users from valuable information, or by including problematic content into their experience [174].

The harms of personalization. But can the unfair outcomes of personalization be harmful for users, and if so, how? Within advertising, consider how prior work has demonstrated that women are shown fewer high paying job ads than men on Google [52, 53]; and fewer STEM-related job ads on Facebook [147], despite clicking on them more often. This is potentially the result of associations the platforms have learnt in their pursuit to optimize for relevance, but it has the inadvertent effect of excluding women from economic opportunities. Not only is such discrimination illegal in the United States [231], it harms individual users by limiting their career prospects.

Prior work also shows that personalization can optimize for individual vulnerabilities, such as optimizing weight loss ads for someone with an eating disorder [100], therefore exposing users to content that is either triggering, or facilitates unhealthy behaviors. Repeated exposure to such problematic content can lead to feelings of distress and a loss of autonomy, and has been likened to "slow violence" in recent sociotechnical work [100, 255]. Again, these effects are a byproduct of the system's design to reach users who find such content relevant, even when that comes at the expense of harming their mental health.

These examples from prior work illustrate that the harms of personalization are not abstract, but are experienced by groups of people, and are quite tangible. When it is used to find the relevant audience for a local event or a sale, personalization carries no adverse user effects; but I argue that the same apparatus can end up harming users in certain domains of content.

As the harms of personalization become apparent, first, we need methods to better measure its adverse effects on users. Such methods can be used by engineers as well as policymakers to concretely understand the harms that users are experiencing. Second, we need to apply these methods to to specific content domains to understand how personalization could harm users, either by excluding them from valuable information, or by including problematic content into their experience [174].

The challenges of measuring harms. However, some key challenges to measuring personalization are that these systems are opaque [179], and online platforms are often resistant to being studied. Therefore, researchers wishing to study these systems have to do so without any privileged access, and without the platform's cooperation. To measure the harms of personalization in this work, I rely on *algorithm auditing*—the process of testing whether an algorithmic system is biased or discriminatory without direct access to its internals [173].

An audit probes an algorithmic system with varied inputs and observes the outputs to identify systematic biased behavior against a subgroup of users or subjects. In the past, algorithm auditing has been used to measure racial disparities in facial recognition algorithms [30], find racial disparities in criminal risk assessment [157], investigate partisan bias in search systems [206], find problematic religious associations in language models [5], among other things. Recent work in auditing, including the work in this dissertation, extends auditing to a sociotechnical setting as well, where users' experience with the system are also measured through surveys [146]. One of the central challenges of this work is measuring the influence of personalization with only black box, external, access to these algorithms. Motivated by the emergent harms of personalization, and the need to measure them, the thesis of this work is:

> *Personalization in online platforms can have adverse effects for users in sensitive content domains, and algorithm auditing methods can be used to measure these effects without privileged access to the platforms.*

Throughout this dissertation, I build novel methods to measure a diverse set of harms caused by personalization—such as discrimination in access to opportunities, political polarization, and overexposure to problematic content. I then employ these methods to conduct measurements for specific domains of content where personalization could be harmful—and where consumer protection might be necessary—such as jobs, housing, political information, and scams. I measure these harms using one of the Internet's largest and richest personalization systems as a case study, Facebook's targeted advertising platform.

1.1 Why study advertising?

Targeted advertising shares many similarities with other personalization systems, as well as differences that enable opportunities for measurement. Similar to other recommender systems, advertising relies on the relevance of each piece of content (ad) to the user. Platforms regularly claim their aim is to show users "relevant" ads: for example, Facebook states "we try to show people the ads that are most pertinent to them" [45]. Intuitively, the goal is to show ads that particular users are likely to engage with, even in cases where the advertiser does not know a priori which users are most receptive to their message. To accomplish this, platforms such as Facebook and Google build extensive user interest profiles and track ad performance to understand how different users interact with different ads. This historical data is then used to steer future ads towards those users who are most likely to be interested in them. However, in doing so, the platforms may inadvertently cause ads to deliver primarily to a skewed or stereotyped set of users, an outcome that the advertiser may not have intended or be aware of, and one that could be harmful for users in certain domains of content.

In contrast to other recommender systems, advertising has three key differences that enable novel approaches for measurement. *First*, advertising systems allows specifying a target audience for each ad. Ad targeting and ad delivery are distinct steps of the advertising process—where the advertiser has more control in the former, but the platform's personalization and optimization determine the ad's audience in the latter phase. While the ad delivery depends on multiple mechanisms such as the advertiser's budget, market competition, and ad quality, personalization plays a central role regardless. For instance, in Facebook's system, the estimated action rate, defined as "an estimate of whether a particular person engages with or converts from a particular ad", is one of the three main components that determines the outcome of an ad auction, alongside bid and ad quality [87]. This distinction between targeting and delivery can be exploited from a methodological standpoint to isolate the impact of personalization alone, as later chapters in this dissertation show.

Second, due to the advertiser-facing nature of the system, advertising platforms also provide fine-grained insights on each ad's performance (e.g. reach, clicks etc.) These insights can be broken down along user attributes such as gender, geography, age etc., and be leveraged as a measurement instrument to understand the ad delivery phase. *Third*, ad platforms—often in response to users' and regulators' concerns—offer transparency mechanisms that help users understand why an ad was shown to them, and what data the platform has inferred about them. Facebook, for example, implements a "Why am I seeing this?" button next to each ad [83], which users can use to inspect the targeting that the advertiser originally specified. Similarly, in another transparency mechanism, Facebook (as well as Google) users are able to view and edit the *interests* inferred about them through the Ad Preference page [41]. These ad transparency tools, in addition to being helpful for users, provide user-facing information about the ad targeting phase, which can be leveraged in an observational setting to disentangle the effect of personalization from targeting in real-world ads.

Taken together, the methodological affordances of advertising provide a vantage point to study the harms of personalization in general, and a better understanding of how these effects could permeate to other systems. Additionally, unlike relying on automated browsers [116] and artificial user profiles [118] to measure personalization, the ability to specify a target audience

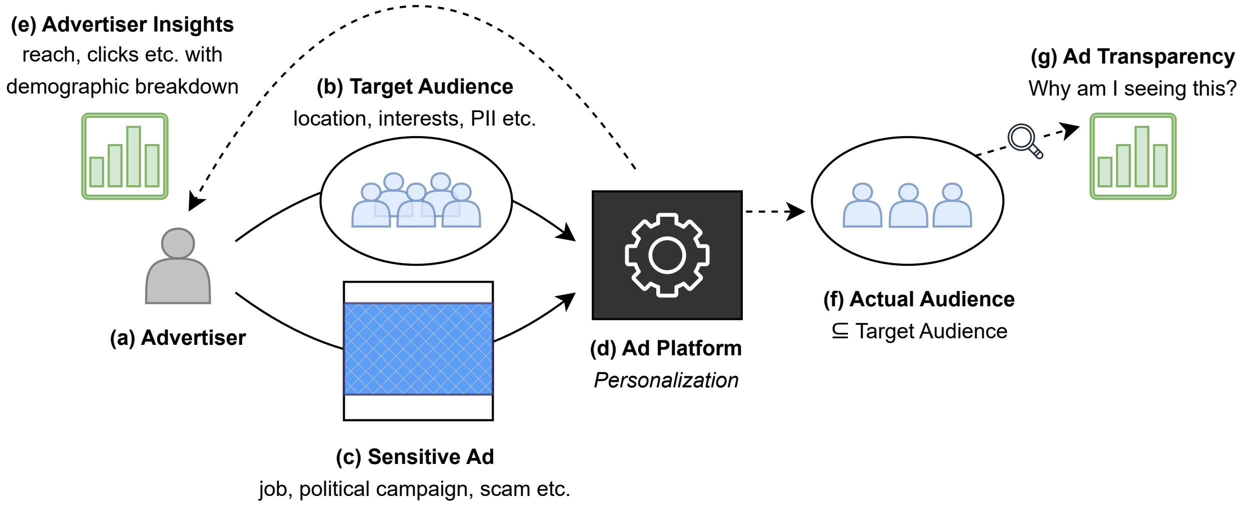


Figure 1.1: High level infrastructure of a targeted advertising platform. (a) The advertiser specifies a (b) target audience for (c) an ad; (d) the ad platform delivers it to the (f) actual audience it judges most relevant based on its internal personalization mechanisms. Advertisers are shown (e) advertiser insights about ad delivery; users are shown (g) ad transparency explanations for why they were shown an ad. Solid lines show inputs into the system that I leverage, dashed arrows indicate outputs from the platform, green charts show points in the pipeline where I collect data in this dissertation.

of real users and to inspect the ad targeting of real-world ads provides broad ecological validity. While I focus on Facebook in this dissertation, the kinds of advertising studied here, and the methods in this dissertation, are easily extensible to other advertising systems. Figure 1.1 shows a high level infrastructure of modern advertising platforms, and highlights the mechanisms that we leverage to conduct our measurements. Section 2.4 describes these components in detail.

Facebook's advertising system, in particular, is the second-largest online advertising platform on the Internet in terms of market share and number of users [124]. It is also a market leader in introducing new advertising methods, such as *custom audiences* and *lookalike audiences* [85], which are subsequently adopted by other platforms, such as Twitter (rebranded now to X), LinkedIn or TikTok [154, 229, 234]. Moreover, owing to the complete profiles users create on Facebook, and the cross-web presence of Facebook's tracking *pixels* [172], the platform has some of the richest user profiles that advertisers can use. Therefore, findings on Facebook are indicative of how a sophisticated personalization platform affects users, and may also be applicable to other ad platforms.

The domains I study in this dissertation include employment, housing, political ads, as well ads perceived as problematic by users themselves. For each of these domains, the nature of user harm differs. Employment and housing are legally protected, and limiting access to these opportunities based on gender and race can result in charges of discrimination in the United States. Political ads, on the other hand, have fewer legal protections, but personalizing such content can reduce the diversity of political discourse and give advertising platforms undue power in the democratic process. Finally, users' own description of problematic ads uncovers content that they themselves find untrustworthy—such as deceptive and scam advertising, clickbait, and ads for individual vulnerabilities such as addictions. Being overexposed to these problematic ads, especially for vulnerable users, can affect users psychologically, or lead to financial and personal information loss [123].

1.2 Thesis Contributions

The highest level contributions of this work include **(a)** experimental and observational methods to measure the effect of personalization in targeted advertising; **(b)** application of these methods to domains where personalization could harm users, resulting in **(c)** concrete measurements of the harms of personalization in Facebook's advertising system. Below, I list the contributions for each content domain in more detail.

1.2.1 Gender and racial biases in job and housing ads

I conduct the first large-scale measurement of algorithmic biases that affect job and housing ads on Facebook, specifically how these ads can be skewed along gender and race due to the platform's optimizations. I design novel methods to isolate the effect of each component of an ad (image, text, URL etc.) on the eventual delivery. I then run dozens of ad campaigns as controlled experiments, and use Facebook's advertiser insights to measure skews that are introduced along gender and race of users. My results show drastic effects on job ads due to personalization: for example, real ads for janitor jobs skew on average to an audience that is nearly 75% female, and nearly 60% Black, compared to ads for jobs in the lumber industry, which skew to an audience over 90% male and nearly 75% white. These effects arise despite targeting the same candidate set of users, and specifying the same budget. I argue that these differences are problematic, as they could enable potentially illegal discrimination, since housing, credit and employment opportunities are protected under U.S. law [1–3, 231]

1.2.2 Ideological biases in political ads

I then measure the impact of personalization on political advertising. I extend the methodology for measuring race and gender skews, to measure skews along political leaning of the targeted users. I run real ads for candidates in the U.S. 2020 Presidential elections (Donald Trump and Bernie Sanders), targeted to voters in several locations within the U.S., to measure the impact of personalization in skewing, or rather limiting, the reach of these ads. My experiments show how ads trying to reach users that do not align with the candidate's political leaning, both reach fewer users than ads for the aligned candidate, and have to pay a price premium of up to 300% more than the aligned candidate. It is important to emphasize that these differences are due to relevance judgments made by the platform's personalization (ad delivery) algorithms, since my methodology is designed to control for all advertiser-dependent variables in experiments. My results provide a precise quantification of the influence personalization can have in delivering political ads. I argue that skews along political leaning are harmful to political advertising online, as it yields significant power to opaque personalization systems in making decisions about important public political messages.

1.2.3 Disparate user experiences with problematic advertising

With an understanding of personalization's role in Facebook's system overall, I turn to investigate variances in *individual experiences*, as seen by users themselves. I create a browser extension capable of collecting ads and targeting data through Facebook's transparency tools, and recruit a panel of active Facebook users to contribute their data for several months. The goal of this effort is to longitudinally observe the real-world experiences of multiple users, and to investigate if some users have a disparately large exposure to problematic ad types. Through participant surveys, I find that clickbait, scams, and sensitive content types such as weight-loss, finance and alcohol are considered *problematic* by users. Modeling the distribution of problematic content over users shows that they are disproportionately shown to a small subset of the panel. I also find that older participants are more likely to see deceptive and clickbait ads, and Black participants are also more likely to see clickbait ads. Inspecting the ad targeting data shows that while advertisers of clickbait target older participants actively, even when they do not, personalization skews these ads towards them. Therefore, I find evidence that personalization is partially responsible for overexposing vulnerable users to problematic content. Through this measurement, I also construct a dataset of over 85,000 unique ads and their targeting data that would support future research in this area.

1.2.4 Publications

Parts of this dissertation have been published as the following full-length conference publications. An asterisk denotes equal authorship.

1. **Muhammad Ali**, Angelica Goetzen, Alan Mislove, Elissa M. Redmiles, Piotr Sapiezynski. "Problematic Advertising and its Disparate Exposure on Facebook". In *Proceedings of the 32nd USENIX Security Symposium*, Anaheim, CA, August 2023.

2. **Muhammad Ali***, Piotr Sapiezynski*, Aleksandra Korolova, Alan Mislove, Aaron Rieke. "Ad Delivery Algorithms: The Hidden Arbiters of Political Messaging". In *Proceedings of the 14th ACM Conference on Web Search and Data Mining (WSDM)*, Virtual, March 2021.

3. **Muhammad Ali***, Piotr Sapiezynski*, Miranda Bogen, Aleksandra Korolova, Alan Mislove, Aaron Rieke. "Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes". In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*, Austin, TX, November 2019.

Work from this dissertation also appears in the following peer reviewed short papers:

1. **Muhammad Ali**, Angelica Goetzen, Alan Mislove, Elissa M. Redmiles, Piotr Sapiezynski. "All Things Unequal: Measuring Disparity of Potentially Harmful Ads on Facebook".

In *Proceedings of the 6th Workshop on Technology and Consumer Protection (ConPro)*, San Francisco, CA, May 2022.

2. **Muhammad Ali**. "Measuring and Mitigating Bias and Harm in Personalized Advertising". Extended Abstract in *Proceedings of the 15th ACM Conference on Recommender Systems (RecSys)*, Amsterdam, The Netherlands, September 2021.

1.2.5 Ethics

All of my experiments and data collection was conducted with careful consideration of ethics about the users and online platforms involved. All of the studies in this dissertation obtained Institutional Review Board (IRB) review from Northeastern University. Two studies (IRB *#18-11-13* and *#19-05-01*) were marked as "Exempt" by the IRB due to not involving human subjects directly and only computing aggregate statistics. One study in this dissertation which involved surveys and human participants' data (IRB *#20-10-11*) was reviewed and formally approved by the IRB.

Beyond a formal approval, I also make sure to minimize harm to users and platforms in our methodology. When running experiments (ads), I never target ads in a discriminatory manner, and never expose users to harmful content. All ads for job, housing and political information actually link to websites that are truthful about the advertised content. When collecting user data, we give informed consent and strictly limit access to the data to just the research team. Within Chapters 4, 5 and 6, I provide more details about the specific ethics considerations for each study on a more fine-grained level.

Positionality Statement. I am a scholar of computer science, which might consciously or subconsciously impact my interpretation of the social science themes in this dissertation, such as discrimination and technological harms. Moreover, I am a South Asian (Pakistani) man, which likely affects my understanding of gender, race, ethnicity, and other demographic attributes that are studied here. I have tried, to the best of my ability, to be respectful and cognizant of the users' identities I study. However, I acknowledge that my gender and ethnic positionality may cause me to not fully understand some nuances.

1.3 Outline

The remainder of this thesis is organized as follows. I first provide necessary background about personalization in general and Facebook's advertising system in particular in Chapter 2. I also describe the technical tools that I use within Facebook's platform to conduct my measurements. In Chapter 3, I summarize prior work that studies discrimination and other harms in personalization, and situate this dissertation within the literature.

Chapter 4 describes how I exploit the advertiser insights provided by Facebook to conduct the first algorithm audit of employment and housing advertising on Facebook. I describe the methods and experimental design in detail, and present measurement results gained from these experiments. Chapter 5 describes my work measuring political advertising on Facebook, specifically how personalization can produce skews along political leaning, which can lead to differential pricing for political candidates. I describe the novel methods I build to isolate the role of personalization, and present results from a measurement study done during the 2020 U.S. presidential election season. In Chapter 6, I present my work on studying individual user experiences with problematic advertising. I describe the methods I use to recruit participants and survey them about their ads to find what they find problematic. I then present measurements on the distribution of problematic content across participants, and discuss the role of personalization in exposing users to such ads. Each chapter concludes with a discussion of the harms uncovered, and recommendations from a technical and policy standpoint to mitigate those issues.

Chapter 7 concludes this dissertation by providing a high level discussion of all the results in this body of work. I also discuss future research directions resulting from the work presented here, and note limitations and opportunities that future work can address.

Since large parts of the work presented here was done with collaborators, in the remainder of this document, I use the collective pronoun *we* to discuss the work.

Chapter 2

Background and Definitions

This chapter describes necessary background on personalization and the advertising ecosystem. We also describe tools within Facebook's advertising system that are used throughout this dissertation.

2.1 Personalization and Recommender Systems

Personalization is implemented in online platforms through algorithms known as *recommender systems*. The core problem in recommender systems, a rich subdiscipline of machine learning, is to recommend *items* to *users* based on previous items they've liked, or users similar to them have liked. The two broad ways of solving this problem is either through *content filtering* or *collaborative filtering* [138]. Collaborative filtering (CF), in particular, has been extremely popular because of its general applicability as a modeling approach. CF approaches are powerful in that they rely on past data, and therefore automatically capture user preferences and item characteristics based on the principle that similar people like similar items.

Especially common are embedding or matrix factorization techniques for collaborative filtering [111]. In this setting, a recommender system learns a vector representation for each item and user; then users are recommended the items with feature vectors most similar to their own (e.g., the items that maximize the inner products between the representations) [111, 138, 159, 211]. Beyond matrix factorization approaches, contrastive learning and other representation learning approaches (e.g. "two tower" neural networks) are also used [156, 181, 248, 263]. The commonality behind many of these CF approaches, however, remains that they learn vector representations that maximize a notion of relevance between users and items.

However, large personalization systems (such as Facebook's advertising system) are often a combination of several smaller recommender systems [179]. These smaller systems often deal with different steps of the pipeline, one model might be generating *candidates* for recommendation, while another might be predicting the relevance of each candidate to the user to constuct a ranking [50, 125]. Within advertising, auction models are also involved in the mix [180]. In this dissertation, we refer not to an individual recommender system, but the combination of all systems used in a pipeline as the larger *personalization* system. While our focus in this dissertation is advertising, due to the algorithmic approaches shared across recommender systems, these methods can be used to audit and measure other personalization systems as well.

2.2 Facebook

Facebook (`https://facebook.com`), founded in 2004, is a social media website and app that allows staying connected with friends through text posts, images, videos, and messaging. As of 2023, it is the largest social media platform on the Internet, with over 2.9 billion active users [188], which means approximately 35% of the world's population uses the platform. Facebook's parent company, Meta, also owns the world's most popular photo sharing app, Instagram. Meta's core revenue stream is through advertising across its family of apps. Ads are presented as posts within a unified *news feed* of content, which includes posts from friends, pages followed by the user, as well as advertisers (marked with "Sponsored" at the top). Advertising on Facebook is particularly successful because of its access to rich content posted by users, which in turn is used to infer attributes and behaviors about the users, which then are provided to advertisers to target ads [78].

2.3 Online Behavioral Advertising

Online advertising, and in particular, *targeted* advertising, supports much of the modern Internet's business model. It is now an ecosystem with aggregate yearly revenues close to $100 billion [57]. Due to the volumes of user data collected by modern social media sites (Facebook, Twitter), search engines (Google, Bing), and content sharing sites (YouTube, TikTok), these platforms are able to build fine-grained profiles on users' interests and behaviors. Platforms then regularly rely on inferring users' interests, and providing advertisers with an interface for these interests, to enable precise targeting of ads based on users' behaviors. This infrastructure often acts as the economic backbone of modern web platforms, allowing them to operate for free. When asked in his 2018 testimony before Congress how Facebook remains free, Mark Zuckerberg, the CEO, famously replied, "Senator, we run ads." [214]

The *contextual* advertising ecosystem, such as that enabled by Google, is a complex set of interactions between ad publishers, ad networks, and ad exchanges, with an ever-growing set of entities involved at each step allowing advertisers to reach much of the web [23]. In contrast, social media platforms such as Facebook and Twitter run advertising platforms that primarily serve a single site, presenting ads in a consolidated feed of content. These platforms serve both as a publisher (as they have ad inventory on their site), and the ad network (as they accept ads from advertisers). In this thesis, we focus on social media advertising, specifically Facebook, but our results may also be applicable to more general contextual advertising on the web, as it also relies on methods to determine relevance of ads to users.

2.4 Facebook's Advertising System

Facebook's advertising platform, in particular, is one of the most data-rich and sophisticated platforms on the internet. Its billions of users, who often sign up with complete personal profiles, alongside the cross-web presence of Facebook's tracking infrastructure [172] make it a market leader within the social media advertising space and second-largest advertising platform on the Internet overall [124]. This section details the specific design of Facebook's advertising system.

2.4.1 Ad Creation

Ad creation refers to the process by which an advertiser submits their ad to the Facebook ad platform. During this stage, advertisers provide the contents of the ad—including the images, videos, text, headline, and destination link collectively called the *ad creative*)—and specify the *target audience* of users on the platform to whom they wish the ad to be shown (see §2.4.2). Each ad placed on Facebook must be linked to a *Page*; advertisers are allowed to have multiple Pages and run ads for any of them. Advertisers often run many ads that are related; collectively, these are called an *ad campaign*.

Before submitting an ad campaign, advertisers also specify an *objective*, i.e., what they want to achieve with the campaign and the *ad budget* they are willing to spend to achieve that objective. For ads linking to an external website, the advertiser can provide a *traffic destination* to send the user to if they click (e.g., a Facebook Page or an external URL); if the advertiser provides a traffic destination, the ad will include a brief description (auto-generated from the HTML `meta` data) about this destination. Examples showing all of these elements are presented in Figure 2.1. After creation, the ads then enter a review process to be approved to run on the platform, where automated systems and moderators ensure it follows the platform's content and ad targeting policies [250]. Below, we describe each of these early steps of creating an ad campaign in detail.

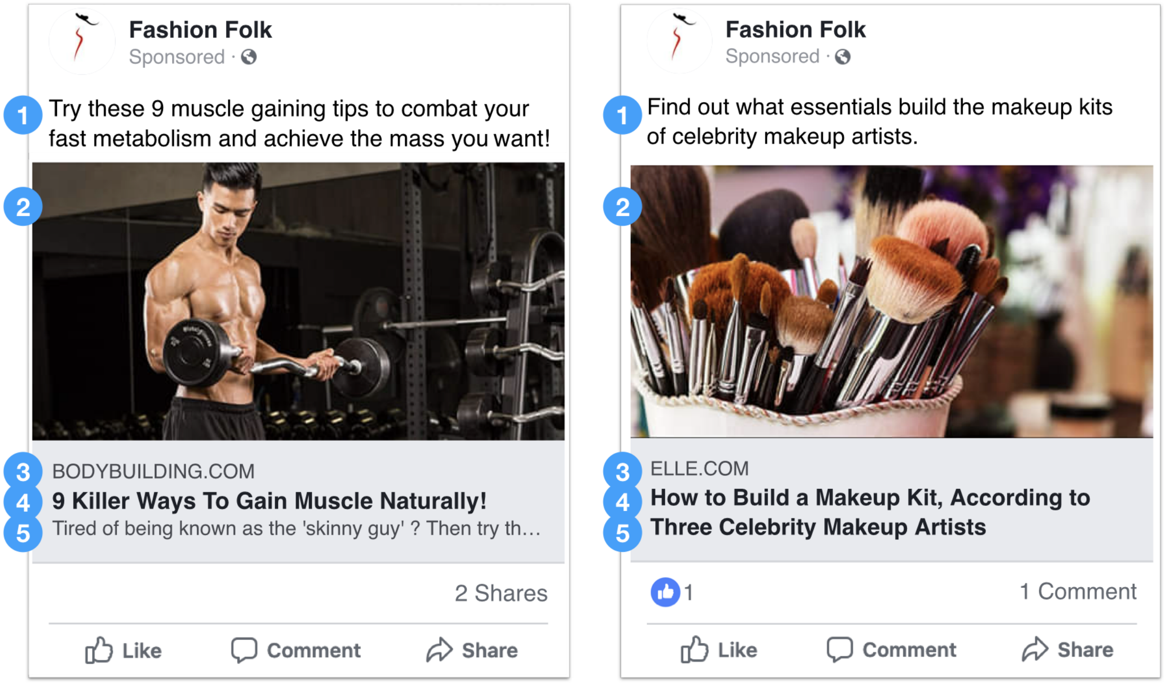
Objective. When creating ad campaigns, Facebook provides advertisers with a number of *objectives* to choose from when placing an ad [38]. Common objectives include "Reach" (showing the ad to as many users as possible), "Traffic" (showing the ad to the users most likely to click), and "App Installs" (showing the ad to the users most likely to install the advertiser's app).

Within each objective, advertisers must also specify an *optimization event*, indicating how Facebook should achieve their objective. For the "Reach" objective, the available optimizations are "Reach" (the default, showing ads to as many *users* as possible) and "Impressions" (showing ads as many *times* as possible). For the "Traffic" objective, the available optimizations are "Link Clicks" (the default, showing ads to the users most likely to click), "Landing Page Views" (showing ads to the users most likely to click *and* visit the destination page), "Daily Unique Reach" (showing ads to users most likely to click, at most per day), and "Impressions" (showing ads to the users most likely to click, but maximizing the number of impressions). Similarly "Conversion" optimizes for sales generated by clicking the ad. For each objective, the advertiser bids on the objective itself, e.g., for "Reach", the advertiser would bid on ad impressions. The bid can take multiple forms, and includes the start and end time of the ad campaign and either a lifetime or a daily budget cap. With these budget caps, Facebook places bids in ad auctions on the advertisers' behalf.

Bidding, budget and billing. When creating an ad, the advertiser also must tell Facebook their bid, which takes the form of an *ad budget*. These budgets are either a daily or lifetime budget for the ad, allowing Facebook to spend the advertiser's money between and within auctions according to an algorithm that is not publicly known. Facebook only says: "Facebook will aim to spend your entire budget and get the most 1,000 impressions using the lowest cost bid strategy" for the "Reach" campaign-level objective, and "Facebook will aim to spend your entire budget and get the most link clicks using the lowest cost bid strategy" for the "Traffic" campaign-level objective [87]. An optional "Bid Control" (maximum bid in each auction) and "Cost control" (the average cost per link click) are also available for the "Reach" and "Traffic" campaign-level objectives, respectively. When using the interface, Facebook also provides advertisers with an "Estimated Daily Reach", which Facebook defines as the estimated number of users who would be exposed to an ad given the targeting criteria and budget; this feature helps advertisers with an understanding of how far their budget will go for the selected audience. Although the bid and cost control options are offered, without the knowledge of the auction, they are difficult to set effectively, and therefore, it is natural to only specify a budget and rely on Facebook to do the bidding. Advertisers can optionally specify a per-bid cap as well, which will limit the amount Facebook would bid on their behalf for a single optimization event.

How the advertiser is actually charged for their ad campaign depends on the objective. For "Reach", the advertiser is charged per impression; for "Traffic" the advertiser is also charged per impression unless the optimization is "Link Clicks" (in which case the advertiser can choose to be charged *per click* if they wish).

Finally, if the advertiser chooses the "Reach" objective and "Reach" optimization, they are allowed to specify a *frequency cap*, which allows them to set the maximum number of



Fashion Folk
Sponsored ·

Try these 9 muscle gaining tips to combat your fast metabolism and achieve the mass you want!

BODYBUILDING.COM
9 Killer Ways To Gain Muscle Naturally!
Tired of being known as the 'skinny guy' ? Then try th…

2 Shares

Like Comment Share

Fashion Folk
Sponsored ·

Find out what essentials build the makeup kits of celebrity makeup artists.

ELLE.COM
How to Build a Makeup Kit, According to Three Celebrity Makeup Artists

1 1 Comment

Like Comment Share

Figure 2.1: Components of an ad creative on Facebook: (1) the ad text, entered manually by the advertiser, (2) the images and/or videos, (3) the domain, pulled automatically from the HTML `meta` property `og:site_name` of the destination URL, (4) the title, pulled automatically from the HTML `meta` property `og:title` of the destination URL, and (5) the description from `meta` property `og:description` of the destination URL. The title and description can be manually customized by the advertiser if they wish.

impressions a single user would see over a specified number of days. We use this feature later in Chapter 5 to force Facebook to deliver our ads to many users in our audience.

2.4.2 Ad Targeting

Facebook provides a wide variety of targeting (or audience selection) options [14, 15, 103, 222]. In general, these options fall into a small number of classes:

User demographics. At the simplest level, advertisers can target based on users' demographics attributes, such as their age, gender, location, and language. This information is usually self-reported by users at the time of profile creation, and therefore often has a high degree of accuracy. While Facebook's minimum age for signing up is 13, advertisers can only target over the age of 18.

Inferred user attributes. Facebook pioneered, and continues to market ways for advertisers to target its users via *user attributes* [222]. Targetable attributes cover a variety of aspects of users' lives, ranging from interests to online activity and even offline information, often acquired without users' explicit consent or knowledge. User attributes can include sensitive information as well, e.g. in the context of politics, Facebook derives attributes that indicate whether users are "interested in" various political candidates (e.g., Donald Trump, Elizabeth Warren), as well as more general attributes about user behaviors, such as "Likely engagement with US political content (Conservative)" or "Likely engagement with US political content (Liberal)". The exact methodology by which Facebook infers such attributes is not disclosed, but likely involves profile data provided directly to Facebook, data from activity on Facebook (e.g., "Liking" Pages or explicit or implicit patterns of interaction with particular content), inferences based on attributes of a user's friends, and data inferred from users' behavior off of Facebook. These inferences can also be made based on a variety of sources beyond the Facebook website and app, including Facebook Pixel tracking [172], app data sharing [40], third-party data brokers [244], and location data [90]. Recent work has shown that Facebook offers over 1,000 well-defined attributes and hundreds of thousands of free-form attributes [222].

Custom audiences. Facebook also allows advertisers to target users directly using their personally-identifiable information (PII) via a tool called *Custom Audiences* [243]. These PII-based audiences are directly tied to a user's identity rather than being inferred from their behavior. Using Custom Audiences, advertisers can upload up to 15 different kinds of PII to Facebook, ranging from names to email addresses to phone numbers to dates of birth. Often, advertisers will upload files with a mix of PII, such as separate columns for email addresses, phone numbers and zip codes, to enable multiple attributes for matching. Facebook then matches these values against their database in order to build an audience for the advertiser; the advertiser is then allowed to target their ads to just the users who match. Beyond Facebook, other platforms such as Google, Twitter, and Pinterest all provide similar features; these are called *Customer Match* [119], *Tailored Audiences*, and *Customer Lists* [186], respectively.

Lookalike audiences. Facebook also enables finding "similar" users to those who they have previously selected. This can be done via a tool called *Lookalike Audiences*, where advertisers can specify a source custom audience, and ask Facebook to essentially expand this list with users who have similar interests. In building a lookalike audience, the platform "identif[ies] the common qualities of the people in it" and creates a new audience with other people who share those qualities [85] Other platforms also offer similar features, Google and Pinterest, for example, offer *Similar Audiences* [107] and *Actalike Audiences* [196] respectively. by starting with a source Custom Audience they had previously uploaded; Facebook then "identif[ies] the common qualities of the people in it" and creates a new audience with other people who share those qualities [85].

Advertisers can often combine many of these features together, for example, by uploading a list of users' personal information and then using attribute-based targeting to further narrow the audience.

2.4.3 Ad Review

Once the ad creation phase is complete, and before the ad enters the ad delivery phase, it is submitted to Facebook for review. Similarly, most ad platforms have a review process, consisting of a combination of automated and manual review, to prevent abuse or violations of their advertising policies [31, 250]. We observed throughout this dissertation that most of our Facebook ads were approved within 30 minutes, some spent hours in review, and a few were never approved. The criteria and internal mechanisms for approval are not entirely clear, and precise reasons for why certain ads are rejected are not given. In this work, we only report on experiments where all necessary ads were approved before their scheduled start time.

Once an ad passes review, it enters the *ad delivery* phase, where it participates in auctions for user impressions that Facebook (in its personalization mechanisms) considers relevant for the ad.

2.4.4 Ad Delivery

Ad delivery refers to the process by which the Facebook ad platform selects which ads get shown to which users. For every opportunity to show a user an ad (e.g., an *ad slot* is available as the user is browsing the service), the ad platform will run an *ad auction* to determine, from among all of the ads that include the current user in the audience, which ad should be shown.

In practice, however, the ad delivery process is somewhat more complicated. *First*, the platforms try to avoid showing ads from the same advertiser repeatedly in quick succession to the same user; thus, the platforms will sometimes disregard bids for recent winners of the same user. *Second*, the platforms often wish to show users relevant ads; thus, rather than relying solely on the bid to determine the winner of the auction, the platform may incorporate a relevance score into consideration, occasionally allowing ads with lower bids but more relevance to win over those with higher bids. *Third*, the platforms may wish to evenly spread the advertiser budget over their specified time period, rather than use it all at once, which introduces additional complexities as to which ads should be considered for particular auctions. The exact mechanisms by which these issues are addressed are not well-described or documented by the platforms. However, we can gain an understanding about the role of personalization in this phase through a better understanding of the auction process.

Facebook's ad auction. Ad platforms routinely use *ad auctions* to select which ads to show to users. Historically, this auction took only the advertiser's bid price into account [64]; modern platforms like Facebook, however also consider other features such as the overall performance of the ad and the platform's estimate of how relevant the ad is to the browsing user [87, 216]. Facebook says as much in its documentation for advertisers [87]:

> [W]e subsidize relevant ads in auctions, so more relevant ads often cost less and see more results. In other words, an ad that's relevant to a person could win an auction against ads with higher bids.

Facebook explains that it measures *relevance* as a composite of *estimated action rates* ("[an] estimate of whether a particular person engages with or converts from a particular ad") and *ad quality* ("[a] measure of the quality of an ad as determined from many sources including feedback from people viewing or hiding the ad") [87].

Facebook defines "Estimated action rates" as "how well an ad performs", meaning whether or not *users in general* are engaging with the ad [37]. They define "Ad quality and relevance" as "how interesting or useful we think a given user is going to find a given ad", meaning how much *a particular user* is likely to be interested in the ad [37].

Thus, it is clear that Facebook attempts to identify the users within an advertiser's selected audience who they believe would find the ad most useful (i.e., those who are most likely to result in an optimization event) and shows the ad preferentially to those users. Facebook says exactly as such in their documentation [36]:

> During ad set creation, you chose a target audience ... and an optimization event ... We show your ad to people in that target audience who are likely to get you that optimization event

Facebook provides advertisers with an overview of how well-matched it believes an ad is with the target audience using a metric called *relevance score*, which ranges between 1 and 10. Facebook says [45]:

> Relevance score is calculated based on the positive and negative feedback we expect an ad to receive from its target audience.

Facebook goes on to say [45]:

> Put simply, the higher an ad's relevance score, the less it will cost to be delivered. This is because our ad delivery system is designed to show the right content to the right people, and a high relevance score is seen by the system as a positive signal.

In short, the personalization algorithms that Facebook employs play a significant role in determining which users see which ads. Essentially, the platform makes its own judgment about which ads are likely to be "relevant" to particular users, its own judgement of how to bid on an advertiser's behalf and distribute the specified budget among auctions, and possibly other considerations connected to its business interests. This role—in the context of sensitive domains of ads—is the key phenomenon that this dissertation seeks to explore in its experiments.

2.4.5 Statistics and Reporting

While the ad campaign is active, Facebook provides advertisers with a feature-rich interface [79] as well as a dedicated API [44] for both launching ads and monitoring those ads as they are in ad delivery. Both the interface and the API give semi-live updates on delivery, showing the number of impressions and optimization events as the ad is running. In particular, Facebook reports the *impressions* (the number of times the ad was shown), the *reach* (the number of unique users who saw the ad), the *clicks* (the number of times users clicked on the ad), and the *spend* (how much money was spent), among other performance insights. Advertisers can also request this data be broken down along a number of different dimensions, including age, gender, and location (such as city or Designated Market Area [182]). Notably, the interface and API *do not* provide a breakdown of ad delivery along racial lines or political leaning. Thus, analyzing delivery along racial lines or political leaning—which is the subject of Chapter 4 and Chapter 5 respectively—necessitates development of a separate methodology that we describe in the respective chapters.

2.4.6 Platform Policies

To maintain the health of its platform, Facebook implements policies on both the content that is allowed, as well as the targeting of ads.

Content policies. Several types of sensitive content have varying level of restrictions on Facebook's platform. For example, ads that are deceptive or misleading, sexually suggestive, contain misinformation, or vaccine discouragement are prohibited [168]. Other ads such as those for financial products, over-the-counter or prescription drugs, alcohol and dating have specific content-specific restrictions, and require additional verifications or targeting restrictions in order to be approved. Ads for alcohol, for instance, cannot be targeted to users under 21 in the United States; similar policies are enforced for other countries as well. However, despite the breadth of content policies, there can still be gaps in enforcement, partially due to the changing landscape of problematic content (e.g. the surge of vaccine misinformation content during COVID-19), and also due to challenges in automated moderation [149].

Anti-discrimination rules. In response to issues of potential discrimination in online advertising reported by researchers and journalists [16], Facebook currently has several policies in place to avoid discrimination for certain types of ads. Facebook has also built tools to automatically detect ads offering housing, employment, and credit, and pledged to prevent the use of certain targeting categories with those ads. [128]. Additionally, Facebook relies on advertisers to self-certify [46] that they are not in violation of Facebook's advertising policy prohibitions against discriminatory practices [80]. More recently, in order to settle multiple lawsuits stemming from these reports, Facebook no longer allows age, gender, or ZIP code-based targeting for housing, employment or credit ads, and blocks other detailed targeting attributes that are "describing or appearing to relate to protected classes" [42, 59, 185]. The latter can be the result of relevance algorithms significantly skewing the the actual audience such that this audience is very different from the eligible audience an advertiser intended to reach. As a result, Facebook in particular has implemented novel systems in response to legislative pressure [167] to minimize variance between eligible and actual audiences, in an effort to ensure fairer delivery of particular ads.

Transparency. Driven by pressure from the public and regulators, Facebook (alongside other ad platforms) has introduced transparency mechanisms to help users understand the data that is collected about them and the content they are shown. One tool that Facebook provides is the *Ad Preferences* page, which allows users to see all the interests and behaviors inferred about them [41]. Through this interface, users can remove annoying or problematic inferences that the personalization system has made to gain more autonomy in their online experience—even though the efficacy of these changes has been questioned in recent case studies [97].

Another mechanism that Facebook provides is allowing users to inspect how a specific ad reached them [14, 83]. Through the "Why am I seeing this?" option, users can see the targeting parameters specified by the advertisers that they match against. However, they are only able to see the attributes that they matched, and not the advertiser's full targeting specification. We exploit this mechanism later in Chapter 6 to isolate the role of personalization by observing real users' ads. For political ads in particular, which have been the subject of many policy debates, Facebook also provides a public *Ad Library* that shows all currently running political ads as well as an archive of all past ads [77].

Chapter 3

Related Work

In this chapter, we detail related work from multiple disciplines of computer science—including fairness in artificial intelligence, human-computer interaction, as well as security and privacy—that this dissertation builds upon.

3.1 Algorithm Auditing for Fairness

Following the ubiquity of algorithms and data-driven software systems in daily life, a growing community has formed around investigating their societal impacts [208], investigating their fairness, accountability and transparency. Questions in this emerging domain often relate to the formal study of algorithmic fairness [62], explainability for algorithmic decisions [145, 205], as well as *auditing* of algorithmic systems [93, 173].

Typically, in algorithm auditing, the algorithms under study are not available to outside auditors for direct examination; thus, most researchers treat them as "black boxes" and observe their reactions to different inputs [173]. Among most notable results, researchers have shown price discrimination in online retail sites [115], gender discrimination in job sites [48, 116], stereotypical gender roles re-enforced by online translation services [27] and image search [133], disparate performance on gender classification for Black women [30], and political partisanship in search results [56, 144, 206]. Although most work has focused exclusively on the algorithms themselves, recently researchers began to point out that auditors should consider the entire socio-technical systems that include the users of those algorithms, an approach referred to as “algorithm-in-the-loop” [109, 210] or “sociotechnical audits” [146].

3.2 Targeted Advertising

Advertising systems have been investigated previously to understand the harms they might cause for users, such as enabling discrimination, privacy leaks, amplifying misinformation, and others.

3.2.1 Discrimination in Advertising

One of the earliest studies identifying the potential for discrimination in advertising came from Sweeney [225], empirically showing that searching for African-American names was more likely to return ads for background-checking websites, compared to searching for European-sounding names. Datta et al. [52] expanded the result, showing through randomized controlled experiments, that setting a participant’s gender to female led to seeing fewer ads for executive career coaching in Google’s advertising system. Furthermore, Datta et al. also train a classifier to predict which demographic group the ads belong to, and establish through permutation testing that the differences learnt by the classifier were significant — suggesting a causal connection between gender and the career coaching ads [52, 53]. While the opacity and complexity of personalized advertising on Google makes it challenging to pinpoint the exact reason of these differences, prior work has attempted to explain them. On Facebook, Lambrecht et al. ran a series of ads for STEM education and found they were consistently delivered more to men than to women, even though there are more female users on Facebook, and they are known to be more likely to click on ads and generate conversions [147]. Factors such as advertiser targeting or manual curation might be an explanation for these differences, but optimizations done for better personalization (often resulting from user behavior) could also clearly be a contributing factor [53].

The magnitude of bias that personalization itself can cause has been highlighted in the literature recently; studies that investigate the bias in ad targeting attributes (used by advertisers) provide a window into these differences. For Facebook’s advertising, prior work has shown the extent of demographic skew that can exist simply during the targeting of the ad, by using custom audiences [223], or by using location [90, 140]. This leads to potential for discrimination on the advertiser's end, where even if racial targeting is prohibited, they can choose features that have high correlation with race to achieve discriminatory ad delivery [223]. The potential to combine these high correlation targeting attributes to compound their effects has also been documented for major platforms like Facebook, Google, and LinkedIn [241]. This leads to situations where malicious advertisers have access to tools that enable discriminatory ads, and sincere advertisers might be unaware of how biased their audiences are to begin with.

3.2.2 Advertising Transparency

In parallel to the developments in detecting and correcting unfairness, researchers have conducted studies and introduced tools with the aim of increasing transparency and explainability of algorithms and their outcomes. For example, much attention has been dedicated to shedding light on the factors that influence the targeting of a particular ad on the web [68, 150, 151, 192] and on specific services [52, 253].

Focusing on Facebook, Andreou et al. investigated the transparency initiative from Facebook that purportedly tells users why they see particular targeted ads [14]. They found that the provided explanations are incomplete and, at times, misleading. Venkatadri et al. introduced the tool called "TREADS" that attempts to close this gap by providing Facebook users with detailed descriptions of their inferred attributes using the ads themselves as a vehicle [242]. Further, they investigated how data from third-party data brokers is used in Facebook's targeting features and—for the first time—revealed those third-party attributes to the users themselves using TREADS [244]. Similar to other recent work [184], Venkatadri et al. found that the data from third-party data brokers had varying accuracy [244].

3.2.3 Political Advertising

Prior work, conducted mostly in the form of laboratory experiments, indicated high efficiency of written persuasion personalized to the psychological profile and motivation of the recipient [120, 175]. More recently, Matz et al. conducted a large-scale experiment in which they showed that Facebook ads tailored to individual's psychological characteristics yielded higher click-through and conversion rates compared to non-personalized ads and mismatching ads [161]. Matz et al. relied on the mechanism first documented by Kosinski et al.: that personality traits of an individual can be accurately inferred from the content that they "Like" on Facebook [141]. Based on that work, they found interests that most correlated with high and low extraversion, as well as high and low openness, then used the detailed targeting functionality on Facebook ad platform to target people who "liked" those interests. Other researchers, in response, pointed out that the unknown optimization mechanisms employed by Facebook might obfuscate the measurement of effectiveness of these personalized ads [63]. Matz et al. refuted these criticisms [162], but work in this dissertation (Chapter 5) indicates that Facebook does, indeed, further refine even precisely targeted audiences, introducing demographic and political biases in the reached audiences, beyond those intended by the advertiser.

Filter bubbles. The extent, or even the existence, of the *filter bubble* effect has been a point of contention both in academia and in popular media. After the initial reports by Eli Pariser [191] scholars have attempted to measure the phenomenon in services including Google Search [94, 114], Google News [113], and Facebook [21]. However, the observed differences could often be explained by user location differences (in the case of Google Search) or attributed to an individual's choice of friends to follow (in the case of Facebook), rather than stemming from algorithmic personalization of an individual's experience. Furthermore, Bail et al. warn against overexposing users to messaging from politicians they do not support as it appears to increase, rather than decrease their partisanship [19].

This dissertation, specifically Chapter 5, provides further evidence that the filter bubble effect is pronounced in the ads users are exposed to; we show that attempting to "burst" the political advertising filter bubble can prove expensive, especially for smaller advertisers. We are the first to study the filter bubble effect due to ad delivery aspects of political messaging; prior work on filter bubbles for political content focused on possibilities of disparate treatment of organic rather than sponsored content or disparate treatment during other parts of the process, such as during the ad review stage [12, 142].

3.2.4 Problematic Advertising

Communication and psychology literature have long explored how traditional mass media (e.g., print, TV, radio) expose consumers to problematic content. Social science theories such as cultivation theory (how exposure to content may influence people's thoughts and behaviors [198]) and agenda-setting theory (how content can be used to shape and filter a consumers' reality [215]) posit ways in which harmful media can produce negative outcomes for consumers. Empirical observations under these frameworks include how violent media teaches violent behaviors (e.g., [218]); how drug usage in media promotes drug usage in viewers (e.g. [247]); how bigoted media reinforces prejudice (e.g., [29, 246]); and how exposure to idealized body images can lead to body image issues (e.g., [17, 197, 240]).

Within advertising as well, prior work has investigated the potential for advertising to expose users to problematic content. This is particularly important, since ad platforms often self-regulate, and set their own policies to define which advertising content they do or do not allow on their sites [106, 168]. These policies are often updated at the platform's discretion, or in response to the changing landscape of problematic content. For instance, social media platforms expanded their definitions of harmful content following the proliferation of misinformation related to the COVID-19 pandemic in 2020 [20]. But despite policies and detection tools that aim to limit problematic content, ads that users find problematic still have a significant presence on popular sites [259] due to both policy inadequacies [158, 190] and technical challenges of catching such content [213]. Recent work has investigated user opinions of problematic content as well; Zeng et al. develop a taxonomy on what users think are the worst qualities of "bad" ads [259, 260], finding that people have consistent reasons of disliking ads, and are particularly likely to dislike ads described as "deceptive," "clickbait", "ugly" and "politicized". Related work has also found that people struggle to identify deceptive ads [256], which can lead to harmful outcomes like software attacks [183, 257]. Additionally, those who suffer from certain mental health disorders or trauma may also experience negative psychological and physical consequences from ads that target these conditions [100].

3.2.5 The Harms of Advertising

An emergent line of work also looks at the concrete harms of advertising—Gak et al. [100] and Wu et al. [255] specifically explore how it relates to Rob Nixon's notion of "slow violence" from the study of environmentalism [187]. Gak et al. find through unstructured interviews with participants who had an eating disorder that repeated exposure to weight-loss ads caused significant mental distress. Their interviews establish that psychological distress and a loss of autonomy are concrete harm that users of advertising can experience. Wu et al. [255] also note how the *surveillance capitalism* [264] enabled by advertising, which leads to the normalization of people's discomfort. Through surveys of multiple participants who describe being harmed by ads, they build a taxonomy of advertising's harms: psychological distress, loss of autonomy, constriction of user behavior, and algorithmic marginalization.

In an alternate line of work, Milano et al. [174] argue that advertising is harmful because of the *epistemic fragmentation* it creates. In their conception, ads are not only harmful because of the content they contain, but also due to the omission of valuable content (e.g. job opportunities) from users. Further, they propose that ads can be harmful universally (e.g. misinformation), as well as contextually, such as when it targets an individual vulnerability.

3.3 Bias and Polarization in Personalization

Within the Recommender Systems community, specifically within the rich multi-armed bandits literature [163, 217, 227], multiple approaches have been proposed to implement fairness and reduce polarization in personalization. Joseph et al. [131] formally introduce the notion of fairnes in bandit optimization, with a particular focus on situations where the algorithm might be choosing between different demographic groups for a decision. Their fairness algorithm focuses on merit, enforcing at each time step that a worse candidate is not preferred over a better one. While a direct translation to advertising in particular isn't provided, Joseph et al.'s results are broadly motivated by Sweeney [225].

Celis et al. [34] provide a general-purpose framework to broadly curtail *polarization* in personalization, which is directly applicable to advertising, alongside other domains that use content selection. They achieve this by specifically defining *groups* of arms in a bandit that might correspond to different content themes at risk of polarization, and constraining the likelihood of picking content at the group level. While generally applicable and useful, their approach requires clearly defined content themes (groups), as well as mapping of all content to these themes, for each kind of harm to avoid. As an alternative to directly controlling the content selection algorithm, modifying the bidding strategies has also been proposed as a solution to avoid gender discrimination [180].

Other approaches also choose to focus on some notion of fairness or user health, while balancing performance tradeoffs. Mehrotra et al. [165] present a group fairness criterion and constrain their personalization system to ensure equity of attention by popularity, in the context of music recommendations. Singh et al. [217] focus on recommendation *trajectories* in particular, defining a notion of user health while watching videos. Their definition of harm isn't based on individual pieces of content, but rather a trajectory of videos that a user might explore. Their proposed method constrains on harm to the worst-off users, and is able to control this worst-case harm without a significant drop in overall performance.

Chapter 4

Gender and Racial Biases in Job and Housing Ads

I now move on to describe the first study done in this dissertation. We investigate how advertising platforms can skew delivery of consequential opportunity ads that have legal protections, such as those for jobs and housing. We specifically investigate skews that arises along demographics of users, particularly self-described gender and race. We focus on Facebook—as it is the most mature platform offering advanced targeting features—and run dozens of ad campaigns, hundreds of ads with millions of impressions, spending over $8,500 as part of our study. The discussion that follows is reproduced from our published work at the *ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW)*, 2019 [8].

Due to the wide variety of targeting features—as well as the availability of sensitive targeting features such as user demographics and interests—researchers have raised concerns about discrimination in advertising, where groups of users may be excluded from receiving certain ads based on advertisers' targeting choices [222]. This concern is particularly acute in the areas of credit, housing, and employment, where there are legal protections in the U.S. that prohibit discrimination against certain protected classes in advertising [1–3]. As ProPublica demonstrated in 2016 [16], this risk is not merely theoretical: ProPublica investigators were able to run housing ads that explicitly excluded users with specific "ethnic affinities" from receiving them.[1] Recently, the U.S. Department of Housing and Urban Development (HUD) sued Facebook over these concerns and others, accusing Facebook's advertising platform of "encouraging, enabling, and causing" violations of the Fair Housing Act [81].

The role of ad delivery in discrimination Although researchers and investigative journalists have devoted considerable effort to understanding the potential discriminatory outcomes of ad targeting, comparatively little effort has focused on ad delivery, due to the difficulty of studying its impacts without internal access to ad platforms' data and mechanisms. However, there are several potential reasons why the ad delivery algorithms used by a platform may open the door to discrimination.

First, consider that most platforms claim their aim is to show users "relevant" ads: for example, Facebook states "we try to show people the ads that are most pertinent to them" [45]. Intuitively, the goal is to show ads that particular users are likely to engage with, even in cases where the advertiser does not know a priori which users are most receptive to their message. To accomplish this, the platforms build extensive user interest profiles and track ad performance to understand how different users interact with different ads. This historical data is then used to steer future ads towards those users who are most likely to be interested in them, and to users like them. However, in doing so, the platforms may inadvertently cause ads to deliver primarily to a skewed subgroup of the advertiser's selected audience, an outcome that the advertiser may not have intended or be aware of. As noted above, this is particularly concerning in the case of credit, housing, and employment, where such skewed delivery might violate antidiscrimination laws.

Second, market effects and financial optimization can play a role in ad delivery, where different desirability of user populations and unequal availability of users may lead to skewed ad delivery [61]. For example, it is well-known that certain users on advertising platforms are more valuable to advertisers than others [134, 155, 207]. Thus, advertisers who choose low budgets when placing their ads may be more likely to lose auctions for such "valuable" users than advertisers who choose higher budgets. However, if these "valuable" user demographics

[1] In response, Facebook banned the use of certain attributes for housing ads, but many other, un-banned, mechanisms exist for advertisers that achieve the same outcome [222]. Facebook agreed as part of a lawsuit settlement stemming from these issues to go further by banning age, gender, and certain kinds of location targeting—as well as some related attributes—for housing, employment, or credit ads [59].

are strongly correlated with protected classes, it could lead to discriminatory ad delivery due to the advertiser's budget alone. Even though a low budget advertiser may not have intended to exclude such users, the ad delivery system may do just that because of the higher demand for that subgroup.

Prior to this work, although hypothesized [61, 147, 238], it was not known whether the above factors resulted in skewed ad delivery in real-world advertising platforms. In fact, in response to the HUD lawsuit [81] mentioned above, Facebook claimed that the agency had "no evidence" of their ad delivery systems' role in creating discrimination [126].

Contributions In this study, we aim to understand whether ads could end up being shown in a skewed manner—i.e., where some users are less likely than others to see ads based on their demographic characteristics—due to the ad delivery phase alone. In other words, we determine whether the ad delivery could cause skewed delivery *that an advertiser did not cause by their targeting choices and may not even be aware of.* We focus on Facebook—as it is the most mature platform offering advanced targeting features—and run dozens of ad campaigns, hundreds of ads with millions of impressions, spending over $8,500 as part of our study.

Answering this question—especially without internal access to the ad delivery algorithm, user data, and advertiser targeting data or delivery statistics—involves overcoming a number of challenges. These include separating market effects from optimization effects, distinguishing ad delivery adjustments based on the ad's performance measured through user feedback from initial ad classification, and developing techniques to determine the racial breakdown of the delivery audience (which Facebook does not provide). The difficulty of solving these without the ad platform's cooperation in a rigorous manner may at least partially explain the lack of knowledge about the potential discriminatory effects due to ad delivery to date. After addressing these challenges, we find the following:

First, we find that *skewed delivery can occur due to market effects alone.* Recall the hypothesis above concerning what may happen if advertisers in general value users differently across protected classes. Indeed, we find this is the case on Facebook: when we run identical ads targeting the same audience but with varying budgets, the resulting audience of users who end up seeing our ad can range from over 55% men (for ads with very low budgets) to under 45% men (for ads with high budgets).

Second, we find that *skewed delivery can occur due to the content of the ad itself* (i.e., the ad headline, text, and image, collectively called the *ad creative*). For example, ads targeting the same audience but that include a creative that would stereotypically be of the most interest to men (e.g., bodybuilding) can deliver to over 80% men, and those that include a creative that would stereotypically be of the most interest to women (e.g., cosmetics) can deliver to over 90% women. Similarly, ads referring to cultural content stereotypically of most interest to Black users (e.g., hip-hop) can deliver to over 85% Black users, and those referring to content stereotypically of interest to white users (e.g., country music) can deliver to over 80% white users, even when targeted identically by the advertiser. Thus, despite placing the same bid on the same audience, the advertiser's ad delivery can be heavily skewed based on the ad creative alone.

Third, we find that *the ad image itself has a significant impact on ad delivery.* By running experiments where we swap different ad headlines, text, and images, we demonstrate that the differences in ad delivery can be significantly affected by the image alone. For example, an ad whose headline and text would stereotypically be of the most interest to men with the image that would stereotypically be of the most interest to women delivers primarily to women at the same rate as when all three ad creative components are stereotypically of the most interest to women.

Fourth, we find that *the ad image is likely automatically classified by Facebook*, and that this classification can skew delivery from the beginning of the ad's run. We create a series of ads where we add an alpha channel to stereotypically male and female images with over 98% transparency; the result is an image with all of the image data present, but that looks like a blank white square to humans. We find that there are statistically significant differences in how these ads are delivered depending on the image used, despite the ads being visually indistinguishable to a human. This indicates that the image classification—and, therefore, relevance determination—is likely an automated process, and that the skew in ad delivery can be due in large part to skew in Facebook's automated estimate of relevance, rather than ad viewers' interactions with the ad.

Fifth, we show that *real-world employment and housing ads can experience significantly skewed delivery.* We create and run ads for employment and housing opportunities, and use our methodology to measure their delivery to users of different races and genders. When optimizing for clicks, we find that ads with the same targeting options can deliver to vastly different racial and gender audiences depending on the ad creative alone. In the most extreme cases, our ads for jobs in the lumber industry reach an audience that is 72% white and 90% male, our ads for cashier positions in supermarkets reach an 85% female audience, and our ads for positions in taxi companies reach a 75% Black audience, even though the targeted audience specified by us as an advertiser is identical for all three. We run a similar suite of ads for housing opportunities, and find skew there as well: despite the same targeting and budget, some of our ads deliver to an audience of over 72% Black users, while others delivery to over 51% Black users. While our results only speak to how our particular ads are delivered (i.e., we cannot say how housing or employment ads *in general* are delivered), the significant skew we observe even on a small set of ads suggests that real-world housing and employment ads are likely to experience the same fate.

Taken together, our results paint a distressing picture of heretofore unmeasured and unaddressed skew that can occur in online advertising systems, which have significant implications for discrimination in targeted advertising. Specifically, due to platforms' optimization in the ad delivery stage together with market effects, ads can unexpectedly be delivered to skewed subsets of the advertiser's specified audience. For certain types of ads, such skewed delivery might implicate legal protections against discriminatory advertising. For example, Section 230 of the U.S. Communications Decency Act (CDA) protects publishers (including online platforms) from being held responsible for third-party content. Our results show Facebook's integral role in shaping the delivery mechanism and might make it more difficult for online platforms to present themselves as neutral publishers in the future. We leave a full exploration of these implications to the legal community. However, our results indicate that regulators, lawmakers, and the platforms themselves need to think carefully when balancing the optimization of ad platforms against desired societal outcomes, and remember that ensuring that individual advertisers do not discriminate in their targeting is insufficient to achieve non-discrimination goals sought by regulators and the public.

Ethics All of our experiments were conducted with careful consideration of ethics. We obtained Institutional Review Board review of our study at Northeastern University (application #18-11-13), with our protocol being marked as "Exempt". We minimized harm to Facebook users when we were running our ads by always running "real" ads (in the sense that if people clicked on our ads, they were brought to real-world sites relevant to the topic of the ad). While running our ads, we never intentionally chose to target ads in a discriminatory manner (e.g., we never used discriminatory targeting parameters). To further minimize the potential for discrimination, we ran most of our experimental ads in categories with no legal salience (such as entertainment and lifestyle); we only ran ad campaigns on jobs and housing to verify whether the effects we observed persist in these domains. We minimized harm to the Facebook advertising platform by paying for ads and using the ad reporting tools in the same manner as any other advertiser. The particular sites we advertised were unaffiliated with the study, and our ads were not defamatory, discriminatory, or suggestive of discrimination.

4.1 Methods

We now describe our methodology for measuring the delivery of Facebook ads. At a high level, our goal is to run groups of ads where we vary a particular feature, with the goal of then measuring how changing that feature skews the set of users the Facebook platform delivers the ad to. To do so, we need to carefully control which users are in our target audience. We also need to develop a methodology to measure the ad delivery skew along racial lines, which, unlike gender, is not provided by Facebook's existing reporting tools. We detail how we achieve that in the following sections.

4.1.1 Audience selection

When running ads, we often wish to control exactly which ad auctions we are participating in. For example, if we are running multiple instances of the same ad (e.g., to establish statistical confidence), we do not want the instances to be competing against each other. To this end, we use random PII-based custom audiences, where we randomly select U.S. Facebook users to be included in mutually-exclusive audiences. By doing so, we can ensure that our ads are only competing against each other in the cases where we wish them to. We also replicate some of the experiments while targeting all U.S. users to ensure that the effects do not only exist when custom audiences are targeted. As we show later in Section 4.2, we observe equivalent skews in these scenarios, which leads us to believe that preventing internal competition between our own ads is not crucial to measure the resulting skews.

Generating custom audiences We create each custom audience by randomly generating 20 lists of 1,000,000 distinct, valid North American phone numbers (`+1 XXX XXX XXXX`, using known-valid area codes). Facebook reported that they were able to match approximately 220,000 users on each of the 20 lists we uploaded.

Initially, we used these custom audiences directly to run ads, but while conducting the experiments we noticed that—even though we specifically target only North American phone numbers—many ads were delivered to users outside of North America. This could be caused by users traveling abroad, users registering with fake phone numbers or with online phone number services, or for other reasons, whose investigation is outside the scope of this study.

Therefore, for all the experiments where we target custom audiences, we additionally limit them to people located in the U.S.

4.1.2 Data collection

Once one of our ad campaigns is run, we use the Facebook Marketing API to obtain the delivery performance statistics of the ad every two minutes. When we make this request, we ask Facebook to break down the ad delivery performance according to the attribute of study (age, gender, or location). Facebook's response to each query features the following fields, among others, for each of the demographic attributes that we requested:

- `impressions`: The number of times the ad was shown
- `reach`: The number of unique users the ad was shown to
- `clicks`: The number of clicks the ad has received
- `unique_clicks`: The number of unique users who clicked

Throughout the rest of the study, we use the `reach` value when examining delivery; thus, when we report "Fraction of men in the audience" we calculate this as the `reach` of men divided by the sum of the `reach` of men and the `reach` of women (see Section 4.1.5 for discussion on using binary values for gender).

4.1.3 Measuring racial ad delivery

The Facebook Marketing API allows advertisers to request a breakdown of ad delivery performance along a number of axes but it does not provide a breakdown based on race. However, for the purposes of this work, we are able to measure the ad delivery breakdown along racial lines by using location (Designated Market Area, or DMA[2]) as a proxy.

Similar to prior work [222], we obtain voter records from North Carolina; these are publicly available records that have the name, address, race, and often phone number of each registered voter in the state. We partition the most populated North Carolina DMAs into

[2] Designated Market Areas [182] are groups of U.S. counties that Neilson defines as "market areas"; they were originally used to signify a region where users receive similar broadcast television and radio stations. Facebook reports ad delivery by location using DMAs, so we use them here as well.

Table 4.1 Overview of the North Carolina custom audiences used to measure racial delivery. We divide the most populated DMAs in the state into two sets, and create three audiences each with one race per DMA set. Audiences A and B are disjoint from each other; audience C contains the voters from A with additional white voters from the first DMA set and Black voters from the second DMA set. We then use the statistics Facebook reports about delivery by DMAs to infer delivery by race.

DMA(s) [182]	**# Records (A)**		**# Records (B)**		**# Records (C)**	
	White	**Black**	**White**	**Black**	**White**	**Black**
Wilmington, Raleigh–Durham	400,000	0	0	400,000	900,002	0
Greenville-Spartanburg, Greenville-New Bern, Charlotte, Greensboro	0	400,000	400,000	0	0	892,097

two sets; for the exact DMAs, please see Table 4.1. We ensure that each racial group (white and Black) from a set of DMAs has a matching number of records of the other group in the other set of DMAs. We sample three audiences (A, B, and C) that fit these constraints from the voter records and upload as separate Custom Audiences to Facebook.[3] Audiences A and B are disjoint from each other; audience C contains the voters from A with additional white voters from the first DMA set and Black voters from the second DMA set. We create audiences in this way to be able to test both "flipped" versions of the audiences (A and B), as well as large audiences with as many users as possible (C); we created audience B as large as possible (exhausting all voters who fit the necessary criteria), and sampled audience A to match its size. The details of the resulting audiences are shown in Table 4.1.

When we run ads where we want to examine the ad delivery along racial lines, we run the ads to one audience (A, B, or C). We then request that Facebook's Marketing API deliver us results broken down by DMA. Because we selected DMAs to be a proxy for race, we can use the results to infer which custom audience they were originally in, allowing us to determine the racial makeup of the audience who saw (and clicked on) the ad. Note that in experiments that involve measuring racial skew all ads target the same target audience. The

[3] Unfortunately, Facebook does not report the number of these users who match as we use multiple PII fields in the upload file [243].

limited number of registered voters does not allow us to create many large, disjoint custom audiences like we do in other experiments. However, as we show with ads targeting all U.S. users, internal competition does not appear to influence the results.

4.1.4 Ad campaigns

We use the Facebook Ad API described in Section 2.4.5 to create all ads for our experiments and to collect data on their delivery. We carefully control for any time-of-day effects that might be present due to different user demographics using Facebook at different times of the day: for any given experiment, we run all ads at the same time to ensure that any such effects are experienced equally by all ads. Unless otherwise noted, we used the following settings:

- *Objective:* Consideration→Traffic[4]
- *Optimization Goal:* Link Clicks
- *Traffic destination:* An external website (that depends on the ads run)
- *Creative:* All of our ads had a single image and text relevant to the ad.
- *Audience selection:* We use custom audiences for many of our ads, as described in Section 4.1.1, and further restrict them to adult (18+) users of all genders residing in the United States. For other ads, we target all U.S. users age 18 or older.
- *Budget:* We ran most ads with a budget of $20 per day, and stopped them typically after six hours.

4.1.5 Measuring and comparing audiences

We now describe the measurements we make during our experiments and how we compute their confidence intervals.

Binary values of gender and race Facebook's marketing API reports "female", "male", and "uncategorized" as the possible values for gender. Facebook's users self-report their gender, and the available values are "female", "male", and "custom". The latter allows the user to manually type in their gender (with 60 predefined gender identities suggested through

[4] This target is defined as: Send more people to a destination on or off Facebook such as a website, app, or Messenger conversation.

auto-complete functionality) and select the preferred pronoun from "female - her", "male - him", and "neutral - them". Across our experiments, we observe that up to 1% of the audiences are reported as "uncategorized" gender. According to Facebook's documentation this represents the users who did not not list their gender [171]. We do not know whether the "uncategorized" gender also features users with self-reported "custom" gender. Thus, in this work we only consider the self-reported binary gender values of "female" and "male".

Further, when considering racial bias, we use the self-reported information from voter records. The data we obtained has 7,560,885 individuals, with 93% reporting their race as either Black or White. Among those, less than 1% report their ethnicity as "Hispanic/Latino". Thus, in this work, we only target the individuals with self-reported race of White or Black. However, when running our experiments measuring race (and targeting specific DMAs), we observe that a fraction (∼10%) of our ads are delivered to audiences outside of our predefined DMAs, thus making it impossible for us to infer their race. This fraction remains fairly consistent across our experiments regardless of what we advertise, thus introducing the same amount of noise across our measurements. This is not entirely unexpected, as we are targeting users directly, and those users may be traveling, may have moved, may have outdated information in the voter file, etc.

We do not claim that gender or race are binary, but choose to focus the analysis on users who self-reported their gender as "female" or "male" and race as "Black" or "White". This way, we report the observable skew in delivery only along these axes. We recognize that delivery can be *further* skewed with respect to gender of non-binary users and/or users of other races in a way that remains unreported in this work.

Measuring statistical significance Using the binary race and gender features, throughout this work, we describe the audiences by the fraction of male users and the fraction of white users. We calculate the lower and upper limits of the 99% confidence interval around this fraction using the method recommended by Agresti and Coull [6], defined in Equation 5.1:

$$
\begin{aligned}
L.L. &= \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n}, \\
U.L. &= \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n},
\end{aligned} \tag{4.1}
$$

where $L.L.$ is the lower confidence limit, $U.L.$ is the upper confidence limit, $\hat{p}$ is the observed fraction of the audience with the attribute (here: male), n is the size of the audience reached by the ad. To obtain the 99% interval we set $z_{\alpha/2} = 2.576$. The advantage of using this calculation instead of the more frequently used normal approximation

$$p \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{4.2}$$

is that the resulting intervals fall in the $(0,1)$ range. Whenever the confidence intervals around these fractions for two audiences are non-overlapping, we can make a claim that the gender or racial makeups of two audiences are significantly different [51]. However, the converse is not true: overlapping confidence intervals do not necessarily mean that the means are not different (see Figure 4 in [51] for explanation). In this work we report all the results of our experiments but for easier interpretation emphasize those where the confidence intervals are non-overlapping. We further confirm that the non-overlapping confidence intervals represent statistically significant differences, using the difference of proportion test as shown in Equation 5.2:

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - 0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{4.3}$$

where $\hat{p_1}$ and $\hat{p_2}$ are the fractions of men (white users) in the two audiences that we compare, n_1 and n_2 are sizes of these audiences, and $\hat{p}$ is the fraction of men (white users) in the two delivery audiences combined. All the results we refer to as statistically significant are significant in this test with a Z-score of at least 2.576.

Note that in experiments where we run multiple instances of an ad targeting disjoint custom audiences, the values of $\hat{p}$ and n are calculated from the sums of reached audiences.

4.2 Experiments and Results

In this section, we explore how an advertiser's choice of ad creative (headline, text, and image) and ad campaign settings (bidding strategy, targeted audience) can affect the demographics (gender and race) of the users to whom the ad is ultimately delivered.

4.2.1 Budget effects on ad delivery

We begin by examining the impact that market effects can have on delivery, aiming to test the hypothesis put forth by Lambrecht et al. [147]. In particular, they observed that their ads were predominantly shown to men even though women had consistently higher click through rates (CTRs). They then hypothesized that the higher CTRs led to women being more expensive to advertise to, meaning they were more likely to lose auctions for women when compared to auctions for men.

We test this hypothesis by running the same ad campaign with different budgets; our goal is to measure the effect that the daily budget alone has on the makeup of users who see the ads. When running these experiments, we keep the ad creative and targeted audience constant, only changing the bidding strategy to give Facebook different daily limits (thus, any ad delivery differences can be attributed to the budget alone). We run an ad with daily budget limits of \$1, \$2, \$5, \$10, \$20, and \$50, and run multiple instances at each budget limit for statistical confidence. Finally, we run the experiment twice, once targeting our random phone number custom audiences, and once targeting all users located in U.S.; we do so to verify that any effect we see is not a function of our particular target audience, and that it persists also when non-custom audiences are targeted.

Figure 4.1 presents the results, plotting the daily budget we specify versus the resulting fraction of men in the audience. The left graph shows the results when we target all users located in the U.S., and the right graph shows the results when we target the random phone number custom audiences. In both cases, we observe that changes in ad delivery due to differences in budget are indeed happening: the higher the daily budget, the smaller the fraction of men in the audience, with the Pearson's correlation of $\rho = -0.88$, $p_{val} < 10^{-5}$ for all U.S. users and $\rho = -0.73$, $p_{val} < 10^{-3}$ for the custom audiences.

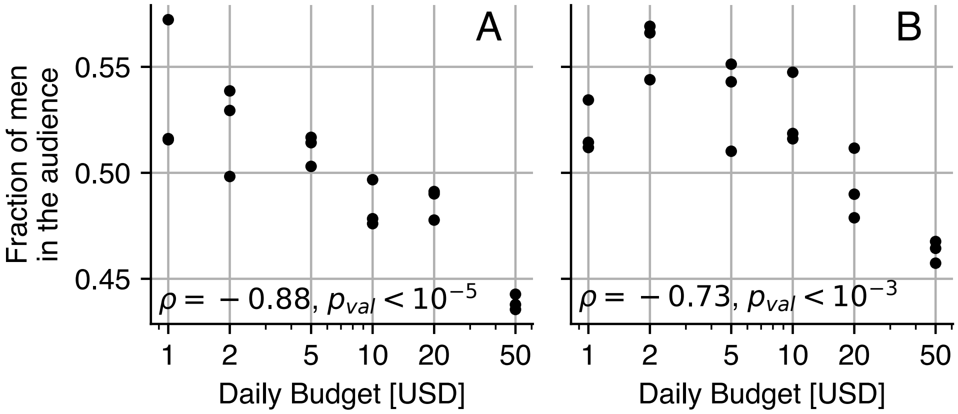


Figure 4.1: Gender distributions of the audience depend on the daily budget of an ad, with higher budgets leading to a higher fraction of women. The left graph shows an experiment where we target all users located in the U.S.; the right graph shows an experiment where we target our random phone number custom audiences.

The stronger effect we see when targeting all U.S. users may be due to the additional freedom that the ad delivery system has when choosing who to deliver to, as this is a significantly larger audience.

To eliminate the impact that market effects can have on delivery in our following experiments, we ensure that all runs of a given experiment use the same bidding strategy and budget limit. Typically we use a daily budget of $20 per campaign.

4.2.2 Ad creative effects on ad delivery

Now we examine the effect that the ad creative (headline, text, and image) can have on ad delivery. To do so, we create two stereotypical ads that we believe would appeal primarily to men and women, respectively: one ad focusing on *bodybuilding* and another on *cosmetics*. The actual ads themselves are shown in Figure 2.1. We run each of the ads at the same time and with the same bidding strategy and budget. For each variable we target different custom audiences, i.e., the "base" level ads target one audience, "text" level ads target another, etc. *Note that we do not explicitly target either ad based on gender; the only targeting restrictions we stipulate are 18+ year old users in the U.S.*

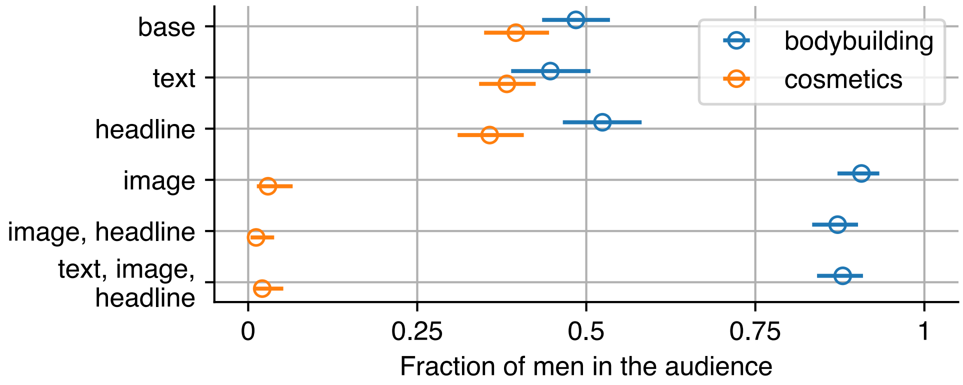


Figure 4.2: “Base” ad contains a link to a page about either bodybuilding or cosmetics, a blank image, no text, or headline. There is a small difference in the fraction of male users for the base ads, and adding the “text” only decreases it. Setting the “headline” sets the two ads apart but the audience of each is still not significantly different than that of the base version. Finally, setting the ad “image” causes drastic changes: the bodybuilding ad is shown to a 91% male audience, the cosmetics ad is shown to a 5% male audience, despite the same target audience.

We observe dramatic differences in ad delivery, even though the bidding strategy is the same for all ads, and each pair of ads target the same gender-agnostic audience. In particular, the bodybuilding ad ended up being delivered to over 75% men on average, while the cosmetics ad ended up being delivered to over 90% women on average. Again, this skewed delivery is despite the fact that we—the advertiser—did not specify difference in budget or target audience.

Individual components’ impact on ad delivery With the knowledge that the ad creative can skew delivery, we dig deeper to determine *which* of the components of the ad creative (headline, text, and image) have the greatest effect on ad delivery. To do so, we stick with the bodybuilding and cosmetics ads, and “turn off” various features of the ad creative by replacing them with empty strings or blank images. For example, the bodybuilding experiment listed as “base” includes an empty headline, empty ad text, and a blank white image; it does however link to the domain `bodybuilding.com`. Similarly, the cosmetics experiment listed as “base” includes no headline, text, or image, but does link to the domain `elle.com`. We then add back various parts of the ad creative, as shown in Figure 2.1.

The results of this experiment are presented in Figure 4.2. Error bars in the figure correspond to 99% confidence intervals as defined in Equation 5.1. All results are shown relative to that experiment's "base" ad containing only the destination URL. We make a number of observations. *First*, we can observe an ad delivery difference due to the destination URL itself; the base bodybuilding ad delivers to 48% men, while the base cosmetics ad delivers to 40% men. *Second*, as we add back the title and the headline, the ad delivery does not appreciably change from the baseline. However, once we introduce the image into the ad, the delivery changes dramatically, returning to the level of skewed delivery discussed above (over 75% male for bodybuilding, and over 90% female for cosmetics). When we add the text and/or the headline back alongside the image, the skew of delivery does not change significantly compared to the presence of image only. Overall, our results demonstrate that the choice of ad image can have a dramatic effect on which users in the audience ultimately are shown the ad.

Swapping images To further explore how the choice of image impacts ad delivery, we continue using the bodybuilding and cosmetics ads, and test how ads with incongruent images and text are delivered. Specifically, we swap the images between the two ads, running an ad with the bodybuilding headline, text, and destination link, but with the image from cosmetics (and vice versa). We also run the original ads (with congruent images and text) for comparison.

The results of this experiment are presented in Figure 4.3, showing the skew in delivery of the ads over time. The color of the lines indicates the image that is shown in the ad; solid lines represent the delivery of ads with images consistent with the description, while dotted lines show the delivery for ads where image was replaced. We make a number of observations. *First*, when using congruent ad text and image (solid lines), we observe the skew we observed before. However, we can now see clearly that this delivery skew appears to exist from the very beginning of the ad delivery, i.e., before users begin viewing and interacting with our ads. We will explore this further in the following section. *Second*, we see that when we switch the images—resulting in incongruent ads (dotted lines)—the skew still exists but to a lesser degree. Notably, we observe that the ad with an image of bodybuilding but cosmetics text delivers closest to 50:50 across genders, but the ad with the image of cosmetics but

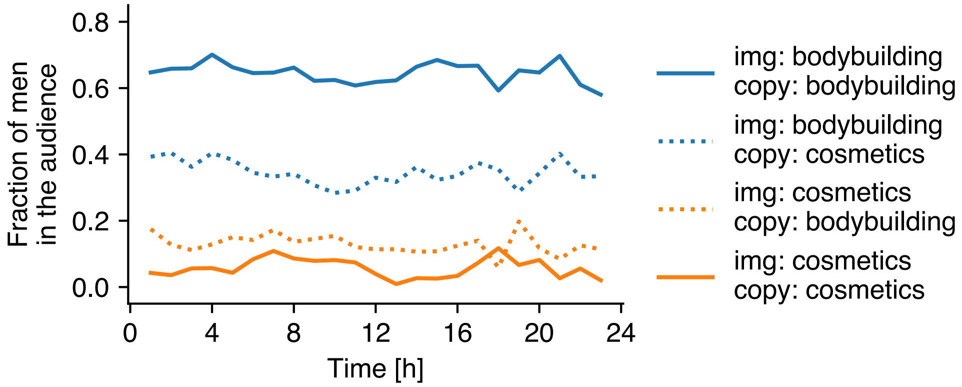


Figure 4.3: Ad delivery of original bodybuilding and cosmetics ads, as well as the same ads with incongruent images. Skew in delivery is observed from the beginning. Using incongruent images skews the delivery to a lesser degree, indicating that the image is not the only element of the ad that drives the skew in delivery.

bodybuilding text does not. The exact mechanism by which Facebook decides to use the ad text and images in influencing ad delivery is unknown, and we leave a full exploration to future work.

Swapping images mid-experiment Facebook allows advertisers to change their ad while it is running, for example, to update the image or text. As a final point of analysis, we examine how changing the ad creative mid-experiment—after it has started running—affects ad delivery. To do so, we begin the experiment with the original congruent bodybuilding and cosmetics ads; we let these run for over six hours. We then swap the images on the running ads, thereby making the ads incongruent, and examine how ad delivery changes.

Figure 4.4 presents the results of this experiment. In the top graph, we show the instantaneous ad delivery skew: as expected, the congruent ads start to deliver in a skewed manner as we have previously seen. After the image swap at six hours, we notice a very rapid change in delivery with the ads almost completely flipping in ad delivery skew in a short period of time. Interestingly, we do not observe a significant change in users' behavior to explain this swap: the bottom graph plots the click through rates (CTRs) for both ads by men and women over time. Thus, our results suggest that the change in ad delivery skew is unlikely to be due to the users' responses to the ads.

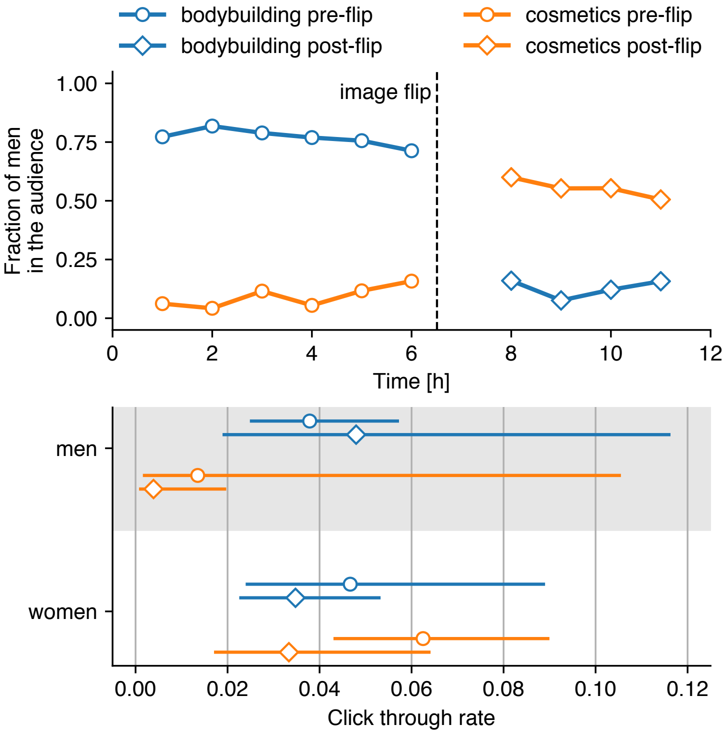


Figure 4.4: When we flip the image in the middle of the campaign, the ad is reclassified and shown to an updated audience. Here, we start bodybuilding and cosmetics ads with corresponding descriptions and after 6 hours and 32 minutes we flip the images. Within an hour of the change, the gender proportions are reversed, while there is no significant difference between the click through rates per gender pre and post flipping of the images.

4.2.3 Source of ad delivery skew

We just observed that ads see a significant skew in ad delivery due to the contents of the ad, despite the bidding strategy and targeting parameters being held constant. However, we observed that the ad delivery skew was present from the very beginning of ad delivery, and that swapping the image in the middle of a run resulted in a very rapid change in ad delivery that could not be explained by how frequently users click on our ads. We now turn to explore the mechanism that may be leading to this ad delivery skew.

Almost-transparent images We begin with the hypothesis that Facebook itself is automatically classifying the ad creative (including the image), and using the output of this classification to calculate a predicted relevance score to users. In other words, we hypothesize that Facebook is running automatic text and image classification, rather than (say) relying on the ad's initial performance, which would explain (a) the delivery skew being present from the beginning of ad delivery, and (b) how the delivery changes rapidly despite no significant observable change in user behavior. However, validating this hypothesis is tricky, as we are not privy to all of Facebook's ad performance data.

To test this hypothesis, we take an alternate approach. We use the *alpha channel* that is present in many modern image formats; this is an additional channel that allows the image to encode the *transparency* of each pixel. Thus, if we take an image and add an alpha channel with (say) 99% opacity, all of the image data will still be present in the image, but any human who views the image would not be able to see it (as the image would show almost completely transparent). However, if an automatic classifier exists, and if that classifier is not properly programmed to handle the alpha channel, it may continue to classify the image.

Test images To test our hypothesis, we select five images that would stereotypically be of interest to men and five images that would stereotypically be of interest to women; these are shown in the second and fourth columns of Table 4.2.[5,6] We convert them to PNG format add an alpha channel with 98% opacity[7] to each of these images; these are shown in the third and fifth columns of Table 4.2. Because we cannot render a transparent image without a background, the versions in Table 4.2 are rendered on top of a white background. As the reader can see, these images are not discernible to the human eye.

We first ran a series of tests to observe how Facebook's ad creation phase handled us uploading such transparent images. If we used Reach as our ad objective, we found that Facebook "flattened" these images onto a white background in the ad preview.[8] By targeting ourselves with these Reach ads, we verified that when they were shown to users on the

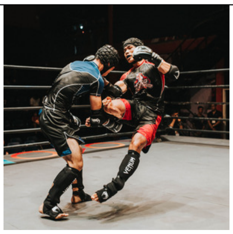
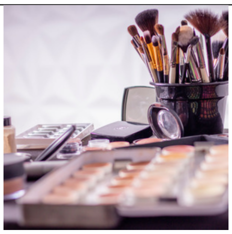
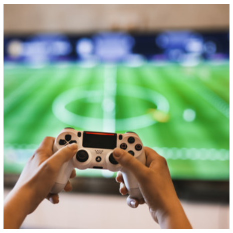
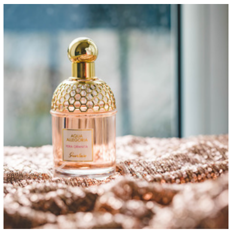
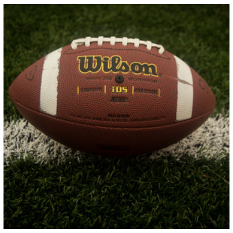
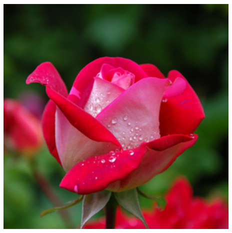
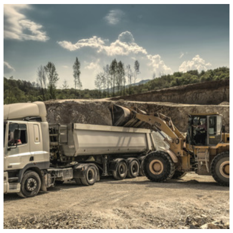
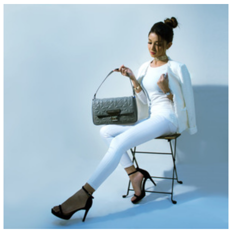
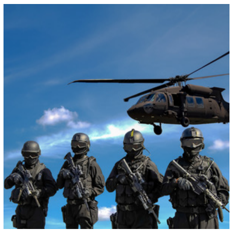
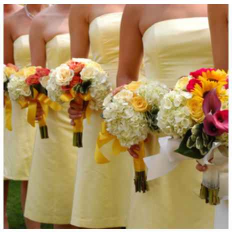
[5] All of these images were cropped from images posted to `pexels.com`, which allow free non-commercial use.

[6] We cropped these images to the Facebook-recommended resolution of 1,080×1,080 pixels to reduce the probability Facebook would resample the image.

[7] We were unable to use 100% transparency as we found that Facebook would run an image hash over the uploaded images and would detect different images with 100% opacity to be the same (and would refuse to upload it again). By using 98% transparency, we ensure that the images were still almost invisible to humans but that Facebook would not detect they were the same image.

[8] Interestingly, we found that if we instead used Traffic as our ad objective, Facebook would both "flatten" these images onto a white background *and then normalize the contrast*. This caused the ads to be visible to humans—simply with less detail that the original ads—thus defeating the experiment. We are unsure of why Facebook did not choose to normalize images with the objective for Reach.

Table 4.2 Diagram of the images used in the transparency experiments. Shown are the five stereotypical masculine and feminine images, along with the same images with a 98% alpha channel, denoted as invisible. The images with the alpha channel are almost invisible to humans, but are still delivered in a skewed manner.

	Masculine		**Feminine**	
No.	**Visible**	**Invisible**	**Visible**	**Invisible**
1				
2				
3				
4				
5				

Facebook mobile app or in the desktop Facebook web feed, the images did indeed show up as white squares. Thus, we can use this methodology to test whether there is an automatic image classifier present by examining whether running different transparent white ads results in different delivery.

Results We run ads with all twenty of the images in Table 4.2, alongside ads with five truly blank white images for comparison. For all 25 of these ads, we hold the ad headline, text, and destination link constant, run them all at the same time, and use the same bidding strategy and target custom audiences in a way that each user is potentially exposed to up to three ads (one masculine image, one feminine image, and one blank image). We then record the differences in ad delivery of these 25 images along gender lines. The results are presented in Figure 4.5A, with all five images in each of the five groups aggregated together. We can observe that ad delivery is, in fact, skewed, with the ads with stereotypically masculine images delivering to over 43% men and the ads with feminine images delivering to 39% men in the experiment targeting custom audiences as well as 58% and 44% respectively in the experiment targeting all U.S. users. Error bars in the plot correspond to the 99% confidence interval calculated using Equation 5.1.

Interestingly, we also observe that the masculine invisible ads appear to be indistinguishable in the gender breakdown of their delivery from the masculine visible ads, and the feminine invisible ads appear to be indistinguishable in their delivery from the feminine visible ads.

As shown in Figure 4.5A, we verify that the fraction of men in the delivery of the male ads is significantly higher than in female-centered and neutral ads, as well as higher in neutral ads than in female-centered ads. We also show that we cannot reject the null hypothesis that the fraction of men in the two versions of each ad (one visible, one invisible) are the same. Thus, we can conclude that the difference in ad delivery of our invisible male and female images is statistically significant, despite the fact that humans would not be able to perceive any differences in these ads. This strongly suggests that our hypothesis is correct: that Facebook has an automated image classification mechanism in place that is used to steer different ads towards different subsets of the user population.[9]

[9] It is important to note we not know exactly how the classification works. For example, the classifier may also be programmed to take in the “flattened” images that appear almost white, but there may sufficient data present in the images for the classification to work.

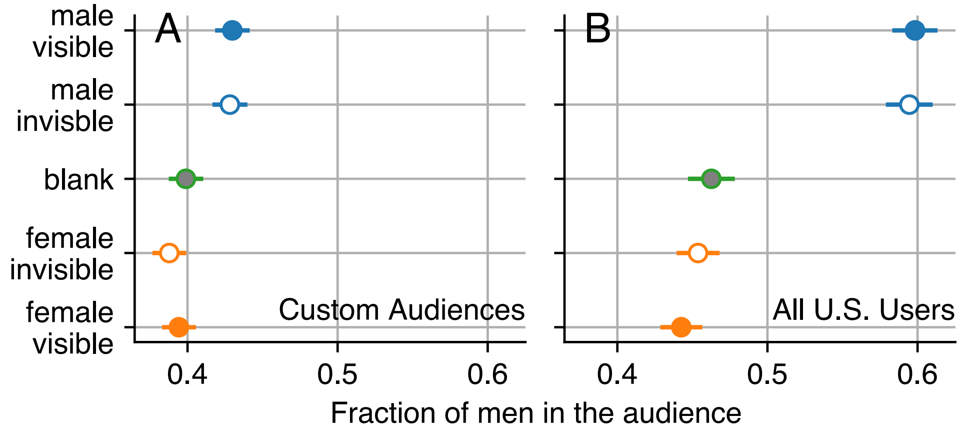


Figure 4.5: Fraction of reached men in the audiences for ads with the images from Table 4.2, targeting random phone number custom audience (A) and US audience (B). The solid markers are visible images, and the hollow markers are the same images with 98% opacity. Also shown is the delivery to truly white images ("blank"). We can observe that a difference in ad delivery exists, and that that difference is statistically significant between the masculine and feminine invisible images. This suggests that automated image classification is taking place.

To confirm this finding, we re-run the same experiment except that we change the target audience from our random phone number custom audiences (hundreds of thousands of users) to all U.S. users (over 320 million users). Our theory is that if we give Facebook's algorithm a larger set of auctions to compete in, any effect of skewed delivery would be amplified as they may be able to find more users for whom the ad is highly "relevant". In Figure 4.5B we observe that the ad delivery differences are, indeed, even greater: the male visible and invisible images deliver to approximately 60% men, while the female visible and invisible images deliver to approximately 45% men. Moreover, the statistical significance of this experiment is even stronger, with a Z value over 10 for the ad delivery difference between the male invisible and female invisible ads.

4.2.4 Impact on real ads

We have observed that differences in the ad headline, text, and image can lead to dramatic difference in ad delivery, despite the bidding strategy and target audience of the advertiser remaining the same. However, all of our experiments thus far were on test ads where we

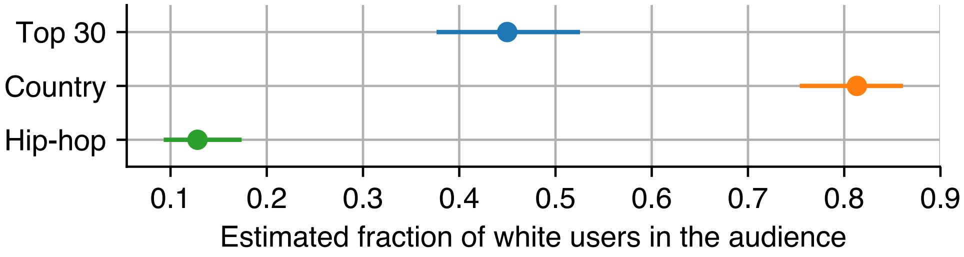


Figure 4.6: We run three campaigns about the best selling albums. *Top 30* is neutral, targeting all. *Country* implicitly targets white users, and *Hip-hop* implicitly targets Black users. Facebook classification picks up on the implicit targeting and shows it to the audience we would expect.

typically changed only a single variable. We now turn to examine the impact that ad delivery can have on realistic ads, where all properties of the ad creative can vary.

Entertainment ads We begin by constructing a series of benign entertainment ads that, while holding targeting parameters fixed (targeting custom audience C from Table 4.1, are stereotypically of interest to different races. Namely, we run three ads leading to lists of best albums in the previous year: general top 30 (neutral), top country music (stereotypically of interest mostly to white users), and top hip-hop albums (stereotypically of interest mostly to Black users). We find that Facebook ad delivery follows the stereotypical distribution, despite all ads being targeted in the same manner and using the same bidding strategy. Figure 4.6 shows the fraction of white users in the audience in the three different ads, treating race as a binary (Black users constitute the remaining fraction). Error bars represent 99% confidence intervals calculated using Equation 5.1.

Neutral ads are seen by a relatively balanced, 45% white audience, while the audiences receiving the country and hip-hop ads are 80% and 13% white, respectively. Assuming significant population level differences of preferences, it can be argued that this experiment highlights the "relevance" measures embedded in ad delivery working as intended. Next, we investigate cases where such differences may not be desired.

Employment ads Next, we advertise eleven different generic job types: artificial intelligence developer, doctor, janitor, lawyer, lumberjack, nurse, preschool teacher, restaurant

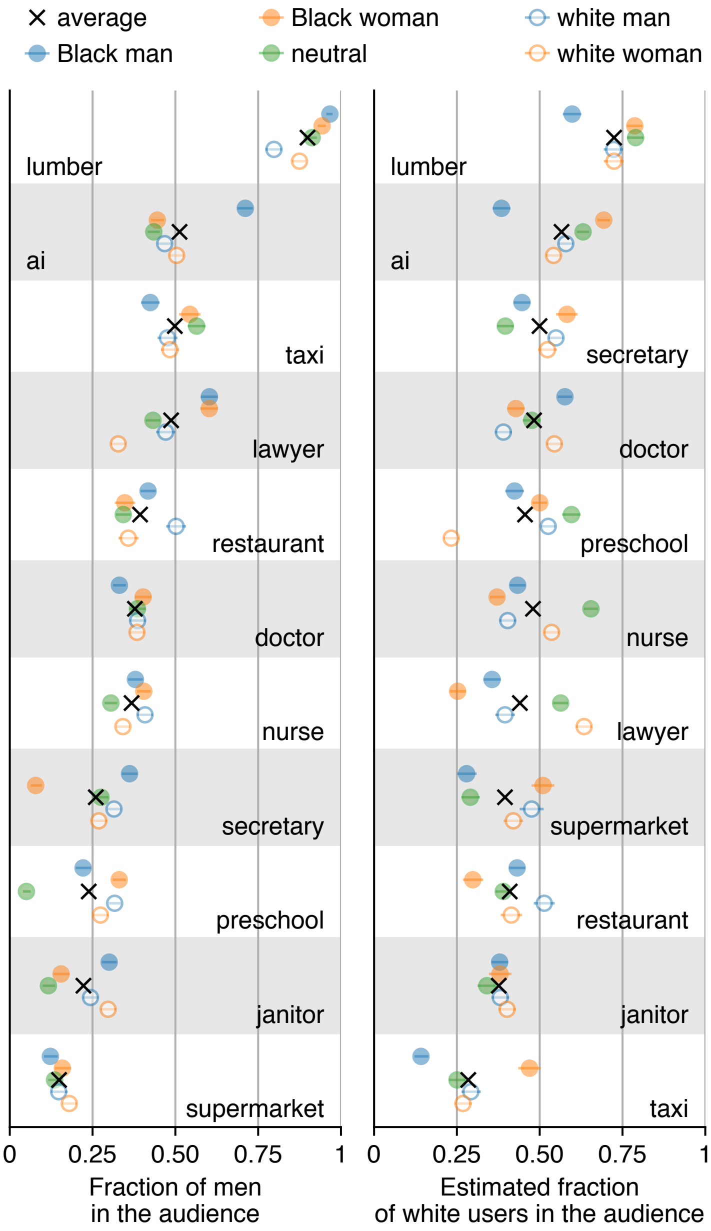


Figure 4.7: Results for employment ads, showing a breakdown of ad delivery by gender (left figure) and race (right figure) in the ultimate delivery audience. The labels refer to the race/gender of the person in the ad image (if any). The jobs themselves are ordered by the average fraction of men or white users in the audience. Despite the same bidding strategy, the same target audience, and being run at the same time, we observe significant skew along on both racial and gender lines due to the content of the ad alone.

cashier, secretary, supermarket clerk, and taxi driver. For each ad, we customize the text, headline, and image as a real employment ad would. For example, we advertise for taxi drivers with the text “Begin your career as a taxi driver or a chauffeur and get people to places on time.” For each ad, we link users to the appropriate category of job listings on a real-world job site.

When selecting the ad image for each job type, we select five different stock photo images: one that has a white male, one that has a white female, one that has a black male, one that has a black female, and one that is appropriate for the job type but has no people in it. We run each of these five independently to test a representative set of ads for each job type, looking to see how they are delivered along gender and racial lines (targeting custom audience C from Table 4.1). We run these ads for 24 hours, using the objective of Traffic, all targeting the same audience with the same bidding strategy.

The results of this experiment are presented in Figure 4.7, plotting the distribution of each of our ads along gender (left graph) and racial (right graph) lines. As before, the error bars represent the 99% confidence interval calculated using Eq. 5.1. We can immediately observe drastic differences in ad delivery across our ads along both racial and gender lines: our five ads for positions in the lumber industry deliver to over 90% men and to over 70% white users in aggregate, while our five ads for janitors deliver to over 65% women and over 75% black users in aggregate. Recall that the only difference between these ads are the ad creative and destination link; we (the advertiser) used the same bidding strategy and target audience, and ran all ads at the same time.

Furthermore, we note that the skew in delivery cannot merely be explained by possibly different levels of competition from other advertisers for white and Black users or for male and female users. Although it is the case that each user may be targeted by a different number of advertisers with varying bid levels, by virtue of running all of our job ads at the same time, targeting the same users, with the same budget, we are ensuring that our ads are experiencing competition from other advertisers identically. In other words, our ad targeting asks that every user who is considered for our “lumberjack” job ad should also be considered for our taxi driver job ad, so these ads should be competing with each other and facing identical competition from other advertisers at auction time. If the content of the ad was not taken into account by the delivery optimization system, then the ads would be expected to have similar—though not necessarily even—breakdowns by race and gender at delivery. Our experiment demonstrates that this is not the case, and thus, aspects of ad delivery optimization, rather than merely advertiser competition, influence the skew in the delivery outcome.

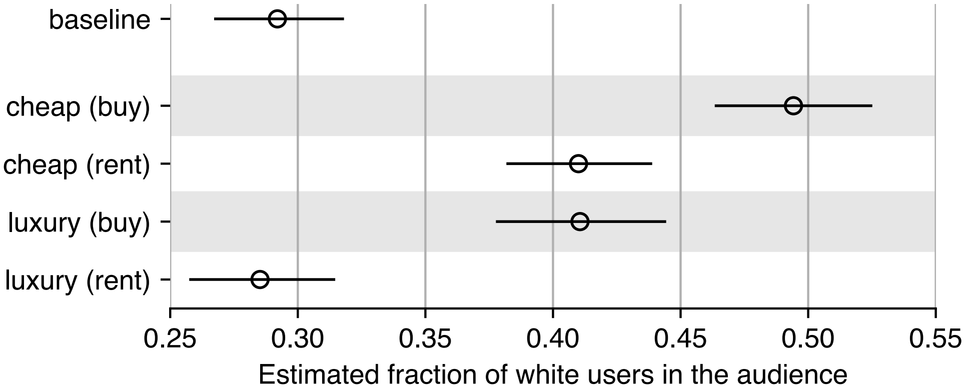


Figure 4.8: Results for housing ads, showing a breakdown in the ad delivery audience by race. Despite being targeted in the same manner, using the same bidding strategy, and being run at the same time, we observe significant skew in the makeup of the audience to whom the ad is delivered (ranging from estimated 27% white users for luxury rental ads to 49% for cheap house purchase ads).

Housing ads Finally, we create a suite of ads that advertise a variety of housing opportunities, as discrimination in online housing ads has recently been a source of concern [81]. We vary the type of property advertised (rental vs. purchase) and the implied cost (fixer-upper vs. luxury). In each ad, the cost is implied through wording of the ad as well as the accompanying image. Each ad points to a listing of houses for sale or rental apartments in North Carolina on a real-world housing site. Simultaneously, we ran a baseline ad with generic (non-housing) text that simply links to `google.com`. All of the ads ran for 12 hours, using the objective of Traffic, all targeting the same North Carolina audiences and using the same bidding strategy. We construct the experiment such that each particular ad is run twice: once targeting audience A and once targeting audience B (see Table 4.1) This way we eliminate any potential geographical effects (for example, users in Wilmington could be interested in cheap houses to buy, and users in Charlotte could be interested in luxury rentals regardless of their ethnicity, but if we only used audience C that effect could appear as racial skew).

We present the results in Figure 4.8 (we found little skew for the housing ads along gender lines, and we omit those results). We observe significant ad delivery skew along racial lines in the delivery of our ads, with certain ads delivering to an audience of over 72% Black users (comparable to the baseline results) while others delivering to an audience of as little as 51% Black users.

As with the employment ads, we cannot make claims about what particular properties of our ads lead to this skew, or about how housing ads in general are delivered. However, given the significant skew we observe with our suite of ads, it indicates the further study is needed to understand how real-world housing ads are delivered.

4.3 Discussion

To date, the public debate and ad platform's comments about discrimination in digital advertising have focused heavily on the targeting features offered by advertising platforms, and the ways that advertisers can misuse those features [60].

In this study, we set out to investigate a different question: *to what degree and by what means may advertising platforms themselves play a role in creating discriminatory outcomes?*

Our study offers an improved understanding of the mechanisms behind and impact of ad delivery, a process distinct from ad creation and targeting. While ad targeting is facilitated by an advertising platform—but nominally controlled by advertisers—ad delivery is conducted and controlled by the advertising platform itself. We demonstrate that, during the ad delivery phase, advertising platforms can play an independent, central role in creating skewed, and potentially discriminatory, outcomes. More concretely, we have:

- Replicated and affirmed prior research suggesting that market and pricing dynamics can create conditions that lead to differential outcomes, by showing that the lower the daily budget for an ad, the fewer women it is delivered to.

- Shown that Facebook's ad delivery process can significantly alter the audience the ad is delivered to compared to the one intended by the advertiser based on the content of the ad itself. We used public voter record data to demonstrate that broadly and inclusively targeted ads can end up being differentially delivered to specific audience segments, even when we hold the budget and target audience constant.

- Demonstrated that skewed ad delivery can start at the beginning of an ad's run. We also showed that this process is likely automated on Facebook's side, and is not a reflection of the early feedback received from users in response to the ad, by using transparent images in ads that appear the same to humans but are distinguishable by automatic image classification tools, and showing they result in skewed delivery.

- Confirmed that skewed delivery can take place on real-world ads for housing and employment opportunities by running a series of employment ads and housing ads with the same targeting parameters and bidding strategy. Despite differing only in the ad creative and destination link, we observed skewed delivery along racial and gender lines.

We briefly discuss some limitations of our work and touch on the broader implications of our findings.

Limitations It is important to note that while we have revealed certain aspects of how ad delivery is accomplished, and the effects it had on our experimental ad campaigns, we cannot make broad conclusions about how it impacts ads more generally. For example, we observe that all of *our ads* for lumberjacks deliver to an audience of primarily white and male users, but that may not hold true of *all ads* for lumberjacks. However, the significant ad delivery skew that we observe for our employment and housing ads strongly suggests that such skew is present for such ads run by real-world advertisers.

Skew vs. discrimination Throughout this study we refer to differences in the demographics of reached audience as “skew” in delivery. We do not claim any observed skew *per se* is necessarily wrong or should be mitigated. Without making value judgements on skew in general, we do emphasize the distinct case of ads for housing and employment. In particular, the skew we observe in the delivery of ads for cosmetics or bodybuilding might be interpreted as reinforcing gender stereotypes but is unlikely to have legal implications. On the other hand, the skew in delivery of employment and housing ads is potentially discriminatory in a legal sense.

Further, for the experiments involving ethnicity, we attempted to create equally sized audiences (50% white and 50% Black). However, solely the fact that ads are not delivered to an evenly split audience does not indicate skew, as there might be differences in matching rates (what fraction of registered voters are active Facebook users) per ethnicity, or the groups could have different temporal usage patterns. Only when we run two or more ads at the same time, targeting the same audience, and these ads are delivered with different proportions to white and Black users, do we claim we observe skew in delivery.

Our focus lies in understanding the extent to which the ad platform’s delivery optimization, rather than merely its targeting tools and their use as implied by Facebook [60], determine the outcomes of ad delivery, and on highlighting that demographic skews presently arise for certain legally protected categories in Facebook, even when the advertiser targets broadly and inclusively.

Skew in traditional media Showing ads to individuals most likely to engage with them is one of the cornerstone promises of online ad platforms. While in traditional media—such as newspapers and television—advertisers can also place their ads strategically to reach particular kinds of readers or viewers, there are three significant differences with implications for fairness and discrimination when compared to advertising on Facebook.

First, when advertising in traditional media, *the advertiser* has the ability to purposefully advertise to a wide and diverse audience, and be assured that their ads will reach that audience. As we show in this work, this is not the case for advertising on Facebook. Even if the advertiser intends to reach a general and diverse audience, their ad can be steered to a narrow slice within that specified audience, that is skewed in unexpected or undesirable ways.

Second, *the individual's* agency to see ads targeted at groups they do not belong to is more severely limited in the hyper-targeted and delivery-optimized scenario of online ad platforms. In traditional media, an individual interested in seeing ads targeted to a different demographic than they belong to has to merely watch programming or read a newspaper that they are not usually a target demographic for. On Facebook, finding out what ads one may be missing out on due to gender, race, or other characteristic inferred or predicted by Facebook is more challenging. A particularly motivated user could change their self-reported gender but might find themselves discouraged from doing so because the account's gender information is always public. Other characteristics, such as race and net worth, are inferred by Facebook (or accessed via third-party companies [244]) rather than obtained through user's self-reported data, which makes them challenging to alter for the purposes of seeing ads. Moreover, although users can remove some of their inferred interests using ad controls on Facebook, they have no ability to control *negative inferences* Facebook may be making about them. For example, Facebook may infer that a particular user is "not interested in working at a lumber yard", and therefore, not show this user ads for a lumberjack job even if the employer is trying to reach them. As a result, Facebook would be excluding them from an opportunity in ways unbeknownst to the user and to the advertiser.

Third, *public interest scrutiny* of the results of advertising is much more difficult in online delivery-optimized systems than in traditional media. Advertising in traditional media can be easily observed and analyzed by many members of society, from individuals to journalists, and targeting and delivery outside the expectation norms can be detected and called out by many. In the case of hyper-targeted online advertising whose delivery is controlled by the platform, such scrutiny is currently outside reach for most ads [88, 176].

Policy implications Our findings underscore the need for policymakers and platforms to carefully consider the role of the optimizations run by the platforms themselves—and not just the targeting choices of advertisers—in seeking to prevent discrimination in digital advertising.

First, because discrimination can arise in ad delivery independently from ad targeting, limitations on ad targeting—such as those currently deployed by Facebook to limit the targeting features that can be used—will not address discrimination arising from ad delivery. On the contrary, to the extent limiting ad targeting features prompts advertisers to rely on larger target audiences, the mechanisms of ad delivery will have an even greater practical impact on the ads that users see.

Second, regulators, lawmakers, and platforms themselves will need to more deeply consider whether and how longstanding civil rights laws apply to modern advertising platforms in light of ad delivery dynamics. At a high level, U.S. federal law prohibits discrimination in the marketing of housing, employment and credit opportunities. A detailed consideration of these legal regimes is beyond the scope of this study. However, our findings show that ad platforms themselves can shape access to information about important life opportunities in ways that might present a challenge to equal opportunity goals.

Third, in the U.S., Section 230 of the Communications Decency Act (CDA) provides broad legal immunity for internet platforms acting as publishers of third-party content. This immunity was a central issue in recently-settled litigation against Facebook, who argued its ad platform should be protected by CDA Section 230 in part because its advertisers are "wholly responsible for deciding where, how, and when to publish their ads." [82] Our research shows that this claim is misleading, particularly in light of Facebook's role in determining the ad delivery outcomes. Even absent unlawful behavior by advertisers, our research demonstrates that Facebook's own, independent actions during the delivery phase are crucial to determining how, when, and to whom ads are shown, and might produce unlawful outcomes. These effects can be invisible to, and might even create liability for, Facebook's advertisers.

Thus, the effects we observed could introduce new liability for Facebook. In determining whether Section 230 protections apply, courts consider whether an internet platform "materially contributes" to the alleged illegal conduct. Courts have yet to squarely consider how the delivery mechanisms described in this study might affect an ad platform's immunity under Section 230.

Fourth, our results emphasize the need for increased transparency into advertising platforms, particularly around ad delivery algorithms and statistics for real-world housing, credit, or employment ads. Facebook's existing ad transparency efforts are not yet sufficient to allow researchers to analyze the impact of ad delivery in the real world.

Potential mitigations Given the potential impact that discriminatory ad delivery can have on exposure to opportunities available to different populations, a natural question is how ad platforms such as Facebook may mitigate these effects. This is not straightforward, and is likely to require increased commitment and transparency from ad platforms as well as development of new algorithmic and machine learning techniques. For instance, as we have demonstrated empirically in Section 4.2.1 (and as [61] have shown theoretically), skewed ad delivery can occur even if the ad platform refrains from refining the audience supplied by the advertisers according to the predicted relevance of the ad to individual users. This happens because different users are valued differently by advertisers, which, in a setting of limited user attention, leads to a tension between providing a useful service for users and advertisers, fair ad delivery, and the platform's own revenue goals; a formal statement of this claim for the theoretical notions of individual fairness [62] and its generalization, preference-informed fairness, can be found in [135].

Thus, more advanced and nuanced approaches to addressing the potential issues of discrimination in digital advertising are necessary. Developing them is beyond the scope of this study; however, we can imagine several different options, each with their own pros and cons. First, Facebook and similar platforms could disable optimization altogether for some ads, and deliver them to a random sample of users within an advertiser's target audience (with or without market considerations). Second, platforms could remove ads in protected categories from their normal ad flows altogether, and provide those listings in a separate kind of marketing product (e.g., , a user-directed listing service like `craigslist.org`). Third, the platforms could allow the advertisers to enforce their own demographic outcomes so long as those desired outcomes don't themselves violate anti-discrimination laws. Finally, the platforms could seek to constrain their delivery optimization algorithms to satisfy chosen fairness criteria (some candidates for such criteria from the theoretical computer science community are individual fairness [62] and preference-informed fairness [135], but a broader discussion of appropriate criteria involving policymakers is needed).

Digital advertising increasingly influences how people are exposed to the world and its opportunities, and helps keep online services free of monetary cost. At the same time, its potential for negative impacts, through optimization due to ad delivery, is growing. Lawmakers, regulators, and the ad platforms themselves need to address these issues head-on.

Chapter 5

Ideological Biases in Political Ads

The results in Chapter 4 show that Facebook's choices during the ad delivery phase—where user personalization happens—can lead to dramatic skews in delivery along gender and racial lines, even when the advertiser aims to reach gender- and race-balanced audiences. In this chapter, we investigate how political ads could be similarly skewed across political leaning (Democrat or Republican in the context of the U.S.) by Facebook. Unlike job or housing ads, political ads do not have legal protections, and a disparity in distribution does not constitute discrimination. However, limiting the reach of such content poses a different form of harm, political candidates trying to reach disagreeing audiences could see their messages stifled by personalization. In aggregate, such decisions could reinforce political filter bubbles [191] for the platform's users. We continue our focus on Facebook as our platform, but our methods are extensible to other platforms that employ personalization as well. The following discussion is from our published work [9] at the *ACM Conference On Web Search and Data Mining (WSDM)*, 2021.

According to Facebook's Ad Library [77], political campaigns have spent over $907M on Facebook ads worldwide since May 2018, with the Trump campaign alone currently spending over $1M each week [232]. Furthermore, a recent study found that, at the state level, ”more than 10 times as many candidates advertise on Facebook than advertise on TV” [98]. This adoption reflects the fact that social media platforms have substantially lowered the cost of advertising, expanding the number of campaigns who can feasibly reach voters through digital channels [98]. Given the growing importance of online ads to the political discourse, it is critical to understand how complex ad platforms like Facebook operate in practice.

Much attention has been had to the *ad creation and targeting* phase, where the advertiser selects their desired audience and uploads their ad creative. Researchers have shown that advanced targeting features on ad platforms can be used to prevent certain ethnic groups from seeing ads [16, 222]. For example, in 2016, the Trump campaign used these techniques to carry out ”major voter suppression operations” aimed at lowering turnout among young women and black voters [110], and there is evidence that Russian organizations used these tools interfere with 2016 U.S. presidential elections [203, 235].

However, the influence of the ad delivery phase in skewing the reach of political ads. As far as we are aware, we are the first to study whether such skews are introduced for political ads on real-world advertising platforms due to the ad delivery phase alone. We focus on Facebook because of its critical importance to today's digital political advertising. We hypothesize that that Facebook may choose to deliver ads only to the subset of the political campaign's target audience that it predicts will be aligned with a campaign's views, *despite* attempts by the campaign to reach a diverse range of voters, and that this practice might play a role in political polarization by creating informational filter bubbles. Specifically, we seek to answer: *Is a political campaign advertising on Facebook able to reach all of the electorate?* Or, *is Facebook preferentially delivering ads to users who it believes are more likely to be aligned with the campaign's political views?* Additionally, *does Facebook vary ad pricing based on its hypothesized match between the target audience and campaign's political views?*

These questions are particularly urgent in light of the debate unfolding over the “microtargeting” of political ads. In late October 2019, Twitter decided to change its policy and ban all political advertising on its platform [137]. In response, U.S. Federal Election Commission (FEC) Chair Ellen Weintraub publicly argued that instead of banning such ads, ad platforms should limit political advertisers' ability to narrowly target ads to ensure that ”a broad public can hear the speech and respond” [249]. Shortly after, Google announced that it will significantly limit election ad targeting in order to ”promote increased visibility of election ads” [108]. At the time of this writing, Facebook is considering reforms of its own ad platform, but details are sparse [104].

Our questions regarding ad delivery are important to these developments for at least two reasons. *First*, skews resulting from ad delivery can raise fundamentally similar concerns to those raised about narrow targeting: an electorate who cannot "hear and respond" to political speech. *Second*, ad delivery algorithms might counteract the goals of restricting microtargeting by redirecting ads according to the choices of the ad platforms (in spite of broader target audiences). Policymakers must be alert to these implications.

To test our hypotheses, we became a political advertiser and ran over $13K[1] of political ads under controlled conditions, and observed how Facebook's algorithms delivered them. Unfortunately, Facebook makes it difficult to understand ad delivery along axes of political affiliation; to measure these results, we had to design careful experiments. *First*, we needed to determine *which users* Facebook was delivering our ads to, and whether skew (along political lines) exists among these users. We re-used techniques published in prior work [8, 222], using proxies based on ground-truth data from the voter records and political donation records. Additionally, we created audiences according to their political leaning—as inferred by Facebook—and used them simultaneously as part of an ad campaign but in a way that we can explicitly see the delivery to each subgroup. *Second*, we needed to determine if it is *possible* for a candidate to reach their entire audience. We used long-running ads, along with the Facebook-provided limits on how frequently a given set of ads can be shown to a user, to "force" the platform to consider delivering our ads to all of our targeted users. This way, we were able to "exhaust" the audience to determine how much of it the platform will allow a given message to be shown to. *Third*, we needed to determine how we were being charged for delivering ads to different sub-populations of the target audience. We used Facebook's advertising reporting features, combined with proxies, in order to understand how our budget is split across users with different political leanings.

Contributions After running our ads and analyzing the results, we present the following contributions:

First, we show that, despite identical targeting parameters, budgets, and competition from other advertisers, the content of a political ad alone can significantly affect which users Facebook will show the ad to. For example, we find Facebook delivers our ads with content from Democratic campaigns to over 65% users registered as Democrats, while delivering ads

[1] Throughout the study we refer to prices in U.S. Dollars.

from Republican campaigns to under 40% users registered as Democrats, despite identical targeting parameters. Moreover, our "control" ads with neutral political content that are run at the same time are delivered to a much more balanced audience (47% Democrats), showing that preferentially delivery is a result of Facebook's ad delivery algorithm.

Second, we find that this effect is surprisingly not present when we target users who *donated* to political campaigns, rather than those who are registered for a given political party.

Third, we find that that the delivery skew is present to an even greater degree when we use Facebook's own political targeting features. For example, when we target an audience of users who Facebook believes have "Likely engagement with US political content (Liberal)", combined with an equal-sized audience who Facebook considers to have "Likely engagement with US political content (Conservative)", we find that our ads from Democratic campaigns deliver to over 60% liberal users (compared to ads from Republican campaigns, which deliver to 25% liberal users).

Fourth, we find that it can be difficult and more expensive for political campaigns to have their content delivered to those who Facebook believes are not aligned with the campaign's views. For example, when re-running ads for Bernie Sanders (a liberal candidate) and Donald Trump (a conservative candidate), we find that when targeting an audience of conservative users, in the first day of the ad campaign, Facebook delivers our Sanders ad to only 4,772 users, while our Trump ad is delivered to 7,588 users.[2] We find that the underlying reason is that our Sanders ads targeting conservative users are charged significantly more by Facebook than our Trump ads ($15.39 versus $10.98 for 1,000 impressions), despite being run from the same ad account, at the same time.[3] Moreover, the difference cannot be attributed to some unknown underlying difference in liberal and conservative users' use of Facebook, as our neutral ad targeting the same audiences and run at the same time is delivered much more uniformly, reaching 6,822 liberal and 6,584 conservative users at a cost of $12.07 and $12.65 for 1,000 impressions, respectively.

Fifth, we find that when an ad creative and landing page shown to the users is neutral, but we "trick" Facebook's algorithm into believing the ad leads to a page with content taken from

[2] We find a similar, but flipped, effect if we target an audience of liberal users.

[3] Again, we see a similar, flipped effect when targeting liberal users.

a particular candidate's campaign web site, the skews in delivery and differential pricing are also present. This suggests that the ad delivery decisions made by Facebook are not driven exclusively by user reactions to the ad (as all such ads appeared identical to the users), but instead are made at least partially *a priori* by Facebook itself.

Taken together, our results indicate that Facebook preferentially shows users political ads whose content it predicts are aligned with their political views, with negative implications for both users and campaigns. For users, such delivery limits users' exposure to diverse viewpoints and—if Facebook's inference is incorrect—may pigeonhole them into a particular slice of political ads. For campaigns, such delivery may inhibit them from reaching beyond their existing "base" on Facebook, as getting ads delivered to users the platform believes are not aligned with their views may become prohibitively expensive. Importantly, these effects may be occurring without users' or campaigns' knowledge or control.

Stepping back, our findings raise serious concerns about whether Facebook and similar ad platforms are, in fact, *amplifying* political filter bubbles by economically disincentivizing content they believe are not aligned with users' political views. Put simply, Facebook is making decisions about which ads to show to which users based on its own priorities (presumably, user engagement with or value for the platform). But in the context of political advertising, Facebook's choice may have significant negative externalities on political discourse in society at large.

Ethics All of our experiments were conducted with careful consideration of ethics. *First*, we obtained Institutional Review Board review of our study, with our protocol being marked as "Exempt". We did not collect any users' personally identifying information from Facebook, and did not collect any information about users who visited our site after clicking on our ads. *Second*, we minimized harm to Facebook users when running our ads by only running "real" ads, i.e., if a user clicked on one of our ads, they were brought to a real-world page not under our control that was relevant to the topic of the ad. In the few cases where the ads pointed to a domain we controlled, the visiting users were automatically and immediately redirected to a real page that we did not control. *Third*, we minimized harm to Facebook itself by participating in their advertising system as any other advertiser would and paying for all of our ads. We registered as an advertiser in the area of "Social Issues, Elections or

Politics" [39], meaning our ads were subject to the same review as the ads of other political campaigns. *Fourth,* we minimized the risk of altering the political discourse through careful choices of the ad content (Section 5.1.2), and running approximately the same number of copies of ads for Republican and Democratic candidates, with the same budgets. The total amount we spent on political advertising while collecting data for this study ($13.7K) is minuscule compared to the ad budgets of real campaigns in the same period (likely in the millions of dollars [77]).

Limitations It is important to note the limitations of our study (see Section 5.2.3 for a detailed discussion). Most importantly, we can only report results of *our own ads*; we are unable to make any statements about the extent to which any effects we observe exist for political ads run by real political campaigns, or political ads *in general.* However, the fact that we observe statistically significant skews in our small set of ads suggests that the effects we observed are likely present in the delivery of other political ads as well.

The rest of this chapter is organized as follows—Section 5.1 gives an overview of our methodology, Section 5.2 presents the results of our experiments, and Section 5.3 offers a concluding discussion.

Table 5.1 Number of uploaded records for Custom Audiences created using publicly available voter records. We divide the DMAs in the state into two sets, and create two audiences, each with voters registered with one party per DMA set (CA$_A$ and CA$_B$). We repeated this process with separate voter records (creating CA$_C$ and CA$_D$), allowing us to run experiments on separate audiences. The number of uploaded records does not match, as we uploaded records so that the Estimated Daily Reach was the same.

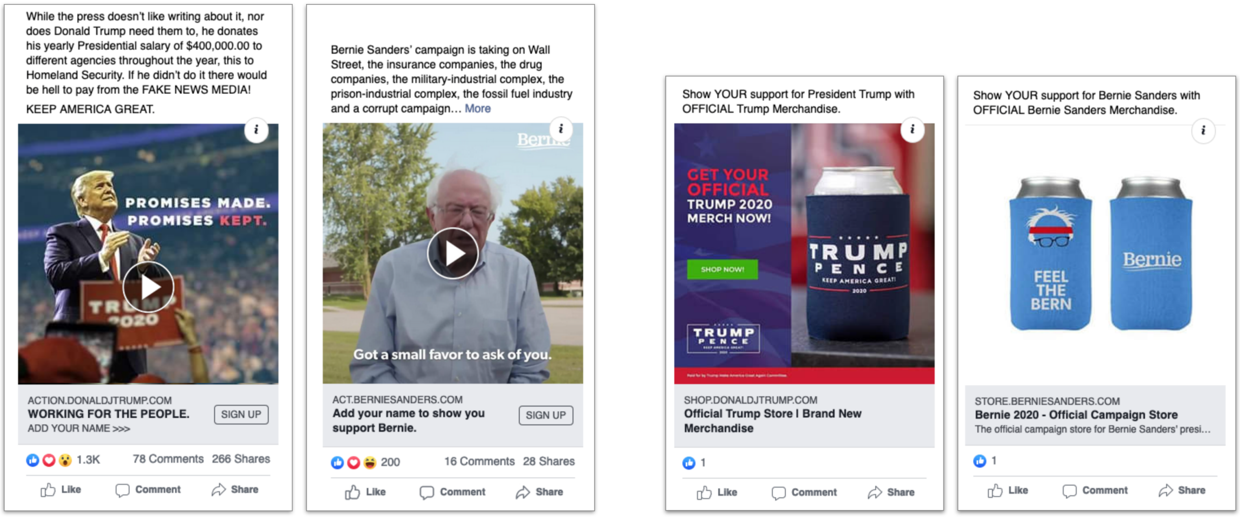
DMA(s) [182]	**CA$_A$**		**CA$_B$**		**CA$_C$**		**CA$_D$**	
	Dem	**Rep**	**Dem**	**Rep**	**Dem**	**Rep**	**Dem**	**Rep**
Greensboro, Charlotte	70,000	0	0	70,000	70,000	0	0	70,000
Wilmington, Raleigh-Durham, Greenville-(New Bern and Spartanburg)	0	63,137	70,000	0	0	54,000	64,166	0

5.1 Methods

In this work we aim to answer two related, but separate questions, and design our experiments accordingly. *First,* we want to verify whether the skew in delivery reported in previous work [8] exists along the lines of political affiliation for political ads. To this end, we replicate the study setup from [8] as closely as possible, including setting the campaign objective to “Traffic”. *Second,* we ask whether a political campaign determined to reach users who may not be aligned with its views—and explicitly requesting such audience from Facebook—can achieve their goal. To be able to better answer this question, we set the campaign objective to “Reach”.

For the sake of clarity, we run ads for only one Democratic presidential candidate (Bernie Sanders) and compare their performance to that of the ads for only one Republican candidate (Donald Trump). We choose these two candidates because at the time of experiment design (early July 2019), they had spent most on Facebook advertising among the major candidates of each party [77]. Therefore, their election performance is least likely to be influenced by ads run on our limited budget.

Next, we provide more details on the audiences and ad campaigns in our experiments, how we measured their performance over time, and the statistical apparatus necessary to interpret the results.



While the press doesn't like writing about it, nor does Donald Trump need them to, he donates his yearly Presidential salary of $400,000.00 to different agencies throughout the year, this to Homeland Security. If he didn't do it there would be hell to pay from the FAKE NEWS MEDIA!

KEEP AMERICA GREAT.

ACTION.DONALDJTRUMP.COM
WORKING FOR THE PEOPLE.
ADD YOUR NAME >>>

Bernie Sanders' campaign is taking on Wall Street, the insurance companies, the drug companies, the military-industrial complex, the prison-industrial complex, the fossil fuel industry and a corrupt campaign… More

ACT.BERNIESANDERS.COM
Add your name to show you support Bernie.

Show YOUR support for President Trump with OFFICIAL Trump Merchandise.

SHOP.DONALDJTRUMP.COM
Official Trump Store I Brand New Merchandise

Show YOUR support for Bernie Sanders with OFFICIAL Bernie Sanders Merchandise.

STORE.BERNIESANDERS.COM
Bernie 2020 - Official Campaign Store
The official campaign store for Bernie Sanders' presi…

Figure 5.1: Ads used in our experiments concerning political issues and promoting candidates' merchandise. The ads were copies of real ads run on Facebook by the official campaigns, with the exception for Bernie Sanders related merchandise (as his store has no official Facebook advertising).

5.1.1 Creating audiences

We use two mechanisms for targeting audiences on Facebook: Custom Audiences and detailed targeting.

Recall that we are interested in studying the skew in delivery along political lines. Since Facebook does not provide ad delivery breakdowns by political leaning, but does provide breakdowns by location (Section 2.4.5), we craft our Custom Audiences in such a way that the statistics about the political leanings of the actual recipients of an ad can be inferred from the statistics about their location. Specifically, we follow the method introduced by prior work [8, 222].

Custom Audiences from voter records We obtained publicly available voter records from North Carolina, which, in addition to PII, include each voter's political party registration (if one exists). We then create Custom Audiences as follows:

1. Divide the DMAs in North Carolina into two sets of roughly equal population sizes.

2. Create a Custom Audience that contains PII of registered Democrats from the first set

of DMAs and another Custom Audience with PII of registered Republicans from the second.

3. Upload the lists to Facebook, and compare their "Estimated Daily Reach" statistics provided by Facebook.

4. If one audience has a higher Estimated Daily Reach, subsample it and re-upload until the estimates match.

5. Repeat steps (2)–(4) with the opposite assignment: registered Republicans from the first set of DMAs and registered Democrats from the second set of DMAs.

Table 5.1 shows the summary statistics of the audiences created from voter records.

There are two distinctions between our method and previous work [8]. *First*, we introduce an additional step in an attempt to create audiences consisting of roughly equal numbers of Democrats and Republicans. We do so for ease and clarity of subsequent analysis; it is not strictly necessary to do so, as we will compare the delivery of two ads *targeting the same audience and run at the same time* to observe differences due to ad delivery (thus, any differences in population, usage, or time of day will affect both campaigns equally). *Second*, Facebook no longer provides advertisers with the number of uploaded records that match Facebook users, as these estimates have been shown to leak private information about individuals [243, 245]. Instead, we use the Estimated Daily Reach provided by Facebook with a budget set to a very high number (e.g., $1M/day), thereby obtaining an estimate of the *total* daily active users in the uploaded audience.

Custom Audiences from donor records We build separate Custom Audiences from publicly available donor records for political campaigns. The U.S. Federal Election Commission (FEC), in particular, makes publicly available the PII of all contributors who have donated a total of $200 or more towards a political campaign [92]. We obtain data from the FEC for individuals who have donated to the "Bernie 2020" or the "Donald J. Trump for President" committees as of July 1, 2019 to craft Custom Audiences of users who actively engage with these campaigns. Because there are fewer donors for Bernie Sanders than for Donald Trump in FEC data, we also use mid-year FEC filing by ActBlue, a popular Democratic fundraising platform [13, 202] to obtain a list of Democratic donors who donated less than $200. Since the

Table 5.2 Overview of Custom Audiences built from public FEC donor records and ActBlue. The number of uploaded records does not match, as we uploaded records so that the Estimated Daily Reach was the same.

DMAs [182]	**CA$_E$**		**CA$_F$**	
	Trump donors	**Sanders donors**	**Trump donors**	**Sanders donors**
DMA Set 1	40,973	0	0	32,000
DMA Set 2	0	32,000	41,458	0

FEC data isn't limited to a particular state or region, we randomly split all 210 U.S. DMA regions [182] into two sets, and then create Custom Audiences of approximately equal numbers of Democratic and Republican donors in each, by relying on the estimated daily reaches (as in Steps (2)–(4)) above. Table 5.2 shows the size and configuration of our audiences created from donor records.

Detailed targeting audiences Although the ability to create Custom Audiences is only granted to advertisers with some history of running and paying for ads (the exact eligibility criteria are not publicly disclosed), all advertisers can specify their audience using detailed targeting (Section 2.4.2). We create a number of audiences this way, selecting a geographic region centered around a town and Facebook's inferred characterization such as "Likely engagement with US political content (Conservative)" and "Likely engagement with US political content (Liberal)". For some of the audiences we further narrowed the targeting by specifying additional required characteristics such as those who are, according to Facebook's characterization, "interested in" topics such as "Donald Trump for President", "Make America Great Again", "Bernie Sanders", or "Elizabeth Warren". We aimed to approximately match the sizes of liberal and conservative audiences for each geographic region by adjusting the targeting radius around a chosen location until the Estimated Daily Reach matches. The Appendix presents the details and size statistics for these audiences.

5.1.2 Creating ad copies

We ran three types of ads throughout our campaigns: (1) merchandise ads for candidates that link to the candidates' online campaign stores, (2) "issues ads" that have detailed content and that link to the candidates' websites, and (3) "neutral" political ads that simply encourage users to vote and link to generic voting information websites.[4] The majority of ads we ran were replicated from real ads run by official political campaigns obtained from the Facebook Ad Library [77]. Ads for Bernie Sanders' merchandise store were the only exception, as—unlike the other campaigns in question—the Bernie Sanders campaign had not advertised merchandise on Facebook; we created the ad creative for this ad. Whenever the replicated ad was written in the first person, we changed it to be a third person reference to the name of the candidate (as we were not running the ads *as* the campaign itself). Examples of the ad copies of types (1) and (2) that we ran are presented in Figure 5.1.

5.1.3 Isolating role of content

Most of our ads link directly to either a candidate's official website or generic voting information websites. In one of the experiments, however, we wanted to isolate the effect that the content of the advertised website has on the delivery skew, while keeping the users' reactions to the ad (such as possible Likes, comments, or reactions) constant. We found that during ad creation, Facebook would automatically crawl the destination link as part of the ad review and classification process. We develop a methodology that would use this feature to create ads that look like they have the same content to users, but different content to Facebook.

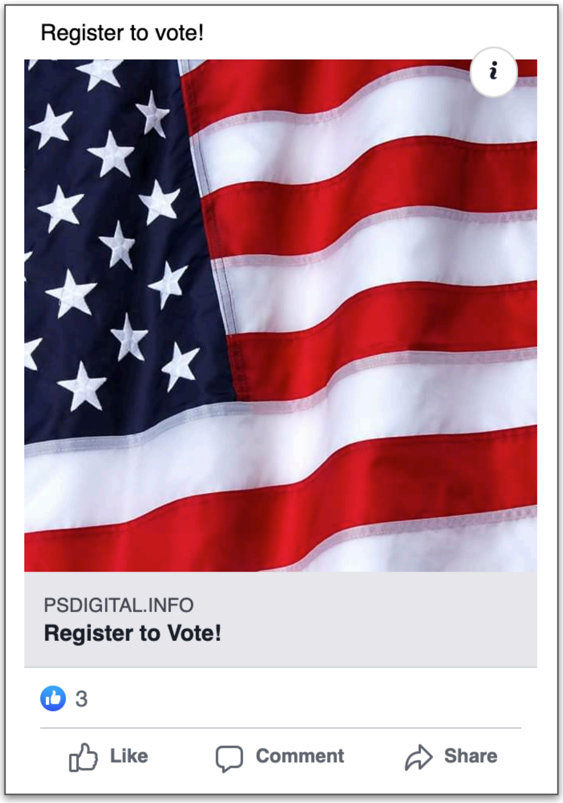
To this end, we created a generic ad with a call to register to vote, a picture of the American flag, and a link to a nondescript domain: `psdigital.info` (see Figure 5.2). We created three copies of this ad, with each copy having a destination link to a *different page* under that domain. We configured our web server to deliver a different response for requests for these pages based on the IP address of the requestor. If the requestor was a Facebook-owned[5] IP address, we served a copy of the HTML[6] from the official Trump campaign website, the official Sanders campaign website, or a generic voting information website,[7] depending on

[4] https://www.usa.gov/election and https://www.usa.gov/register-to-vote

[5] We determined Facebook IP addresses by using the IP address blocks advertised by Autonomous Systems numbers owned by Facebook.

[6] Only the HTML code was served from our server; we modified the HTML so that all images, JavaScript, and stylesheets would be downloaded from the corresponding official websites.

[7] https://www.usa.gov/register-to-vote



Register to vote!

PSDIGITAL.INFO

Register to Vote!

Figure 5.2: Ads that have a destination link to our webserver (`psdigital.info`), which serves HTML from the candidate webpages to requests from Facebook's IP addresses, but redirects all other traffic to a generic voting information site. The ads look identical to users, but different to Facebook.

the particular page under our domain requested. Otherwise, if the requestor was from any other IP address, the user would be immediately redirected to the generic voting information website. In this way, all three ads would appear identical to users (and those users would all be brought to the same voting information site if they clicked on the ad), but Facebook's algorithm believed they linked to pages with different political content.

5.1.4 Collecting performance statistics

As mentioned in Section 2.4.5, Facebook provides semi-live statistics on how the ad is delivering. Once an ad starts running, we query Facebook every five minutes in order to get these statistics over the lifetime of the ad. For ads where we use Custom Audiences with

DMAs as a proxy for political leaning, we request these delivery statistics be broken down by DMA.

5.1.5 Statistical analysis

The core questions in this work revolve around comparisons of the fractions of Democrats (or Republicans) among the users exposed to two ads that differ in their content. The comparison process consists of two steps and is based on previous work [8].

First, we estimate the fraction of Democrats in each ad and the 99% confidence interval around that estimate as shown in Equation (5.1):

$$\begin{aligned} L.L. &= \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n}, \\ U.L. &= \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + z^2_{\alpha/2}/n}, \end{aligned} \tag{5.1}$$

where $L.L.$ is the lower confidence limit, $U.L.$ is the upper confidence limit, $\hat{p}$ is the observed fraction of Democrats in the audience, n is the total size of the audience exposed to the ad. To obtain the 99% interval we set $z_{\alpha/2} = 2.576$.

Second, we compare whether the fractions in two scenarios are statistically significantly different. If their confidence intervals do not overlap (easily judged visually from the subsequent figures), the difference is statistically significant. If the intervals do overlap, we need to perform a difference of proportion test as shown in Equation (5.2):

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - 0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \tag{5.2}$$

where $\hat{p_1}$ and $\hat{p_2}$ are the fractions of Democrats in the two audiences, n_1 and n_2 are the total sizes of these audiences, and $\hat{p}$ is the fraction of Democrats in the two audiences combined. If the resulting Z-score is above 2.576 (corresponding to 99% confidence) the difference in proportion is statistically significant.

5.2 Experiments and Results

We now present the detailed set-ups and results of our experiments. Recall that our aim is to study both (a) whether the content of a political campaign's ad could lead to skew in delivery along political lines, and, if so, (b) whether a political campaign can successfully reach users who Facebook believes are not aligned with the campaign's views. In the two subsections below, we address each of these questions in turn, before discussing the implications and limitations of our study.

5.2.1 Ad content and skew

We begin by examining whether the content of an ad can lead to skew in delivery along political lines.

Voter records Similar to methodology of prior work [8] for studying skews along race and gender, we use the Custom Audiences CA_A, CA_B, CA_C, and CA_D described in Table 5.1 that are based on voter records. These audiences are designed so that asking Facebook to report delivery statistics by DMA serves as a proxy for obtaining delivery statistics by political affiliation.

We create three ad creatives: one taken from the official Donald Trump campaign, another from the Bernie Sanders campaign (both found in Facebook's Ad Library [77], shown in Figure 5.1 and linking to the respective campaign's web site), and a "neutral" political ad that simply encourages users to vote and links to a generic election website.[8] We then run one copy of each ad targeting each of the four Custom Audiences, for a total of 12 individual ads. Our ads are run with a daily budget of $20 per ad set and use the objective "Traffic" and optimization "Link Clicks" (Section 2.4.1).

Figure 5.3 (top row) presents the overall delivery statistics for these three ads, with the delivery statistics of all four instances of each ad aggregated together. We can immediately observe significant differences in delivery: The neutral ad delivers to 47% Democrats, while the Trump ad delivers to less than 40% Democrats. The Sanders ad, on the other hand, delivers to almost 70% Democrats. Note that this difference in delivery is despite the fact that

[8] https://www.usa.gov/election

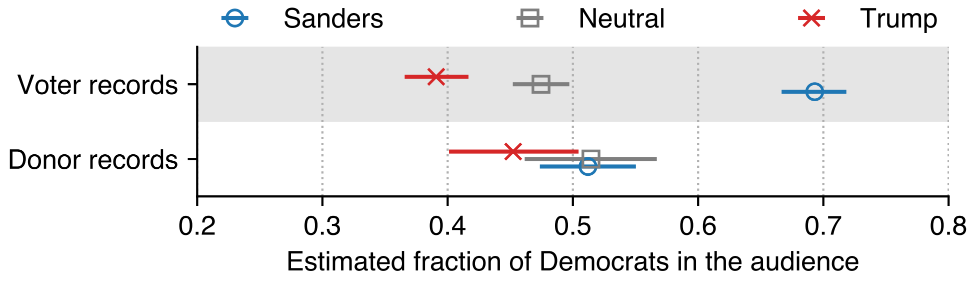


Figure 5.3: The estimated fraction of Democrats who were shown our ads, targeting both registered voters in North Carolina and political donor records. In the case of voter records, the ad delivery to Democrats ranges from approximately 69% for Sanders' ad to only 39% for the Trump's ad. In the case of donor records, we do not see statistically significant differences in ad delivery.

all ads are run from the same ad account, at the same time, targeting the same audiences, and using the same goal, bidding strategy, and budget; *the only difference between them is the content and destination link of the ad.*

Donor records Having observed that delivery skew along political lines can occur due to the content of the ad, we next turn to examine whether that skew is amplified if we choose users who recently engaged with politics. In particular, we examine whether recent *donors* to political campaigns are estimated by Facebook to have greater relevance for our ads, when compared to users who are simply registered as Democratic or Republican voters. Thus, we use our Custom Audiences of donor records (CA_E and CA_F, described in Table 5.2); however, due to the limited size of the donor record databases, we are only able to run our experiment on two, and not four, audiences. Thus, we run six ads in the same manner as the experiment we just described.

The results of this experiment are presented in Figure 5.3 (bottom row). Surprisingly, we do not find statistically significant differences in the ad delivery between the three ads targeting political donors. While we can only speculate as to why we observe a skew with voter records but not with donor records, the absence of a skew for the donor record audiences might suggest that Facebook does not have sufficient information about these users to do accurate relevance estimation for political ads.

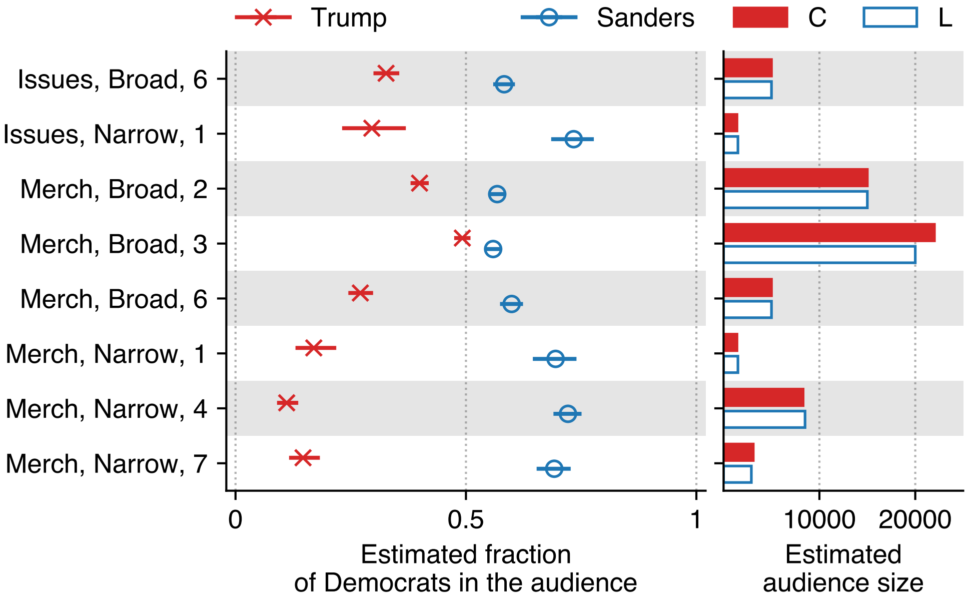


Figure 5.4: We ran merchandise and issue ads with two levels of targeting specificity (Broad: users with "Likely engagement with US political content (Conservative)" or "... (Liberal)"; Narrow: additional detailed targeting for inferred interest in Donald Trump or Bernie Sanders), and targeting different regions (1: Celina, OH; 2: Dutchess, NY; 3: Lorain, OH; 4: Macclenny, FL; 5: McCormick, SC; 6: Richlands, VA; 7: Saginaw, MI; 8: Slinger, WI). In all cases, Sanders' ads deliver to a larger fraction of Democrats than Trump ads even though they are targeting the same audiences at the same time using the same budgets. The effect is more pronounced for smaller audiences (compare, for example Merch, Broad, 3 and Merch, Broad, 6).

Detailed targeting To further explore the role of Facebook's use of inferences about its users in delivery and its impact on political ad skew, we next use audiences where we *know* that Facebook has inferred the political affiliation of its users. We do so using detailed targeting (Section 2.4.2), selecting attributes "Likely engagement with US political content (Conservative)" for one audience and "Likely engagement with US political content (Liberal)" for another. As discussed in Section 5.1.1, we then geographically limit our targeting to regions where we can ensure an approximately equal number of users in each audience (as shown by Facebook's audience size estimates). Then, over a course of six hours we concurrently ran two ad copies, each to two audiences (a total of four ads): one Sanders ad targeting liberal users and another targeting conservative users, and one Trump ad targeting liberal users and another targeting conservative users. For this, and all further experiments, we optimize for “Reach”, not “Traffic”. To calculate the delivery skew of a politician’s ads, we sum reach across the two audiences, and calculate the fraction of deliveries to the liberal audience.

The result of our first experiment is shown in the top row of Figure 5.4. We can immediately observe similar skews in delivery to the ones observed for voter records, with the content of the ad causing delivery skew along political lines. This indirectly suggests that our hypothesis for the reasons behind differences for voter vs donor records could have some merit.

Next, we explore this finding in depth, varying three aspects of our experiment:

1. The size of the audience, as reported by Facebook’s Estimated Daily Reach,
2. The “specificity” of the audience (narrowing the detailed targeting further by attributes such as users’ inferred interest in “Donald Trump for President” or “Bernie Sanders” according to Facebook), and
3. The specific topic of the ad (adding ads that advertise small campaign-branded merchandise that users can purchase, as shown in Figure 5.1).

The results for these experiments are shown in Figure 5.4 (remaining rows), with each row representing a separate experiment. Experiments described as “issues” are run with the first two ad creatives from Figure 5.1 and “merch” ads correspond to the third and fourth creative in Figure 5.1.

We make a number of observations from this experiment. *First*, we observe statistically significant skews in ad delivery along political lines for *all* of our ad configurations. This suggests that such skew is a pervasive property of Facebook’s ad delivery system. *Second*, we observe that the skews tend to be less pronounced when the ads are targeting larger audiences (more than 10,000 daily active users). While we do not know the underlying cause of this phenomenon, we hypothesize that the larger audiences provide the platform with a big enough pool of users to afford “relevant” users regardless of their inferred political leaning. On the other hand, we suspect that when running our ads with smaller audiences, Facebook “exhausts” the (small) subset of users in the non-aligned audience (e.g., Sanders advertising to a conservative audience) for whom Facebook believes the ad is, in fact, relevant, and thus pauses or raises the price for delivery, but continues the delivery among the aligned audience. We explore this hypothesis in more detail in the next experiment.

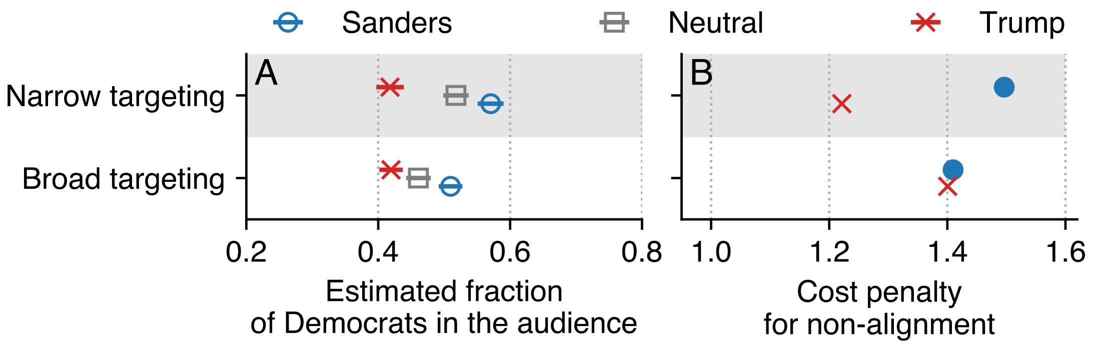


Figure 5.5: Delivery statistics for ads that look identical to users, but appear partisan to the Facebook classification mechanism. (A) The skew in delivery is consistent with that observed in visibly distinct ads. (B) There is a financial penalty for trying to show an ad that Facebook deems non-aligned; reaching the same number of people in the same audience is up to 1.5 times more expensive.

Overall, our findings strongly support our hypothesis that the content of an ad could lead to skew in its delivery along political lines whenever the platform has enough information (or thinks it has enough information) about the political leanings of the users being targeted, and that the skew is due to ad delivery optimization algorithms run by Facebook, rather than to other factors. As discussed in the introduction, this has profound implications for political advertisers, users, and society.

5.2.2 Longitudinal delivery

We now explore what happens if a political campaign aims to reach users who Facebook believes are not aligned with the campaign's views. Specifically, we "force" the Facebook ad platform to consider showing ads to all users in the political advertiser's targeting set, including users for whom Facebook may believe the ad is not relevant. We do so in two steps: *First*, on a small audience we measure whether skew appears even if the ads look the same to users but differently to Facebook. *Second*, we run near-copies of real ads of political campaigns on a larger audience to measure the total effect. In both cases, we configure the campaigns using the objective "Reach" and optimization "Reach", which enables us to tell Facebook to only show the ad once to each user each week, thus forcing the delivery

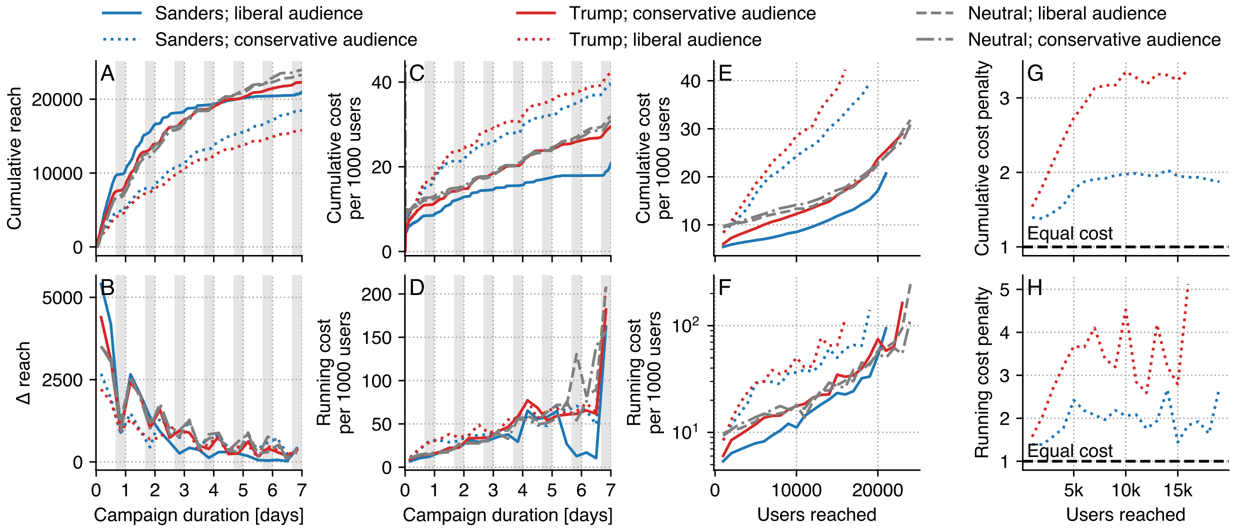


Figure 5.6: Ads for a political campaign deliver to more users and for a lower cost if the targeted users have the same partisanship. A and B - the delivery rates are the highest in the beginning of the ad runtime and for aligned audiences. C and D - the cost of reaching non-aligned audiences is higher, especially in the beginning of the experiment. E and F - the more people have already seen the ad, the more expensive it becomes to show it to even more people; that growth is log-linear (see F). G and H show the ratio between the cost of a political campaign advertising to a non-aligned audience and their competitor advertising to the same audience.

mechanism to 'exhaust' the audience rather than showing the ad to the same subset of users.

Generic ads We begin by applying the method described in Section 5.1.3, setting up ads that look identical to Facebook users (Figure 5.2) and, when clicked, redirect the user to a governmental website with instructions to register to vote. However, when visited by Facebook's web crawler, each ad's landing website shows different HTML: one serves Trump's campaign HTML, another Sanders', and a third a generic voting information site. Since all ads look identical to users, any skew can only be attributed to Facebook's optimization based on the content of the linked website.

Each ad copy is targeted at four audiences: two that are Broad and two that are Narrow. The Broad audiences target "Likely engagement with political content (Liberal)" and "...Conservative", and the Narrow audiences additionally target users with interest in Bernie Sanders and Donald Trump, respectively. Table 5.3 entries for Oxford, NC and

Scranton, PA provide detailed audience parameters and size statistics for the Broad and Narrow audiences respectively. Since we are attempting to reach everyone in the targeting set here, we set a higher budget $40 per day for each ad. These ads were run for two days and did not fully exhaust the audiences.

The results of this experiment are presented in Figure 5.5A. Even though the users see the same ad in all three cases, and therefore their explicit or implicit reactions to them are not more different than chance, the delivery is still skewed according to what Facebook's crawler sees. We also present the price differentiation in Figure 5.5B. For a given audience, we measure how much it cost for the aligned ad to reach the same number of users as the non-aligned ad did. For example, it cost 1.5× more for the ad linking to Sanders' campaign page (as perceived by Facebook) to reach the same number of users in the Broad conservative audience than the ad linking to Trump's campaign page. Conversely, it cost 1.2× more for the ad linking to Trump's campaign page to reach the same number of people in the Broad liberal audience than the ad linking to Sanders' campaign page.

These results show that the contents of the destination link—and not users' reaction or engagement with the ad—play a role in Facebook's decision for skewed delivery and differential pricing. An implication of this finding is that two campaigns running an ad about the same issue to the same target audience might reach different fractions of that audience and at different prices, only because the destination links are different. This differential delivery and pricing may be particularly damaging for local political campaigns, where candidates may agree on some issues but not others.

Real ads We now turn to explore how this effect plays out for real-world ads that differ in content and destination link. In this experiment, we run three ads (Trump, Sanders, and neutral issue ads as before), each to two Narrow audiences over a period of seven days and with a daily budget of $100 for each ad and audience combination.[9] The ad copies are the first two presented in Figure 5.1, and details about the target audiences are provided in Table 5.3 (the audiences from Michigan and Wisconsin). The conservative and liberal audiences were selected such that they had approximately the same daily active reach, and such that we expected our ads to have reached *almost everybody* in the audience by the end of the seven

[9] Our total spend over the week ended up being $4,228.19 distributed roughly equally among Sanders, Trump, and neutral ads.

days.

The results of this experiment are presented in Figure 5.6. We first focus on panel A, which shows the cumulative number of users reached over seven days (along with its derivative in panel B). We can observe that the delivery increases rapidly for all ads during the first day and then slows down quickly. However, we can observe two notable outliers in panel A: the Trump ad targeting the liberal audience and the Sanders ad targeting the conservative audience. Both of these non-aligned ads end up delivering to over 25% *fewer* users than their aligned counterparts. In other words, when the Trump ad is advertised to the conservative audience, it delivers to a total of 21,792 users; when the Sanders ad is run at the same time and targeted to the same conservative audience, it delivers to only 17,964 users. Note that this difference cannot be attributed to some unknown underlying difference between Facebook use between the users categorized as liberal and conservative by Facebook because the neutral ad delivers equally to liberals and conservatives, reaching approximately 23,000 users.

We turn to panel C, which shows the cumulative cost per thousand unique users to help explain why this is occurring. We can immediately notice an increasing cost trend for all ads: as the ads run longer, their cost increases substantially. Presumably, this is because Facebook first delivers the ad to the "cheaper" users in the target audience before deciding to spend our budget on the more "expensive" users. However, we can observe that the non-aligned ads are again outliers here: both show a substantially *higher* cost per thousand users, a difference noticeable from the first day of the experiment. By the end of the experiments, when the liberal ad is shown to the liberal audience, it is charged $21 per thousand users; when the conservative ad is delivered to the same audience, it is charged over $40 per thousand users.

Because the delivery rates slow down after the first day plots C and D make the growth of cost per 1,000 also appear to slow down. Therefore, we turn to plots E and F, which show this growth as a function of the size of reached audience, rather than time. We observe that the growth is rapid, and the running cost per 1,000 is growing exponentially as a function of the number of users reached. Finally, in plots G and H we show that the ratio between the cost a political campaign pays to show their ad to the non-aligned audience and the cost of their competitor showing to the same audience is relatively stable, between 2:1 and 4:1.

Overall, Figure 5.6 emphasizes three findings: *First*, the penalty for reaching a non-aligned audience remains at a relatively stable ratio between 2:1 and 4:1 as a function of audience already reached (see Figure 5.6H). *Second*, that while the cost per thousand viewers grows linearly with time (Figure 5.6D), it grows super-linearly with the number of users already reached (Figure 5.6F). For example, panel F shows that the cost of showing the Sanders ad to the first 1,000 liberal users (solid blue line) is approximately $5, and the cost of reaching the first 1,000 conservative users with this ad is approximately $10. However, once 10,000 users in each of these audiences are already reached, reaching another thousand of liberal users costs approximately $15 (a three-fold increase) and reaching another thousand of conservative users costs approximately $37 (nearly a four-fold increase compared to the first thousand conservatives). *Third*, at the end of the experiment, both neutral ads and the two aligned partisan ads reached over 20,000 users, while the non-aligned ads reached significantly fewer. This demonstrates the core phenomenon: it is cheaper and more effective for a political campaign to reach audiences that are politically aligned (as inferred by Facebook) with their agenda, and as the campaign progresses it becomes more expensive to reach additional viewers.

Finally, we run a similar experiment targeting Broad audiences (i.e., audiences without specific interest in candidates). The results of this experiment are presented in Figure 5.7. We find that the core phenomenon holds there as well, and note that in some cases, the conservative audience is cheaper for Trump than the liberal audience is for Sanders.

5.2.3 Limitations

We now discuss limitations of our study and briefly mention the steps we took to mitigate them when possible. Controlling for all possible variables that may affect political ad delivery is beyond the scope and financial capabilities of our work, and is better suited to be performed by Facebook itself or by an independent third-party auditor that would be granted broader data and algorithms' access than what is available through the ad interface. Similarly, it is important to note that we can only report on delivery skew that we observed for our own ads; we *cannot* draw any conclusions about how political ads in general (or all ads run by a particular campaign) are delivered. Nonetheless, the fact that we observe strong and statistically significant effects in our small set of ads suggests that the potential negative outcomes for individuals, political campaigns, and society in the context of ad delivery

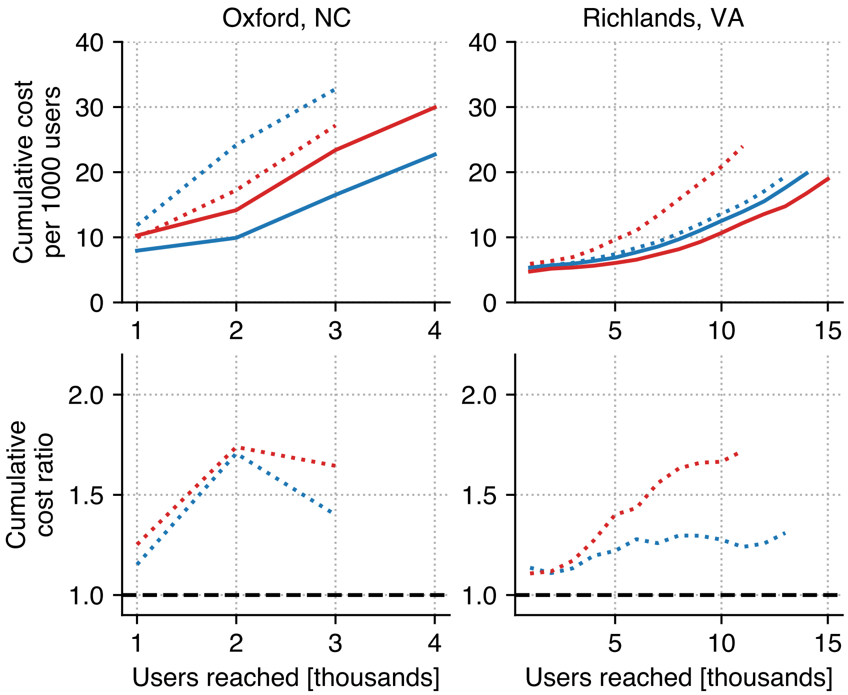


Figure 5.7: The price penalty for showing to non-aligned audience is not just an effect of a particular location or targeting more “extreme” (Narrow) audiences. Findings from Figure 5.6 hold also in other cities and with weaker targeting (here: only political alignment, without additional interests). We note that in some cities Trump’s ads are cheaper for his aligned audience than Sanders’ ads are for his.

optimization of political advertising are not mere hypotheticals and warrant further scrutiny (Section 5.3).

Role of advertiser’s identity We have repeated a subset of our experiments using another advertising account registered as an advertiser in the area of “Social Issues, Elections, and Politics” and linked to a Facebook page unrelated to the first. Our results were quantitatively and qualitatively similar. This suggests that the effects we observed were not tied to our particular advertising account. Nevertheless, we do not make any statements about the extent to which the observed effects hold when run by real political campaigns with a more established history than ours.

Role of budget We also re-ran a subset of our experiments with varying lifetime budgets ranging from $10–$100 per campaign, and with a generous bid cap of $10 in each auction.

Our ads ran on consecutive weekdays at similar times; we observed qualitatively similar skews regardless of the budget. Although \$100 per ad may seem small compared with the total political ad spending, such ads are representative of practice: recent work [11] that analyzes data from Facebook's political ad archive has found that 82% of all political ads spend less than \$100.

Role of competition We ran each pair of campaigns targeting a particular audience representing two different political campaigns at the same time and with the same budget. Such a set-up is designed to ensure that both campaigns have the same users available for delivery (i.e., if run at different times, the skews could be attributed to different Facebook use patterns by liberals or conservatives) and both are experiencing the same competition from other advertisers (i.e., that it would not be the case that one campaign is under-performing because it happened to run at the same time that another large and wealthy advertiser was targeting those users, whereas another campaign avoided such a collision). Thus, running campaigns simultaneously is an effective strategy to isolate the effects of delivery optimization from other extraneous factors. However, to verify that the skews are not merely the effect of our ads competing with each other, we also re-ran a subset of campaigns separately. The qualitative and quantitative skew effects for those campaigns we similar.

Audience sizes We aimed to match our constructed liberal and conservative audiences in size as closely as possible, but the matches are inevitably imprecise as Facebook only provides *estimates* of daily reach [35] rather than exact audience sizes.

User engagement with our ads There are a number of ways users can engage with the ads we present, each of which potentially influences future delivery and pricing: reactions ('like', 'love', 'haha', 'wow', 'sad', 'angry'), commenting, and sharing. Facebook advertising interface reports all such engagements. Additionally, Facebook might be collecting and using *telemetric information*; for example, how long each user spent looking at the ad. This telemetric information is not available to the advertisers (and thus, neither to us), but might still play a role in ad delivery optimization algorithms.

Some of our ads received reactions, comments, and re-shares from the users they were delivered to. We note four important, related observations, that emphasize that our findings about skew in delivery and differential pricing are not merely a function of the ad delivery algorithm's use of user engagement. *First*, we observe consistent skew and price differences in ads that look identical to users, yet trick Facebook into classifying them as partisan (Figure 5.5). Users do not react differently to ads that appear identical, and, therefore, the entire observed difference can be attributed to Facebook's pre-delivery classification (and some random effects). *Second*, we observe consistent skew in delivery of ads that had virtually no engagement since they were run on small budgets and only for a few hours, as shown in Figure 5.4. *Third*, longitudinal ads with neutral content proved less engaging than either aligned or not-aligned ads, yet they eventually reached larger audiences and at lower prices (Figure 5.6). Specifically, non-aligned ads were shared at higher rates than neutral ads (0.34% vs 0.03% for conservative audience, 0.19% vs 0.05% for the liberal audience). This leads us to believe that the relatively lower costs of aligned ads compared to non-aligned ads do not stem from "free" exposures originating from re-shares. *Finally*, we do find a negative correlation between the fraction of positive reactions ("like" and "love") among all reactions and the price in the longitudinal ads with $\rho = -0.91$, $p_{val} = 0.01$. Taken together, our work demonstrates although the skew in delivery as well as differential pricing can be further amplified during the course of delivery by users' reactions, the primary reason stems from Facebook's ad delivery optimization's use of classification of an ad and its landing page content.

Table 5.3 Overview of the audiences created using Facebook's inferred interests.

	Liberal Audience		Conservative Audience	
Location	**Targeting**	**Size**	**Targeting**	**Size**
Celina, OH	**Engagement**: liberal **Interests**: Bernie Sanders **Radius**: +25mi	1,500	**Engagement**: conservative **Interests**: Donald Trump for President **Radius**: +21mi	1,400
Dutchess County, NY	**Engagement**: liberal	15,000	**Engagement**: conservative	15,000
Loraine, OH	**Engagement**: liberal **Radius**: +13 mi	20,000	**Engagement**: conservative **Radius**: +10 mi	22,000
Macclenny, FL	**Engagement**: liberal **Interests**: Bernie Sanders **Radius**: +24 mi	8,500	**Engagement**: conservative **Interests**: Donald Trump for President **Radius**: +30 mi	8,300
McCormick, SC	**Engagement**: liberal **Radius**: +20 mi	3,000	**Engagement**: conservative **Radius**: +17 mi	3,400
Richlands, VA	**Engagement**: liberal **Radius**: +34 mi	5,000	**Engagement**: conservative **Radius**: +10mi	5,000
Saginaw, MI	**Engagement**: liberal **Radius**: +10mi	13,000	**Engagement**: conservative **Radius**: +10mi	13,000
Slinger, WI	**Engagement**: liberal **Interests**: Bernie Sanders, U.S. Senator Bernie Sanders **Radius**: +21mi	2,900	**Engagement**: conservative **Interests**: Donald Trump for President **Radius**: +24mi	3,100
Oxford, NC	**Engagement**: liberal **Radius**: +12mi	3,000	**Engagement**: conservative **Radius**: +10mi	3,600
Scranton, PA	**Engagement**: liberal **Interests**: Bernie Sanders, Democratic Party (United States), U.S. Senator Bernie Sanders, Barack Obama, Joe Biden **Radius**: +45 mi	3,000	**Engagement**: conservative **Interests**: Donald Trump for President, Make America Great Again **Radius**: +50mi	3,200
Michigan	**Engagement**: liberal **Interests**: Democratic Party (United States), Bernie Sanders, U.S. Senator Bernie Sanders, Joe Biden, Barack Obama	34,000	—	—
Wisconsin and Michigan	—	—	**Engagement**: conservative **Interests**: Donald Trump for President, Republican Party (United States), Make America Great Again or Mike Pence	38,000

5.3 Discussion

Our findings suggest that Facebook is wielding significant power over political discourse through its ad delivery algorithms without public accountability or scrutiny.

Implications *First,* Facebook limits political advertisers' ability to reach audiences that do not share those advertisers' political views in ways that are significantly different from traditional broadcast media. The existence and extent of this skew may not be apparent to advertisers and varies based on their ad's message and the destination link used by the campaign. For example, a campaign targeting a certain geographic region might reasonably expect to reach an audience whose political views are representative of users in the region. To discover otherwise would require careful research, as we have demonstrated in this study. Furthermore, the strength of delivery skews vary for campaigns of different political leanings and targeting different populations, making digital advertising inequitable for political campaigns with identical budgets.

Second, recent moves to restrict political advertisers' targeting options [104, 108, 137], although valuable from a user privacy perspective [90, 139, 222], might be undermined by the operation of ad delivery algorithms, and even give companies like Facebook *more* control over selecting which users see which political messages. This selection occurs without the users' or political advertisers' knowledge or control. Moreover, these selection choices are likely to be aligned with Facebook's business interests, but not necessarily with important societal goals.

Third, today, researchers, regulators, and campaigns lack access to algorithms and data required for a more thorough study of ad delivery skews and their likely impacts. In particular, although much has already been said about the inadequacy of current ad transparency tools provided by ad platforms [88, 176, 237], our work draws attention to the need to expand these efforts to account for ad delivery algorithms as well.

Policy analysis Today, U.S. law cannot do much, if anything, to *directly* change how ad platforms deliver political ads. For the foreseeable future, it is likely that the primary regulator of digital political advertising will not be the government, but rather ad platforms themselves.

The U.S. Congress has addressed conceptually similar "ad delivery issues" in the past, albeit in a different domain. For example, the Federal Communications Commission (FCC) enforces the so-called Equal-Time Rule [4], which originated in 1927 in response to worries that broadcast licensees could unduly influence the outcome of elections. The rule requires that licensees make air time available to all candidates for the same office on equivalent terms. However, the rule only applies to broadcast licensees, and has only narrowly survived constitutional scrutiny in part because it implicates government interests in managing limited broadcast spectrum [33].

Prevailing interpretations of the First Amendment are likely to block efforts to extend the logic of the Equal-Time Rule to digital advertising platforms, which are not regulated like broadcast licensees. As an initial matter, the First Amendment strongly protects political speech, and generally tolerates only narrowly-tailored government regulations [254]. This protection is so strong that legal scholars cannot even be confident that lighter-touch kinds of regulations—for example, a requirement that social media users be entitled to opt-in to micro-targeted political advertising—would survive constitutional scrutiny. Moreover, the Supreme Court recently declared that "the creation and dissemination of information" constitutes speech under the First Amendment [220]. This reasoning, which might expand the "commercial free speech" rights of companies, creates some uncertainty about the government's ability to restrict corporations' use of data in the context of digital advertising.

Looking ahead, it is clear that government regulation of digital political advertising is on firmest legal footing when it requires disclosure about who is speaking to whom, when, and about what [254]. Accordingly, Congress and the FEC can consider transparency requirements that will enable detailed auditing and research about ad targeting and the delivery of political ads.

Mitigations As an initial data, the public and the campaign managers need more information about the operation of ad delivery algorithms and their real-world effects. Ad platforms could increase transparency around political ads (including key metrics such as targeting criteria, detailed ad metadata, ad budgets, and campaign objectives) to enable further study of the effects of ad targeting and delivery. And they could provide access to and insight into the ad delivery algorithms themselves (including those involved in running the auction, relevance measurement and estimation, and bid and budget allocation on advertisers' behalf), allowing third parties greater ability to study and audit their performance and effect on political discourse. Without these and similar steps, policymakers and the public will be unable to formulate appropriate responses.

Ad platforms could also disable delivery optimization for political content, or a least allow advertisers to do so. They could also introduce more nuanced user-facing controls for political content delivery and expand public ad archives to make them more accessible and usable by everyone.

Finally, we call on ad platforms to acknowledge the central role they play in the delivery of political ads, and to collaborate with other key stakeholders—including researchers, political campaigns, journalists, law, policy and political philosophy scholars—to address that role when it is not aligned with public interests.

Chapter 6

User Experiences with Problematic Advertising

We saw in Chapter 4 that in an effort to find users who would find a job relevant, personalization can lead to discriminatory outcomes along race and gender. Chapter 5 shows how that the same mechanisms can limit the delivery of political advertising to non-aligned audiences. Given these effects, could personalization also find vulnerable users when an advertiser has a deceptive offer to advertise? Due to the scale of marketplaces like Facebook, an average user runs into ads on a vast variety of topics—ranging from neutral product ads, to opportunity ads, and even to problematic clickbait ads and scams. To investigate the kinds of problematic ads that exist on Facebook, and personalization's role in exposing users to such content, we turn our attention to studying *individual* user experiences in this chapter.

To do so, we recruit a panel of 132 paid participants, who we select across a variety of demographic categories. We longitudinally observe participants' Facebook ad experiences over a period of three months, collecting the ads they receive, and the revealed targeting information for each ad. We use a combination of (1) logged data and (2) quantitative surveys to measure our participants' ad experiences. *First*, we instrument our participants' web browsers to collect all Facebook ads they are shown in their desktop browsers, alongside the detailed targeting information Facebook provides for these ads. *Second*, using a combination of inductive qualitative coding, and deductive analysis of computational and social science research, as well as existing platform policies, we develop a *codebook* of ad categories, covering a variety of potentially problematic ad types. *Third*, using human raters, we classify over 32,000 ads shown to our participants using this codebook. With this coded data, we regularly survey our participants to assess which types of ads—within the set of ads that they are shown by Facebook and which we annotated—they find problematic and why. The following discussion is from our published work at the *USENIX Security Symposium*, 2023 [10].

Given the wide variance in ads a user may potentially receive, it is important to consider whether some users' *overall ad experience* might be worse than others. Prior work has illustrated the impact of harmful media [17, 29, 197, 218, 240, 246, 247], has theorized about the ways in which digital ads may harm users [100, 153, 183, 194, 200, 212, 258] has asked users themselves to express why they find certain ads problematic [260]. However, a complete understanding on the online ad experiences of individual users, along with a breakdown of the kinds of ads different users find problematic, remains elusive.

In this chapter, we build on prior work to systematically identify which categories of ads people perceive as problematic, evaluate if there are skews in the delivery of problematic categories of ads, and determine the roles of advertisers and personalization algorithms in the distribution of problematic ads. Thus, we aim to answer the following research questions:

RQ1: What categories of ads are perceived as problematic?

RQ2: Are there skews in the distribution of problematic ads?

RQ3: Who is responsible for any observed skews?

Using novel collected data from active Facebook users, we first examine the content that participants dislike (RQ1). We identify four categories of ads that participants find problematic (i.e., are disliked more than ads of any other category): deceptive ads, content that is prohibited by Facebook itself, clickbait, and ads considered sensitive by platform or government policy (e.g., ads for weight loss, gambling or alcohol).

We then statistically model the distribution of problematic ads across our panel (RQ2). Our results show that problematic content makes up a relatively small fraction of all ads our users see on Facebook—a median of 10%—but a subset of our panel is exposed to problematic ads over three times more often than the median participant. Looking at which participants tend to receive more problematic ads, we find that participants who are older are more likely to see deceptive and clickbait ads, and those who are Black are also more likely to see clickbait ads. Men are additionally more likely to see financial ads, a complex category that is (i) considered sensitive by U.S. regulation and Facebook policy, as it may include offers for exploitative financial products, (ii) disliked by participants more than neutral ads, but (iii) which also may include beneficial financial products.

Finally, we investigate the extent to which the advertisers and the platform personalization algorithms are responsible for these biases (RQ3). We find that certain categories of ads (e.g., opportunity ads and ads for sensitive topics) tend to be much more narrowly targeted than neutral ads, suggesting that advertisers carefully choose which users are eligible to see these ads. On the other hand, we identify a subset of ads that are not targeted at all (i.e., the advertisers make all adult U.S. users eligible to see the ad), and find that demographic skews still persist for ads across different problematic categories. For example, we find that older users receive a much higher fraction of problematic ads overall, even when the advertiser did not use any targeting options. Similarly, we find that Hispanic users are shown a higher fraction of deceptive ads, and women see fewer financial ads. Together, our results shed light on users' overall ad experiences on a major platform, and illuminate disparities in those experiences caused by a combination of advertiser choices and platform algorithms.

In the remainder of this chapter, Section 6.1 describes the methods we use for collecting and analyzing user data. Section 6.2 describes our process for annotating collected ads in detail. Section 6.3 presents results and observed skews from user experiences; Section 6.4 provides a concluding discussion about the impact of the observes skews, and also provides recommendations to mitigate the effects of personalization for problematic content.

6.1 Methods

Below, we describe our methods for recruiting a diverse and demographically balanced panel and for collecting the desktop ads our participants are shown by Facebook.

6.1.1 Panel Recruitment

We recruited our panel of Facebook users from two sources: by listing tasks on Prolific, an online crowd-work and survey platform, and by advertising on Facebook.

Participants were screened via a short survey.[1] Our criteria to be eligible for the study were that participants must (1) have an active Facebook account that (2) they use for at least 10 minutes per day (3) on a desktop or laptop computer (4) via either the Google Chrome or Mozilla Firefox browsers (5) without using ad blockers or tools for anonymous browsing (e.g., Tor). Additionally, we went to significant lengths to recruit a diverse panel across select demographic variables: race and ethnicity (white; Black; Hispanic; Asian), gender (men; women), age (younger than Generation X [58]; Generation X or older) and educational attainment (below a bachelor's degree; bachelor's degree and above). We sought to balance our panel among all combinations of our chosen demographic variables (e.g., representation for Generation X Hispanic women with high educational attainment) but we struggled with recruitment and retention of some demographics, partly due to the distribution of users who participate in online studies or use the platforms we recruited on [199, 219, 226]. We made a continuous effort to balance our sample by accepting participants on a rolling basis and not screening in those with demographics we were saturated with. Table 6.1 shows the ultimate demographic breakdown of our participants.

Unfortunately, while all participants were screened based on their Facebook usage, not all users contributed a significant number of ads during the 3 month study period. Of the 184 participants originally enrolled in the study, 132 were *active* participants, which we define as those who contributed at least 30 ads (on average 10 per month) over the course of the three months of their participation in the study.

[1] Prolific participants were compensated with a base pay of $8.04 per hour for completing the screening survey while those recruited via Facebook advertisements were not compensated for the screening survey as there is no mechanism to do so. Demographics which we initially struggled to recruit were offered marginally more compensation. The survey took a median of 6 minutes and 9 seconds to complete.

Table 6.1 Demographics of panel participants. Active participants are those who contributed at least 30 ads over the course of the study. Population percentages from the U.S. 2020 Census shown for comparison of representativeness.

Variable	Value	Recruited		Active		Census
		n	%	n	%	%
Gender	Female	96	52.17	71	53.79	50.5
	Male	86	46.74	59	44.70	49.5
	Non-binary	2	1.09	2	1.52	–
Age	Younger than Gen-X	134	72.83	88	66.67	33.6
	Gen-X and older	50	27.17	44	33.33	47.8
Race / Ethnicity	White	105	57.07	82	62.12	75.8
	Latino/Hispanic	21	11.41	16	12.12	18.9
	Black	53	28.80	32	24.24	13.6
	Asian	21	11.41	16	12.12	6.1
	Other	3	1.63	3	2.27	–
Education	Below Bachelor's	72	39.13	51	38.64	58.5
	Bachelor's or above	112	60.87	81	61.36	32.9
Total		**184**		**132**		

6.1.2 Data Collection

Logged Data. Our study collected the ads that were shown to our participants on their Facebook news feeds while using Facebook on a desktop computer over a 3 month period. In order to collect our participants' ads, we used a browser extension, based on the NYU Ad Observer project [84, 189]. We modified Ad Observer to include unique participant IDs along with the ads reported to our server, and we introduced an additional "Surveys" tab that serves participants monthly surveys to collect their sentiments for their individual ads. Across all of our recruited participants, we collected 165,556 impressions to 88,509 unique ads. Repeat impressions of ads are relatively sparse in our data—a median of twice per ad per participant—and only 5.33% of our ads are shown more than 3 times to a participant.

Targeting Data. We also collected ad targeting information provided by Facebook through its "Why am I seeing this?" API [83], which reveals information about how the advertiser selected their target audience [14]. While prior work has shown that Facebook's targeting explanations can be incomplete, and include only one targeting criteria in each ad explanation [14], we find empirically that the system has changed since. We also observe differences between the summarized targeting data which is shown on the user interface, and what is reported through the API. Our data includes several instances of multiple targeting criteria—62.7% of ads in our data with interest targeting include more than one interest.

Survey Data. Every month, we prepared a survey that assesses participant sentiments toward the ads they saw on Facebook during the prior month. Specifically, for each ad that we showed to a user in the survey, we asked them: "Which of the following, if any, describe your reasons for *disliking* this ad?" and present the following non-mutually exclusive answer choices:

- It is irrelevant to me, or does not contain interesting information.
- I do not like the design of the ad.
- It contains clickbait, sensationalized, or shocking content.
- I do not trust this ad, it seems like a scam.
- I dislike the advertiser.
- I dislike the type of product being advertised.
- I find the content uncomfortable, offensive, or repulsive.
- I dislike the political nature of the ad.
- I find the ad pushy or it causes me to feel anxious.
- I cannot tell what is being advertised.
- I do not dislike this ad.

We then ask: "Which of the following, if any, describe your reasons for *liking* this ad?" and present the following non-mutually exclusive answer choices:

- The content is engaging, clever or amusing.
- It is well designed or eye-catching.
- I am interested in what is being advertised.
- It is clear what product the ad is selling.
- I trust the ad, it looks authentic or trustworthy.

- I trust the advertiser.
- It is useful, interesting, or informative.
- It clearly looks like an ad and can be filtered out.
- I do not like this ad.

Answer choices for these questions are drawn from Zeng et al.'s taxonomy of reasons for users' like or dislike of ads [260], with the exception of one item. In a small pilot version of this survey, in which we allowed participants to also provide free-text answers of their reasons for liking and disliking Facebook ads with 300 respondents, we identified an additional reason for liking an ad, "This ad is filterable", so we included it to capture a broader spectrum of reasons users like ads.

We survey participants about at most 5 ads from each of our seven ad categories (Section 6.2). We limit the monthly surveys to up to 35 ads each so that it did not become prohibitively long (more than 20 minutes) for participants to complete.

Study Deployment. We began data collection in November 2021, with participants recruited on a rolling basis. Each participant was a part of our study for three months. The final participant completed the study in September 2022. We compensated our participants by paying them up to $60: $5 when they signed up, $15 for each month they kept the plugin installed and completed the monthly sentiment survey, and upon completing all three months of the study, they were rewarded with a $10 bonus payment. Those participants who dropped out of our study were compensated using the scheme above based on how long they did participate. Since we deployed surveys directly through our extension, we were not able to assess average time of completion, but pilot tests of the survey averaged a completion time of about 15 minutes.

6.1.3 Analysis

Here, we describe the quantitative methods we employ to analyze survey responses, logged ad observations, and ad targeting data. We limit all our analyses to the 32,587 ads that we annotated (see Section 6.2), and to our list of active participants (Table 6.1).

RQ1. For survey responses, we use Chi-squared (χ^2) tests for equality of proportions to compare rates of ad dislike. We also report Cohen's ω as the effect size of the Chi-squared tests to characterize the scale of differences. As a general guideline, $\omega = 0.1$ is considered a small effect, 0.3 is a medium effect, and 0.5 and above is considered a large effect [49]. We examine the association between the reasons for dislike mentioned in the surveys and the ad type through mixed-effects logistic regression models. To control for variance in participants' individual preferences, we include a random effect term for each participant. In line with statistical best practice [102], we do not correct our regression models as each model represents a purely complementary comparison (e.g., contains a distinct dependent variable).
RQ2. To understand disparities in the distribution of ad types, we treat number of ad types observed for each participant as a frequency distribution. To quantify inequality in this distribution, we compute skewness [252], a measure of asymmetry for a probability distribution, computed via its third standardized moment. A positive skew implies a distribution with a long right tail, while a negative skew means the left tail is longer. We also compute the Gini coefficient [251] to measure inequalities across participants. To understand inequities between demographic groups, we use linear regression models to model the fraction of each ad category in participants' ad diet, as a function of their demographics.
RQ3. To disentangle ad delivery's influence from ad targeting in our observations, we use the advertising interface to obtain audience size estimates for each ad. Concretely, we query Facebook's advertising API for monthly "reach" estimates for the targeting specifications of every ad in our dataset. Note that these estimates are not accessible for ads that use Custom Audiences (CAs), such as phone number uploads or cookie re-targeting; those are only known to the owners of these CAs. We use linear regressions similar to RQ2 to identify differences between demographic groups that appear due to the platform's ad delivery practices.

6.1.4 Ethics

Given the sensitivity of the data we were collecting, we took care to follow best practices, maximizing beneficence while minimizing harm to our participating users and Facebook itself. First, our research project was approved by our institution's Institutional Review Board (IRB). Second, we collected the minimal data on our participating users necessary to conduct the study; we only collected personally-identifiable information where necessary to facilitate payments, and we used unique, random identifiers for all survey responses and ads collected. Third, we controlled access to the uploaded pseudonymous data to just the research team, and we do not plan on making this data generally available to protect the privacy of our users. Finally, we minimized the harm to advertisers and Facebook itself by not causing any ad impressions that would not have otherwise occurred; the only additional requests to Facebook were to fetch the targeting specifications, and to later retrieve audience sizes of these specifications.

While Facebook prohibits collection of data using automated means in its terms of service (ToS), we argue that the public benefits of our work outweigh the risks posed to Facebook. Further, violating ToS by scraping content that is otherwise available through non-automated means is not considered a violation of the U.S. Computer Fraud and Abuse Act [239]. Platforms, however, reserve the right to ban users who scrape or have done so in the past.

6.2 Categorizing Ads

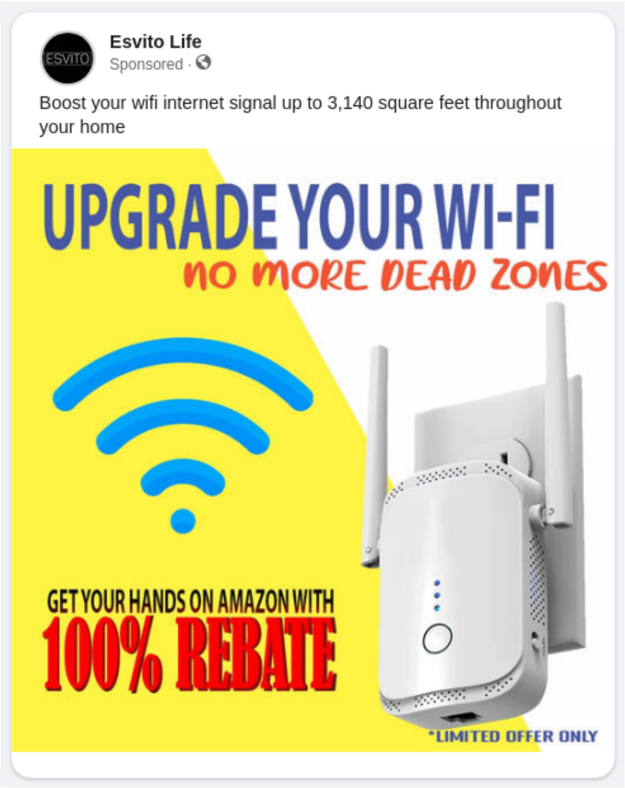
In order to evaluate whether there are inequities in participants' exposure to problematic ads, we first evaluate which of our collected ads are problematic. To do so, we develop a codebook to categorize the ads our participants see, and then use that codebook to annotate a significant subset of their ads.

Creating the codebook. We use a combination of inductive qualitative coding [47, 228], and deductive analysis [18] of prior work and platform policies to develop a robust categorization of participant ads. To create our initial inductive categorization of Facebook ads, we conducted pilot data collection with 7 participants, collecting their ads with our browser extension between June and July 2021. We then cross-referenced our initial codebook with platform and governmental policies and empirical research to develop our final ad categories. Our categorization particularly focuses on capturing problematic ads, though we also make sure our codebook captures content that users might find unproblematic, such as products, events, or local businesses. Below, we define our categories, describe how we reason about them, and provide examples from our dataset.

6.2.1 Deceptive

Ads that may overtly or deceptively lead users to engage with fraudulent offers, potential scams, false or misleading claims, or predatory business practices (e.g. recurring billing). This includes fraudulent offers, potential scams, false or misleading claims, predatory business practices. *Examples:* Guaranteed monthly income, sign-up flows for personal information ("clickfunnels"), non-descript offers with requests for direct messages.

Deceptive advertising and its breadth is notoriously hard to capture (see, e.g., a review of definitions [95] and a diversity of FTC reports on the subject [99]). Therefore we define this code broadly, to be able to capture multiple forms of deceptive and scam content. We categorize financial and personal information scams, fraudulent offers, and a diverse array of misleading content as Deceptive. Many aspects that we cover in this definition are covered by Facebook's policies for unacceptable business practices [71], unrealistic outcomes [72], and broadly under the platform's deceptive content policy [74]. Prior work has documented deceptive ads in contexts such as malicious web advertising [153], social engineering attacks [183], and distributing malware [200, 258].



Esvito Life
Sponsored ·

Boost your wifi internet signal up to 3,140 square feet throughout your home

Figure 6.1: Example Deceptive ad; multiple reviews on Facebook page mention a rebate was never issued.

Qualifies:

- Ads that have potential to harm users financially, such as:
  - Payday loans, paycheck advances, bail bonds, or any short-term loans.[2]
  - Debt settlement services, especially with strongly worded guarantees.
- Services which are highly unlikely to result in the advertised outcome, e.g. fat burning pills, guaranteed monthly income etc.
- Scams, either for money or personal information, that we can confirm via Facebook reviews or Better Business Bureau reports.

[2] Also considered deceptive in Facebook's policies [70].

- Ads that employ deceptive tactics, such as:
  - Containing false or exaggerated claims.
  - Designed to look like they are advertising a different product than the linked webpage.
- Overly pushy or manipulative ads.
- Predatory business practices such as: requests for direct messages, recurring/non-cancellable billing as mentioned by users.

Does not qualify:

- Sketchy product ads for which we cannot find evidence of deception.
- Visibly low quality products that are not inherently harmful e.g. clothing, jewellery, novels.

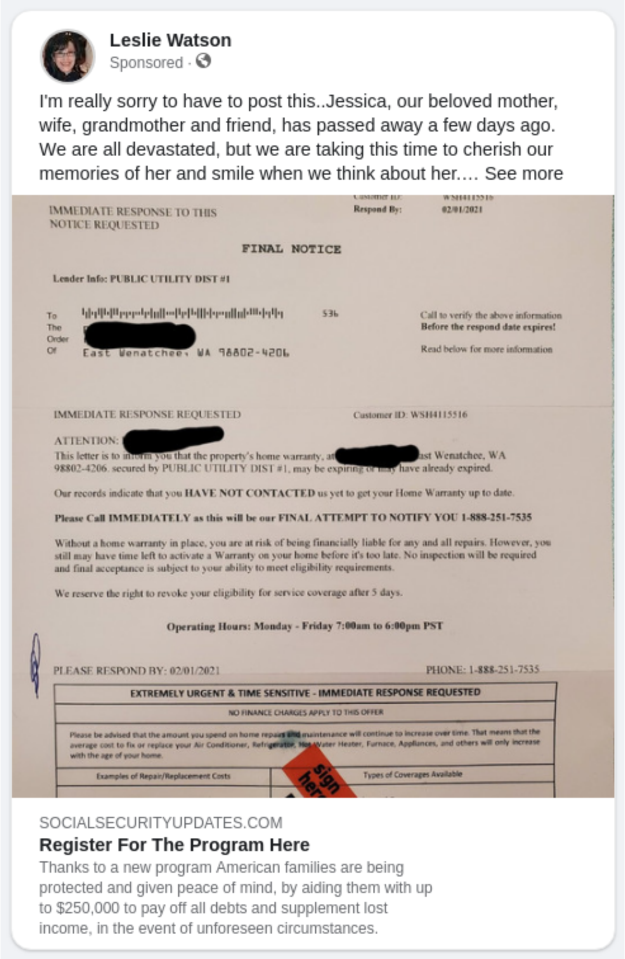
6.2.2 Clickbait

Ads that intentionally omit crucial information or exaggerates the details of a story to make it seem like a bigger deal than it really is [43]. Such ads often have three distinct characteristics [260]: the ad is attention grabbing, the ad does not tell the viewer exactly what is being promoted to "bait" the viewer into clicking it, and the landing page of the ad often does not live up to people's expectations based on the ad. *Examples:* Provocative news headlines, celebrity gossip, incomplete offers ("Click to find out").

Prior work has documented how clickbait ads are attention grabbing by being unclear, and do not live up to users' expectations [195, 260]. It has also been found to waste users' time [212], contain provocative content [194], and act as a vehicle for misinformation [129, 194, 261]. Facebook's policies also recognize the misleading and annoying nature of clickbait, and they enforce policies to reduce exposure to such content [43, 76, 170].

Qualifies:

- Ads where the text, headline, and description together don't clarify what precisely is being advertised.
- Ads that omit information to entice users.



Leslie Watson
Sponsored ·

I'm really sorry to have to post this..Jessica, our beloved mother, wife, grandmother and friend, has passed away a few days ago. We are all devastated, but we are taking this time to cherish our memories of her and smile when we think about her.... See more

IMMEDIATE RESPONSE TO THIS NOTICE REQUESTED

Customer ID: WSH4115516
Respond By: 02/01/2021

FINAL NOTICE

Lender Info: PUBLIC UTILITY DIST #1

To The Order Of
[redacted]
East Wenatchee, WA 98802-4206

536

Call to verify the above information
Before the respond date expires!
Read below for more information

IMMEDIATE RESPONSE REQUESTED

Customer ID: WSH4115516

ATTENTION: [redacted]

This letter is to [redacted] you that the property's home warranty, at [redacted]ast Wenatchee, WA 98802-4206. secured by PUBLIC UTILITY DIST #1, may be expiring [redacted] have already expired.

Our records indicate that you **HAVE NOT CONTACTED** us yet to get your Home Warranty up to date.

Please Call IMMEDIATELY as this will be our FINAL ATTEMPT TO NOTIFY YOU 1-888-251-7535

Without a home warranty in place, you are at risk of being financially liable for any and all repairs. However, you still may have time left to activate a Warranty on your home before it's too late. No inspection will be required and final acceptance is subject to your ability to meet eligibility requirements.

We reserve the right to revoke your eligibility for service coverage after 5 days.

Operating Hours: Monday - Friday 7:00am to 6:00pm PST

PLEASE RESPOND BY: 02/01/2021

PHONE: 1-888-251-7535

EXTREMELY URGENT & TIME SENSITIVE - IMMEDIATE RESPONSE REQUESTED

NO FINANCE CHARGES APPLY TO THIS OFFER

Please be advised that the amount you spend on home repairs and maintenance will continue to increase over time. That means that the average cost to fix or replace your Air Conditioner, Refrigerator, Hot Water Heater, Furnace, Appliances, and others will only increase with the age of your home.

Examples of Repair/Replacement Costs	Types of Coverages Available

SOCIALSECURITYUPDATES.COM

Register For The Program Here

Thanks to a new program American families are being protected and given peace of mind, by aiding them with up to $250,000 to pay off all debts and supplement lost income, in the event of unforeseen circumstances.

Figure 6.2: Example Clickbait ad.

- Have very dense text in the image.
- Invite the users to tap a section of the image/button in image for results.
- Ads for products that are actually affiliate marketing links or data collection forms, e.g. home renovation and solar panel surveys.

Does not qualify:

- Ads that use loud language but are clear about what the advertised product, e.g. e-commerce ads that advertise 200% growth in business but are upfront that they're advertising an online course.

6.2.3 Potentially Prohibited

Ads that may not be allowed on the platform according to Facebook's prohibited content policies [168] on unacceptable content, dangerous content and objectionable content. *Examples:* Tobacco, drugs, unsafe dietary supplements, multi-level marketing, weapons.

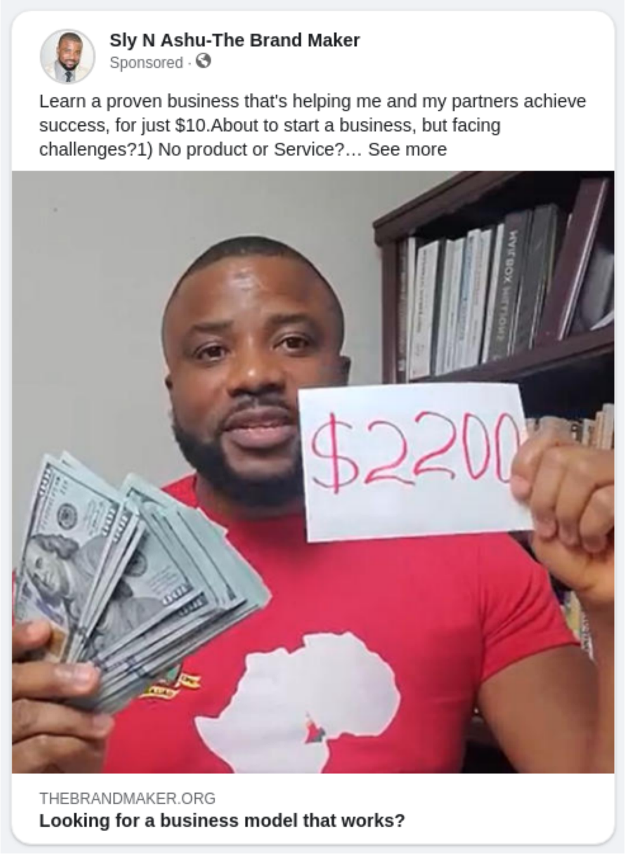
Facebook's policies prohibit several types of ads [74], including but not limited to ads for tobacco, adult content, body parts, payday loans, and multi-level marketing. Ads that pose a security threat to users, such as spyware or malware, non-functional landing pages, and efforts circumventing review systems, are also prohibited [169]. Even with an extensive policy, Facebook's ability to accurately detect content and enforce policies is limited (see, e.g., prior work documenting challenges in detection and enforcement of political advertising policies [66, 149]). We therefore code for ads whose content match any of Facebook's prohibitive policies. We note that only Facebook can enforce these policies – therefore we refer to our annotations as *potentially* prohibited.

Qualifies:

- Prohibited substances—such as illegal prescription and recreational drugs, tobacco, and related products.
- Unsafe dietary supplements and medical treatments.
- Weapons, ammunition or explosives.
- Adult products or services.
- Instant loans, pre-financing and security deposits.
- Selling human body parts or fluids.
- Multilevel marketing, or income opportunities that offer quick income with low investment.
- Spyware or malware.
- Ads against vaccinations.

Does not qualify:

- Products for sexual or reproductive health, such as medical devices for family planning and contraception.



Sly N Ashu-The Brand Maker
Sponsored ·

Learn a proven business that's helping me and my partners achieve success, for just $10.About to start a business, but facing challenges?1) No product or Service?... See more

THEBRANDMAKER.ORG
Looking for a business model that works?

Figure 6.3: Example Potentially Prohibited ad; advertises a business plan that offers quick income with low investment.

6.2.4 Sensitive

Ads that fall under Facebook's content-specific restrictions policy [74]: such content isn't prohibited but, given its sensitive nature, it must comply with additional guidelines, including written permissions and targeting restrictions.

Facebook subjects ads for sensitive topics to additional scrutiny on their content and targeting practices [74]. For example, ads for weight loss programs can only be targeted to people at least 18 years or older; financial advertisers must provide authorization by regulatory authorities, and online pharmacies require an additional certification [69].

In addition to platform policies, sensitive ads closely relate to prior work on content that targets user's vulnerabilities [97, 100] — such content may be benign to some users but may foster negative thoughts or behaviors for others [122, 174]. Gak et al. [100], for instance, found that among people with a history of unhealthy body stigmatization, dieting, or eating disorders, being targeted with weight-loss-related ads had negative emotional and physical outcomes.

Within Sensitive ads, we find an increased prevalence (more than two-thirds) for Financial ads, so we break this code into two sub-codes — Sensitive: Financial and Sensitive: Other.

Sensitive: Financial: Ads that contain products or services related to managing finances, building credit, and other financial tools. Such ads are subject to content-specific restrictions on Facebook [74], and must comply with additional targeting and authorization restrictions. *Examples:* Credit cards, loans, mortgage financing.

Qualifies:

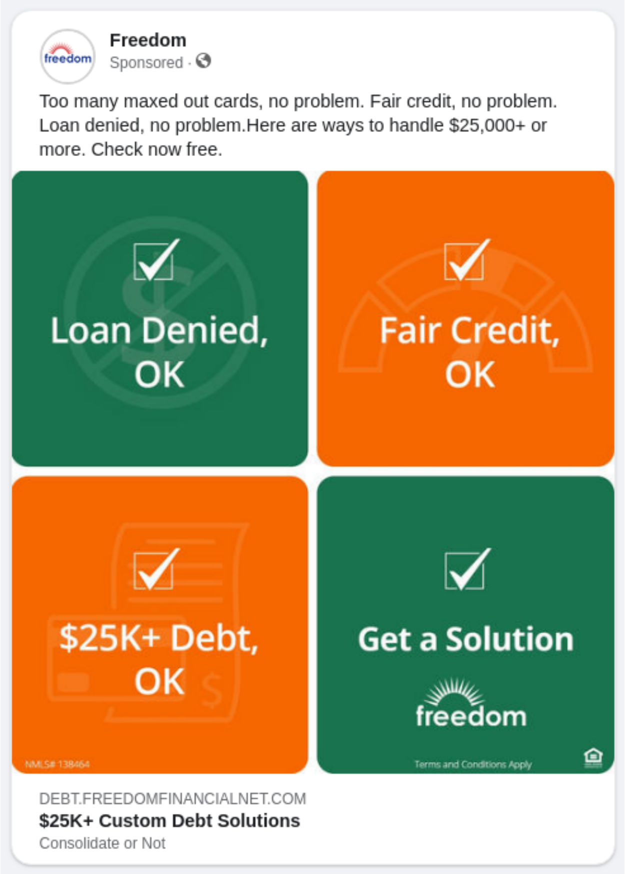
- Loans and mortgage financing ads.
- Checking, savings and brokerage accounts ads.
- Credit card ads.
- Ads regarding financial and investment advice.
- Cryptocurrency or stock investment opportunities.

Does not qualify:

- Ads that reference saving money, but whose products or services are not inherently financial (e.g. browser extensions for deal-hunting).
- Insurance advertising.
- Ads for selling homes.

Sensitive: Other: Ads that contain subject matter that may be sensitive or triggering for users to view, or they may contain content that is harmful for vulnerable groups (e.g. those suffering from an addiction). Such ads are subject to content-specific restrictions on Facebook, and must comply with additional targeting and authorization restrictions. *Examples – Sensitive: Other:* Weight loss programs, online mental health prescription services, online slot machines.

Qualifies:



Freedom
Sponsored ·

Too many maxed out cards, no problem. Fair credit, no problem. Loan denied, no problem.Here are ways to handle $25,000+ or more. Check now free.

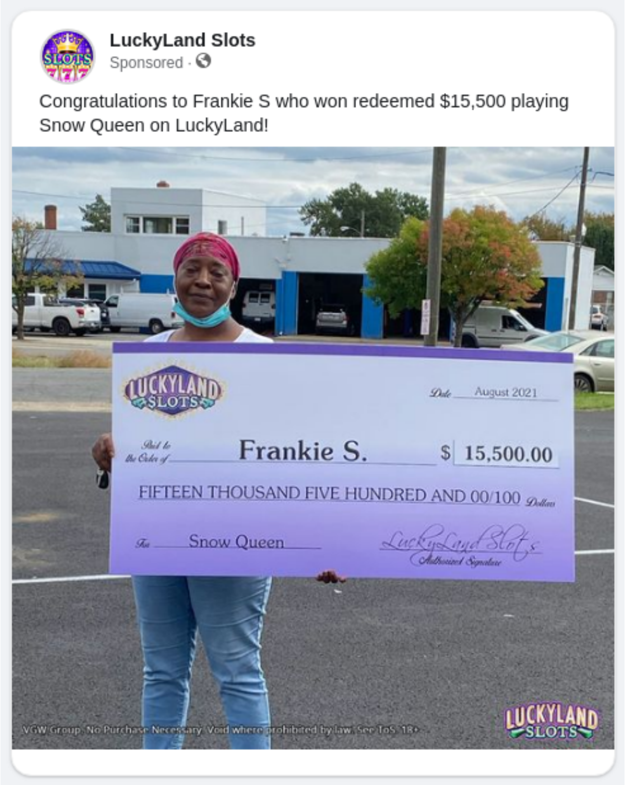
DEBT.FREEDOMFINANCIALNET.COM
$25K+ Custom Debt Solutions
Consolidate or Not

Figure 6.4: Example Sensitive: Financial ad.

- Alcohol.
- Gambling, casinos, online slot machines.
- Dieting, weight loss treatments, anything related to body image.
- Prescription and over-the-counter drugs.
- Online pharmacies and services for mental and physical health.

Does not qualify:

- Ads for meal plans.
- Popular brick and mortar pharmacy ads e.g., CVS and Walgreens.



LuckyLand Slots
Sponsored ·

Congratulations to Frankie S who won redeemed $15,500 playing Snow Queen on LuckyLand!

LUCKYLAND SLOTS
Date August 2021
Paid to the Order of Frankie S.
$ 15,500.00
FIFTEEN THOUSAND FIVE HUNDRED AND 00/100 Dollars
For Snow Queen
LuckyLand Slots
Authorized Signature
VGW Group. No Purchase Necessary. Void where prohibited by law. See ToS. 18+
LUCKYLAND SLOTS

Figure 6.5: Example Sensitive: Other ad for an online slot machine.

6.2.5 Opportunity

Ads that present any employment, housing, or educational opportunity to users. *Examples:* Degree programs, jobs or gig-work, fellowships, scholarships.

We coded for ads that displayed opportunities for users, such as a job or gig, higher education, or apartments and homes for sale. Facebook's own policies prohibit discrimination in targeting of opportunities, or advertising fraudulent or misleading opportunities [75]. Further, cases of discrimination in the delivery of online opportunity ads [8, 53, 136] led us to code these ads to examine their distribution among our participants.

Qualifies:

- Ads for a job or gig, such as
  - Full- or part-time employment opportunities.
  - Gig-work opportunities like Uber, Doordash etc.

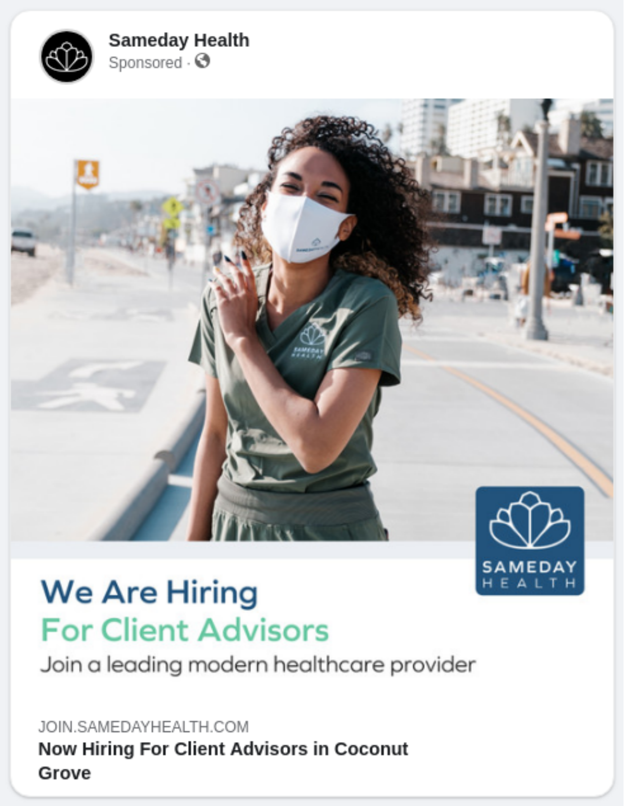


Figure 6.6: Example Opportunity ad; advertises a job.

  - Local job fairs.

- Ads for educational opportunities (for-profit universities and online degrees included).
- Fellowships, scholarships, writing contests, etc.

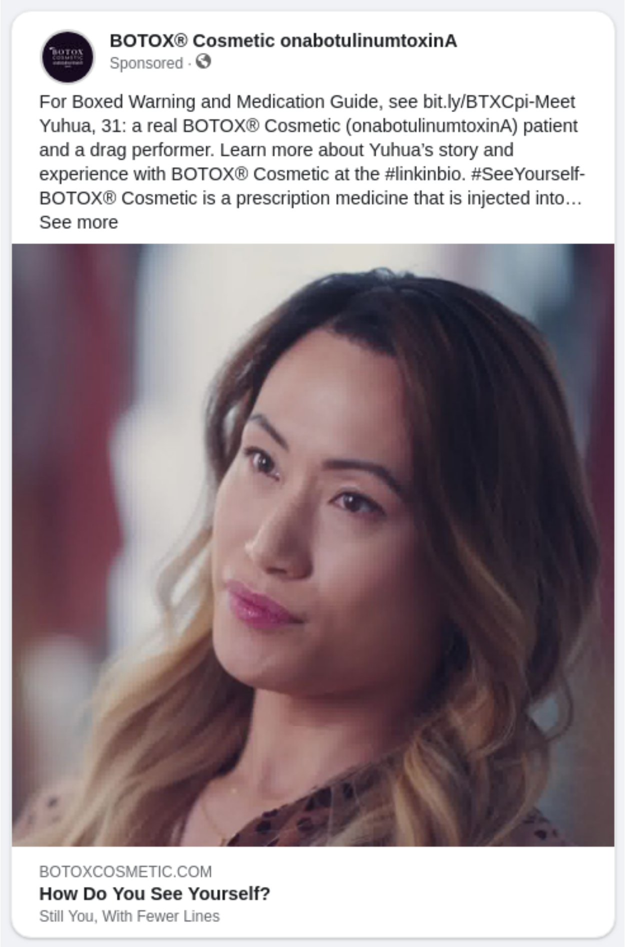
Does not qualify:

- Product sweepstakes or cash-back promotions.
- Ads with financial opportunities (savings, credit cards etc.), regardless of how big the sign-up rewards are.
- Online studies or market research opportunities, regardless of compensation.

6.2.6 Healthcare

Ads that contain products, services or messages related to healthcare, fitness, mental and physical wellness. *Examples:* Medical devices, gym equipment, public health announcements, fitness programs, health insurance.

We find a wide array of healthcare-related ads that are broader than the content covered by Facebook's content-specific restrictions (Sensitive), and we use a separate code to capture such content. These ads are diverse in nature, ranging from helpful to possibly problematic.



BOTOX® Cosmetic onabotulinumtoxinA
Sponsored ·

For Boxed Warning and Medication Guide, see bit.ly/BTXCpi-Meet Yuhua, 31: a real BOTOX® Cosmetic (onabotulinumtoxinA) patient and a drag performer. Learn more about Yuhua's story and experience with BOTOX® Cosmetic at the #linkinbio. #SeeYourself-BOTOX® Cosmetic is a prescription medicine that is injected into… See more

BOTOXCOSMETIC.COM
How Do You See Yourself?
Still You, With Fewer Lines

Figure 6.7: Example Healthcare ad; advertises a cosmetic treatment.

Qualifies:

- Fitness products or services, and gym memberships.

- Vitamins or supplements.
- Physical appearance-related products or services, like hair growth supplements.
- Health insurance.
- Dieting products or services (also annotated as Sensitive).
- Online mental health clinics and prescription services.
- At-home medical monitoring devices.
- Public-health announcements (e.g. CDC, WHO).
- Ads for children's health.

Does not qualify:

- Ads for meal plans.
- Ads for pet health.

6.2.7 Neutral

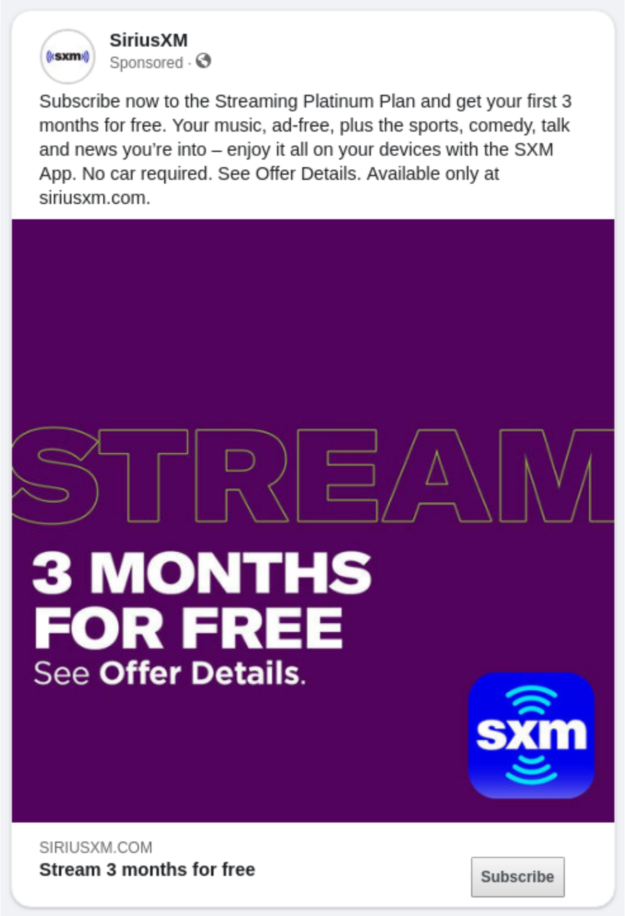
Ads that simply seek to advertise a product, service, local event, apolitical news etc. Ads that don't fall into other categories, and seem benign based on their impact on users, should be marked Neutral. This code is mutually exclusive from all others. *Examples:* Sales, product deals, local events.

Qualifies:

- Product ads.
- Services and subscriptions.
- News and information ads, from news outlets as well.
- Religious ads.
- Insurance ads.

Does not qualify:

- Ads that fit strongly into one of our other specific categories.
- Ads with either opportunities or potential harmful outcomes for users.



SiriusXM
Sponsored ·

Subscribe now to the Streaming Platinum Plan and get your first 3 months for free. Your music, ad-free, plus the sports, comedy, talk and news you're into – enjoy it all on your devices with the SXM App. No car required. See Offer Details. Available only at siriusxm.com.

SIRIUSXM.COM
Stream 3 months for free

Figure 6.8: Example Neutral ad.

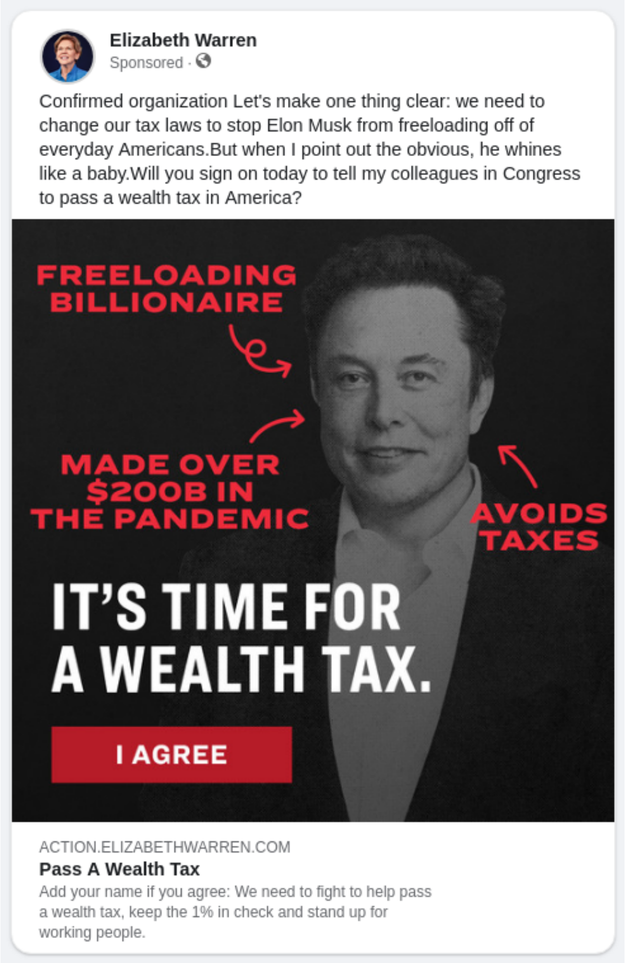
6.2.8 Political

Ads that contain any overt references to political subject matters. *Examples:* Political campaign ads, petitions for political causes.

Qualifies:

- References to any political candidates or figures.
- References to any bills, laws, legislation, etc.
- Petitions, causes, events, or fundraisers that are politically affiliated or motivated.
- References to political parties.

Does not qualify:



Elizabeth Warren
Sponsored ·

Confirmed organization Let's make one thing clear: we need to change our tax laws to stop Elon Musk from freeloading off of everyday Americans.But when I point out the obvious, he whines like a baby.Will you sign on today to tell my colleagues in Congress to pass a wealth tax in America?

ACTION.ELIZABETHWARREN.COM
Pass A Wealth Tax
Add your name if you agree: We need to fight to help pass a wealth tax, keep the 1% in check and stand up for working people.

Figure 6.9: Example Political ad.

- Ads for political merchandise, e.g. t-shirts.
- Religious ads.

While we initially coded for political ads, we exclude them from our analysis. We consider ads for political content to be outside of our scope for this study due to challenges in measuring user perceptions of political ads [221]; further, problematic content [261], delivery [8] and policy [149] surrounding political ads are well-addressed in recent prior work.

The prevalence of each category in our annotated data is shown in Table 6.2. Figure 6.10 also shows concrete examples of each category. We leave a small fraction of ads (122, 0.41%) in our dataset uncategorized because they do not fit into our codebook, but are also not

Table 6.2 Prevalence of each code in our annotated dataset.

Code	**Count**	**%**
Neutral	20,596	68.52
Healthcare	3564	11.86
Opportunity	2267	7.54
Sensitive: Financial	1429	4.75
Sensitive: Other	631	2.10
Clickbait	1182	3.93
Deceptive	542	1.80
Potentially Prohibited	253	0.84
Political	263	0.87

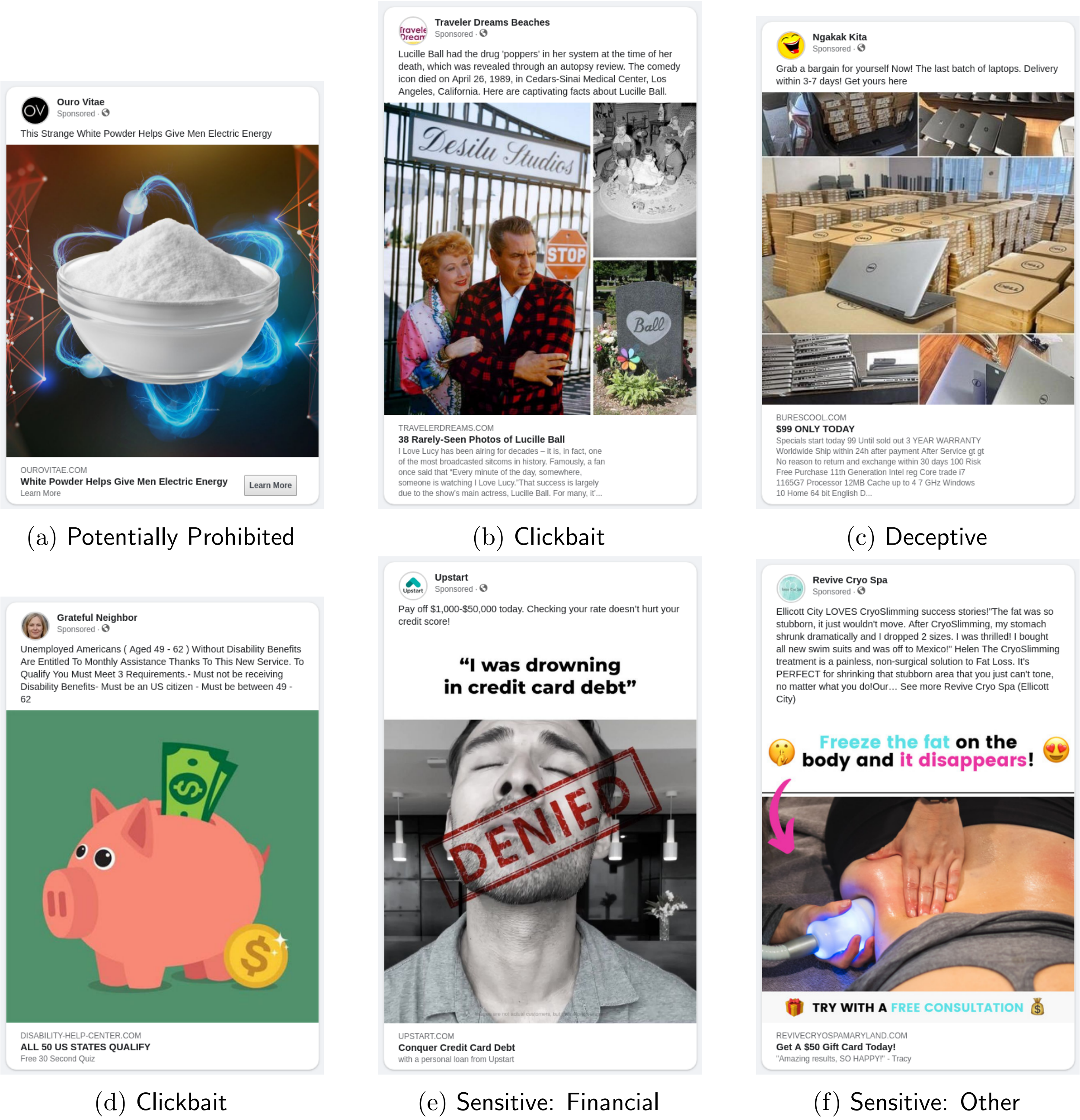
benign; often, these are potentially deceptive offers which we are unable to verify. Since some of our participants are recruited from Facebook, we observe an increased prevalence of research-study-related ads (2558, 7.85%). We use an auxiliary code "Study" to annotate all such ads, and remove them from all subsequent analyses.

In our annotation, we allowed for double-coding when an ad fell into two or more categories (e.g., an unclear ad for "5 Steps My Clients Use to Overcome Anxiety" falls into both Healthcare and Clickbait). However, we do not allow multiple codes when an ad is categorized as Neutral.

6.2.9 Coding Ads

Across all of our recruited participants, we collected 165,556 impressions to 88,509 unique ads. Out of these, 83,507 (94.3%) ads and 156,213 (94.3%) impressions were contributed by the participants ultimately deemed active (and considered in the remainder of the study). Due to the high volume of ads, we annotated a random subset of up to 200 ads per participant per month. Since we repeated this sampling strategy every month for each participant, we avoid introducing time- or participant-related sampling biases to the subset of our data we annotated. Through this sampling process, we were able to annotate 32,587 out of our collected 88,509 ads, or ≈36.8% of them.

The authors annotated the first two months of data. For the remaining months, we hired two students from our institute as external annotators. We choose to hire annotators locally instead of crowd-workers to be able to train them to use our codebook properly and



Ouro Vitae
Sponsored ·
This Strange White Powder Helps Give Men Electric Energy
OUROVITAE.COM
White Powder Helps Give Men Electric Energy
Learn More

(a) Potentially Prohibited

Traveler Dreams Beaches
Sponsored ·
Lucille Ball had the drug 'poppers' in her system at the time of her death, which was revealed through an autopsy review. The comedy icon died on April 26, 1989, in Cedars-Sinai Medical Center, Los Angeles, California. Here are captivating facts about Lucille Ball.

TRAVELERDREAMS.COM
38 Rarely-Seen Photos of Lucille Ball
I Love Lucy has been airing for decades – it is, in fact, one of the most broadcasted sitcoms in history. Famously, a fan once said that "Every minute of the day, somewhere, someone is watching I Love Lucy."That success is largely due to the show's main actress, Lucille Ball. For many, it'...

(b) Clickbait

Ngakak Kita
Sponsored ·
Grab a bargain for yourself Now! The last batch of laptops. Delivery within 3-7 days! Get yours here
BURESCOOL.COM
$99 ONLY TODAY
Specials start today 99 Until sold out 3 YEAR WARRANTY Worldwide Ship within 24h after payment After Service gt gt No reason to return and exchange within 30 days 100 Risk Free Purchase 11th Generation Intel reg Core trade i7 1165G7 Processor 12MB Cache up to 4 7 GHz Windows 10 Home 64 bit English D...

(c) Deceptive

Grateful Neighbor
Sponsored ·
Unemployed Americans (Aged 49 - 62) Without Disability Benefits Are Entitled To Monthly Assistance Thanks To This New Service. To Qualify You Must Meet 3 Requirements.- Must not be receiving Disability Benefits- Must be an US citizen - Must be between 49 - 62
DISABILITY-HELP-CENTER.COM
ALL 50 US STATES QUALIFY
Free 30 Second Quiz

(d) Clickbait

Upstart
Sponsored ·
Pay off $1,000-$50,000 today. Checking your rate doesn't hurt your credit score!

UPSTART.COM
Conquer Credit Card Debt
with a personal loan from Upstart

(e) Sensitive: Financial

Revive Cryo Spa
Sponsored ·
Ellicott City LOVES CryoSlimming success stories!"The fat was so stubborn, it just wouldn't move. After CryoSlimming, my stomach shrunk dramatically and I dropped 2 sizes. I was thrilled! I bought all new swim suits and was off to Mexico!" Helen The CryoSlimming treatment is a painless, non-surgical solution to Fat Loss. It's PERFECT for shrinking that stubborn area that you just can't tone, no matter what you do!Our… See more Revive Cryo Spa (Ellicott City)

Freeze the fat on the body and it disappears!
TRY WITH A FREE CONSULTATION

REVIVECRYOSPAMARYLAND.COM
Get A $50 Gift Card Today!
"Amazing results, SO HAPPY!" - Tracy

(f) Sensitive: Other

Figure 6.10: Example images of categories identified as problematic by our participants.

communicate in case of errors. The annotators were shown the ad's text and a screenshot of the ad (e.g. Figure 6.10) during annotation tasks.

Since our annotation task consists of multiple labels and we consider agreement for more than two annotators, we use Krippendorf's Alpha with the Jaccard set distance function to evaluate agreement between annotators. External annotators were first trained to use the codebook on a pilot task using the authors' gold standard annotations. Subsequently, every month, we picked a 5% subset of the month's ads to overlap across both annotators and the first author. If agreement on this common subset was low ($\alpha < 0.70$), we went over discrepancies and re-calibrated our use of the codebook. We repeated this exercise each month to ensure annotation quality remained high. The final agreement on our annotated data, $\alpha = 0.726$, is considered 'substantial' [148].

We specifically avoided using machine learning to avoid mis-labeling points in our data. Deceptive content, in particular, requires a level of investigation that would not be possible with automation. To investigate whether an ad is indeed deceptive, annotators are asked to visit the advertised web page, look at the advertiser's Facebook page, and inspect reviews on Facebook and Better Business Bureau.

Post-processing. Finally, while we annotate multiple codes per ad for a richly described dataset, we post-process our coding to translate into one code per ad. We do this for easier interpretation of the following results (Section 6.3), particularly in regression analyses. In line with the severity of restrictions in Facebook's policies [74], we translate sets of codes to a single code in the following precedence order:

Potentially Prohibited > Deceptive > Clickbait > Sensitive > Opportunity > Healthcare > Neutral.

6.3 Results

We now summarize our study's results. Section 6.3.1 identifies which categories of ads participants find problematic (RQ1). Section 6.3.2 investigates the distribution of problematic ads (RQ2). Section 6.3.3 examines the reasons for the discovered discrepancies (RQ3).

6.3.1 What do participants find problematic?

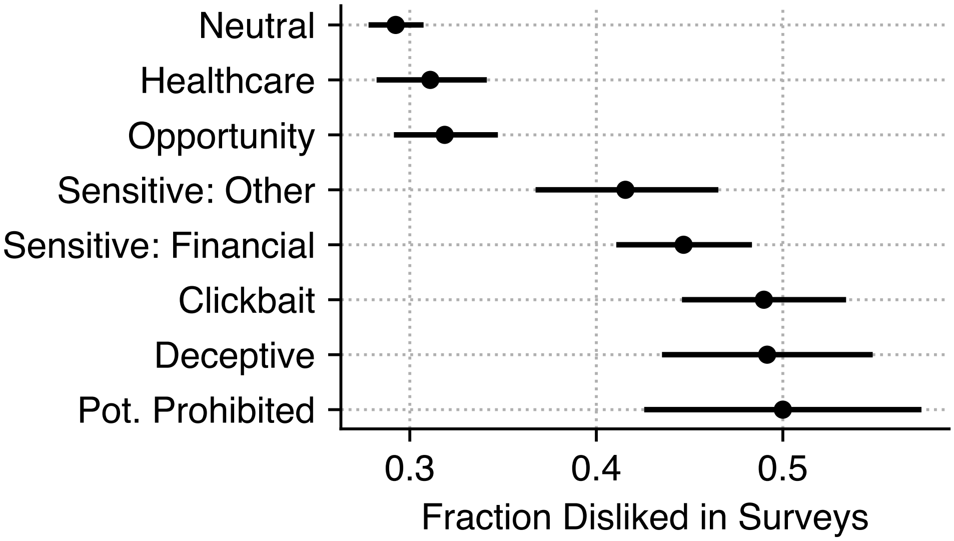


Figure 6.11: Fraction of responses where participants showed dislike for an ad category (i.e., chose "I do not like this ad" in the survey). 95% confidence intervals for (binomial) proportions are estimated via normal approximation.

To evaluate whether our participants found certain ad categories problematic, we first examine general dislike: whether participants dislike a higher fraction of particular ads. We then evaluate reasons for disliking: whether participants have different reasons for disliking each category in our codebook. Specifically, to evaluate general dislike, we use χ^2 proportion tests to evaluate differences in the proportion of ads in each category that participants marked as "I do not like this ad" in the second question of our survey (Section 6.1.2).

Figure 6.11 shows the fraction of responses, for each category, that were disliked by participants. Across our surveys, participants reported disliking nearly half of the ads we

Table 6.3 Odds ratios and 95% confidence intervals for mixed-effects logistic regression models, with a random effect term for respondents. Each model examines association between ad category and dislike reasons in survey responses. Each column shows shows one model, where dependent variable is the category (boolean) in the column header. Independent variable (rows) are respondents' binary responses for different dislike reasons. Each model is fit on responses for the category in the column and Neutral ads, so odds ratios should be interpreted as comparisons with the Neutral baseline. All highly disliked categories from Figure 6.11 are also modeled together in the "Problematic" column. $p < 0.001^{***}$; $p < 0.01^{**}$, $p < 0.05^{*}$.

Dislike Reason	Odds Ratio [95% CI]							
	Pot. Prohibited	**Deceptive**	**Clickbait**	**Sensitive: Financial**	**Sensitive: Other**	**Problematic**	**Opportunity**	**Healthcare**
intercept	0.03*** [0.02, 0.06]	0.036*** [0.02, 0.07]	0.103*** [0.07, 0.16]	0.16*** [0.11, 0.24]	0.052*** [0.03, 0.09]	0.403*** [0.29, 0.55]	0.236*** [0.17, 0.32]	0.231*** [0.17, 0.32]
advertiser	0.438 [0.16, 1.21]	0.701 [0.36, 1.37]	0.937 [0.55, 1.6]	0.679 [0.4, 1.14]	**2.101**** [1.22, 3.63]	0.99 [0.71, 1.39]	1.437 [0.95, 2.17]	1.098 [0.69, 1.76]
clickbait	0.657 [0.27, 1.58]	**2.465**** [1.37, 4.43]	**1.983**** [1.26, 3.13]	0.93 [0.58, 1.5]	1.265 [0.69, 2.32]	**1.472*** [1.07, 2.03]	1.124 [0.74, 1.7]	0.721 [0.44, 1.18]
design	0.734 [0.36, 1.49]	1.098 [0.58, 2.08]	0.935 [0.58, 1.5]	0.743 [0.48, 1.15]	1.168 [0.67, 2.05]	0.884 [0.65, 1.2]	1.175 [0.81, 1.71]	1.091 [0.73, 1.62]
irrelevant	1.677 [0.99, 2.85]	1.071 [0.68, 1.69]	1.043 [0.74, 1.48]	**1.574**** [1.15, 2.16]	1.327 [0.87, 2.02]	**1.34*** [1.07, 1.68]	**1.406*** [1.05, 1.87]	1.23 [0.91, 1.67]
politicized	2.754 [0.91, 8.37]	1.759 [0.61, 5.04]	1.359 [0.59, 3.16]	0.78 [0.3, 2.02]	0.15 [0.02, 1.27]	1.128 [0.62, 2.05]	0.818 [0.38, 1.75]	1.797 [0.85, 3.78]
product	0.832 [0.41, 1.7]	0.954 [0.56, 1.64]	1.078 [0.68, 1.7]	0.997 [0.66, 1.52]	**1.734*** [1.05, 2.88]	1.048 [0.78, 1.4]	0.987 [0.67, 1.45]	0.705 [0.45, 1.09]
pushy	0.747 [0.28, 1.97]	1.209 [0.6, 2.45]	0.682 [0.37, 1.26]	1.367 [0.83, 2.26]	0.499 [0.22, 1.16]	1.008 [0.7, 1.46]	**0.572*** [0.34, 0.95]	1.505 [0.96, 2.36]
scam	1.749 [0.98, 3.12]	**1.972**** [1.21, 3.21]	**1.473*** [1.01, 2.14]	**1.45*** [1.03, 2.05]	**2.078**** [1.34, 3.21]	**1.643***** [1.28, 2.1]	1.314 [0.96, 1.8]	0.894 [0.63, 1.28]
unclear	**1.891*** [1.02, 3.5]	0.566 [0.29, 1.12]	1.387 [0.91, 2.11]	1.109 [0.75, 1.64]	0.798 [0.46, 1.38]	1.137 [0.86, 1.51]	**0.55**** [0.37, 0.81]	**0.592*** [0.39, 0.9]
uncomfortable	1.603 [0.59, 4.39]	0.798 [0.33, 1.95]	1.382 [0.71, 2.69]	0.491 [0.22, 1.12]	0.642 [0.23, 1.82]	0.915 [0.56, 1.48]	1.274 [0.72, 2.26]	0.631 [0.3, 1.33]
N	1152	1213	1308	1386	1227	2018	1408	1359

had classified as Clickbait (48.98%), Deceptive (49.16%), and Potentially Prohibited (50%). Participants reported disliking 43.58% of the ads we coded as Sensitive, while Neutral, Healthcare and Opportunity ads were disliked less: 29.24%, 31.09%, and 31.87%, respectively.

These differences across ad categories are significant ($p < 0.001$, omnibus $\chi^2 = 186.25$; $\omega = 0.15$). In a series of pair-wise χ^2 proportion tests comparing each of our coded categories with Neutral, with Benjamini & Hochberg correction [24], we observe that Potentially Prohibited, Deceptive, Clickbait, and both types of Sensitive ads (Financial and Other) are all disliked significantly more than Neutral ads ($p < 0.001$, $\chi^2 > 24$; $0.07 \leq \omega \leq 0.13$). Opportunity ($p = 0.121$, $\chi^2 = 2.60$; $\omega < 0.10$) and Healthcare ($p = 0.28$, $\chi^2 = 1.14$; $\omega < 0.10$) ads, on the other hand, are not significantly more or less disliked than Neutral ads. To identify whether any of the ad categories are disliked more than each other (rather than just more than Neutral) we conduct an additional series of pair-wise corrected tests, comparing differences between sequential ad categories (e.g., comparing Potentially Prohibited, the most disliked category, with Deceptive, the next most disliked). This testing finds only one significant difference, between Sensitive: Other and Opportunity ($p = 0.003$, $\chi^2 = 11.34$; $\omega < 0.10$). In combination, our statistical results suggest that Clickbait, Deceptive, Potentially Prohibited, and Sensitive ads form an equivalence class of potentially problematic ads.

To understand *why* participants dislike these ad categories, we investigate the specific reasons they reported for disliking in the first survey question. Table 6.3 shows the odds ratios (exponentiated regression coefficients) of eight mixed-effects logistic regression models, with a random intercept for the participant. The odds ratios (O.R.) give the relative odds that an ad category was described with a certain dislike reason in survey responses, compared to the same dislike reason for our baseline (Neutral). For each ad category (column), an O.R. of 1 means a given dislike reason (row) is not used to describe the ad category more often than Neutral. Values greater than 1 correspond to increased odds of participants describing that ad category with the given reason, while values less than 1 indicate lower odds.

We first observe in Table 6.3 that participants are significantly more likely to describe the combined most highly disliked ad categories ("Problematic" column) as irrelevant (O.R. = 1.34, $p = 0.011$), clickbait (O.R. = 1.47, $p = 0.018$) and scam (O.R. = 1.64, $p < 0.001$). Looking at the disliked categories individually, we find that Deceptive, Clickbait and Sensitive ads are also significantly more likely to be described as scams (all O.R. $\geq$ 1.45, $p < 0.05$). The odds of Sensitive: Other ads, in particular, being described as scams are more than twice the odds of Neutral ads being described as scams (O.R. = 2.08, $p = 0.001$). Also for these ads, participants' odds of disliking the advertiser (O.R. = 2.10, $p = 0.007$) or product (O.R. = 1.73, $p = 0.032$) are significantly higher. Further, respondents find Potentially Prohibited ads to be unclear in their description (O.R. = 1.89, $p = 0.042$). Finally, our results find evidence that participants recognize the clickbait nature of the ads we categorize as Clickbait (O.R. = 1.98, $p = 0.003$), as well as those we categorize as more broadly Deceptive (O.R. = 2.46, $p = 0.002$), the latter of which are likely to use attention-grabbing content to lure people to click [132, 201].

Comparatively, the odds of Opportunity and Healthcare ads being described by participants as unclear are lower than the odds of a Neutral (all O.R. $\leq$ 0.55, $p < 0.05$). We also note that Opportunity ads, despite having higher odds of being described as irrelevant (O.R. = 1.4, $p = 0.020$), have lower odds of being described as pushy than Neutral ads.

Overall, we find differences in both rates of dislike, and reasons for disliking across our defined ad categories. Potentially Prohibited, Deceptive, Clickbait, and Sensitive ads are found to be disliked at a higher rate than other categories, and for more severe reasons beyond irrelevance: participants recognize their clickbait-y and scammy nature; dislike the sensitive products they advertise and the advertisers selling those products; and find them unclear, potentially due to advertisers evading platform prohibitions. As such, for the remainder of this study we refer to the collection of these four ad categories as Problematic.

6.3.2 How are Problematic ads distributed?

To understand how each ad category is distributed over our panel, we investigate the skew in its distribution over our participants: Figure 6.12 shows a cumulative distribution function (CDF) for all ads in each category. We also employ the Gini coefficient to precisely quantify this inequality. While highly recurrent impressions of ads are relatively sparse in our data—a median of two impressions per ad per participant—we account for the frequency of impressions in this analysis as well.

First, we observe that Neutral ads are not uniformly distributed, as observed by the distance from a uniform distribution. Because of this inherent skew in ad distribution, we treat Neutral (Gini = 0.48) as the baseline for comparison. Second, we see that Healthcare (Gini = 0.60) and Opportunity (Gini = 0.59) ads are more skewed (i.e., less uniformly distributed) than Neutral. This may be because Healthcare and Opportunity ads focus on narrower themes, and may be more personalized to users by advertisers or the platform. Third, we find that all five Problematic categories are more skewed across participants than Neutral. In these categories, we note the following order from least to most skewed: Sensitive: Other (Gini = 0.62), Sensitive: Financial (0.65), Clickbait (0.66), Potentially Prohibited (0.67), and Deceptive (0.69). To offer a concrete example of this skew: 80% of the Deceptive ad impressions (0.8 on y-axis) are delivered to just 36 participants (x-axis), compared to Healthcare, where the same fraction of impressions are delivered to 47 participants (or 60 participants in the case of Neutral).

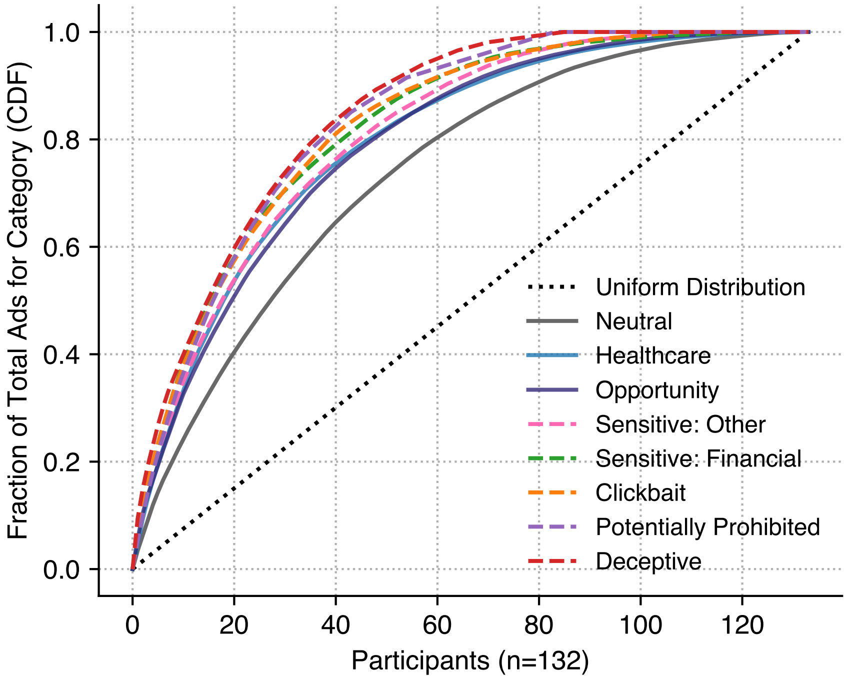


Figure 6.12: Cumulative Distribution Function (CDF) of impressions, showing what fraction of each ad category's total (y-axis) is contributed by how many participants (x-axis), given 132 total active participants.

Next, we focus on how individual-level exposure to Problematic ads vary for our participants. First, we note that data contributions themselves are inherently skewed, since participants have varying rates of Facebook use. To control for these differences, we look at the fraction of every participant's *ad diet*, i.e., all ads seen by them during the study, that consisted of Neutral vs. Problematic categories. Figure 6.13 shows the frequency distribution of these fractions across our panel.

We first observe that on average, a higher fraction of our panel's ad diet is composed of Neutral ads ($\mu = 0.71$, $\sigma = 0.12$), compared to Problematic ($\mu = 0.12$, $\sigma = 0.08$). Confirming our findings in the prior section, the distribution of Problematic has a heavier tail, suggesting that certain participants in our panel have increased exposure to these ads compared to the average. This observation is supported by measuring the skewness of these distributions, a statistical measure of asymmetry of a probability distribution. Recall that positive skew implies a distribution has a long right tail, while a negative

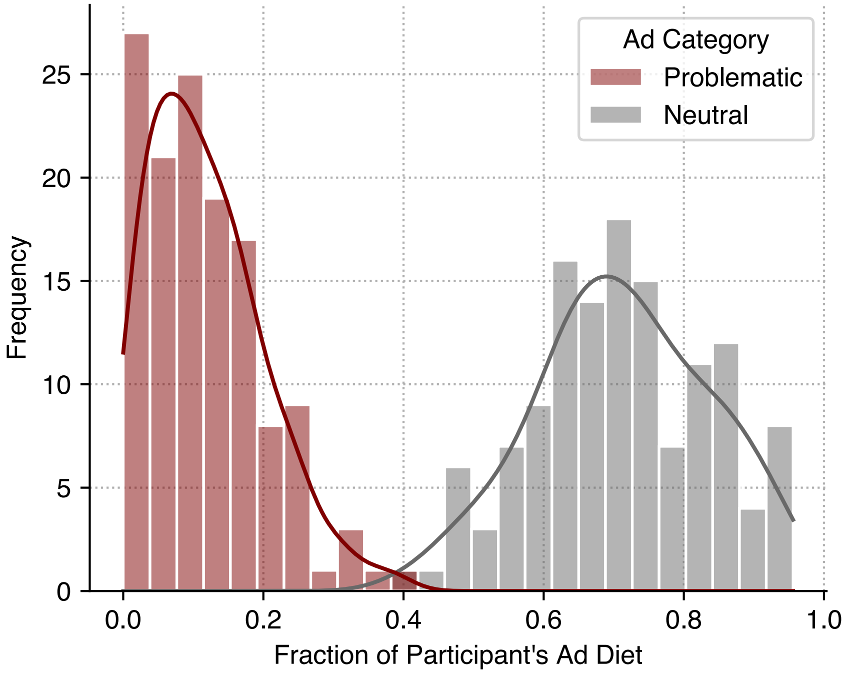


Figure 6.13: Fractions of exposure to Neutral and Problematic ads, out of participants' overall ad diet. We factor in frequency of seeing an ad while computing fractions. Smoothed lines are kernel density estimates (KDE) of the probability distribution.

skew means the left tail is longer. We measure the skewness for Neutral in Figure 6.13 as -0.11, and for Problematic as 0.84. These differences imply that despite the average exposure to Neutral ads in our panel being 71%, certain participants exist at the long left tail of this distribution, who are shown fewer Neutral ads, and a higher fraction of Problematic ads.

We next examine these participants who are shown a higher fraction of Problematic ads. Specifically, we investigate whether for any particular demographic groups, the Problematic ads constitute a higher fraction of ad diets. Table 6.4 shows coefficients of six linear models that we build to examine the relationship between participant demographics and fraction of Problematic ads among the ads they encountered. The intercept shows the average fraction in the ad diets of participants for whom all independent demographic variables are `false`, i.e., white, non-Hispanic men, born in 1980 or after, without a college degree. The proportion of these participants' ad diets that is composed of Problematic ads is 12% (first column in

Table 6.4 Coefficients of linear regression models, with 95% confidence intervals, modeling the relationship between exposure to Problematic ads and participants' demographics. Dependent variable (columns): fraction of ad type, out of total ad diet. Independent variable (rows): participant demographics. Union of all problematic ad types modeled in the Problematic column.
$p < 0.001^{***}$; $p < 0.01^{**}$, $p < 0.05^{*}$.

Variable	**Estimate (β)** [95% CI]					
	Problematic	**Pot. Prohibited**	**Deceptive**	**Clickbait**	**Sensitive: Financial**	**Sensitive: Other**
Intercept	0.12^{***} [0.09, 0.15]	0.01^{***} [0.01, 0.01]	0.008 [0, 0.02]	0.012 [0, 0.02]	0.07^{***} [0.04, 0.1]	0.02^{**} [0.01, 0.03]
Gender: Woman	$\mathbf{-0.064}^{***}$ [-0.09, -0.04]	-0.002 [0, 0]	-0.005 [-0.01, 0]	-0.008 [-0.02, 0]	$\mathbf{-0.045}^{***}$ [-0.07, -0.02]	-0.004 [-0.02, 0.01]
Race: Black	0.025 [-0.01, 0.06]	-0.001 [0, 0]	0.006 [0, 0.02]	$\mathbf{0.013}^{*}$ [0, 0.02]	0.004 [-0.02, 0.03]	0.002 [-0.01, 0.02]
Race: Asian	-0.002 [-0.04, 0.04]	0.001 [0, 0.01]	-0.003 [-0.02, 0.01]	0.005 [-0.01, 0.02]	-0.007 [-0.04, 0.03]	0.002 [-0.02, 0.02]
Ethnicity: Hispanic	0.023 [-0.03, 0.08]	$\mathbf{-0.007}^{*}$ [-0.01, 0]	0.005 [-0.01, 0.02]	-0.007 [-0.03, 0.01]	0.036 [-0.01, 0.08]	-0.003 [-0.02, 0.02]
Education: college and above	0.01 [-0.02, 0.04]	-0.002 [0, 0]	0.004 [-0.01, 0.01]	0.01 [0, 0.02]	-0.003 [-0.03, 0.02]	0 [-0.01, 0.01]
Age: Gen-X and older	$\mathbf{0.051}^{***}$ [0.02, 0.08]	$\mathbf{-0.003}^{*}$ [-0.01, 0]	$\mathbf{0.011}^{*}$ [0, 0.02]	$\mathbf{0.017}^{**}$ [0.01, 0.03]	0.017 [-0.01, 0.04]	0.009 [0, 0.02]

Table 6.4). All statistically significant coefficients in the table mark biases in comparison to that baseline.

We find that the ad diets of older participants, born before 1980, are (additively) composed of 5.1% more Problematic ads (CI: 2-8%) than younger participants. Women's ad diets are composed of 6.4% fewer Problematic ads (CI: 4-9%) than those who do not identify as women—largely because women see 4.5% fewer Sensitive: Financial ads (CI: 2-7%). We also note that older participants' ad diets are composed of higher fractions of Deceptive (1.1%, CI: 0-2%), and Clickbait ads (1.3%, CI: 1-3%). Ad diets of Black participants contain 1.3% (CI: 0-2%) more Clickbait ads than those of white or Asian participants in our panel. However, older participants and Hispanic participants ad diets have slightly lower fraction of Potentially Prohibited ads, 0.3% (CI: 0-1%) and 0.7% (CI: 0-1%) respectively, potentially because these ads target products assumed by advertisers or the platforms not to be of interest to these groups. To account for possible variance in participants' privacy behavior (e.g. changing ad preferences), we model their awareness of privacy settings as an additional independent variable in Table 6.5. We find that privacy awareness does not have any significant effect on the disparate exposure that we observe, and demographic skews similar to those in Table 6.4

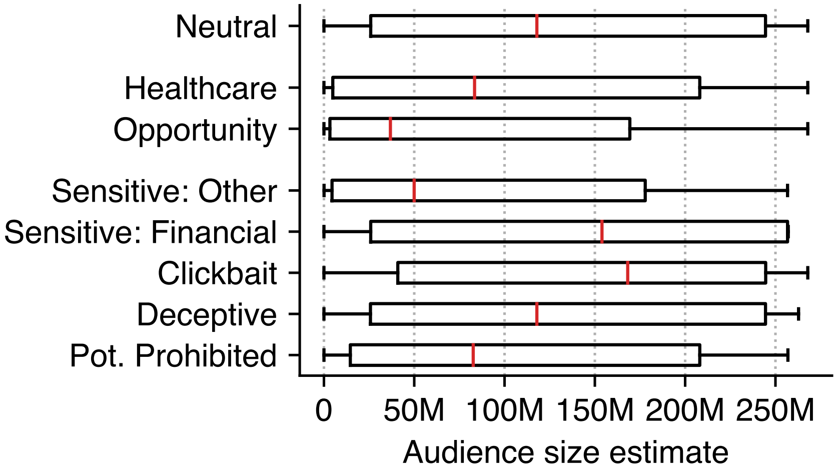


Figure 6.14: Audience size distributions of different ad categories. The red vertical lines mark the median audience size, the box indicates the 25th and 75th percentile, and the whiskers extend from the box by 1.5x of the inter-quartile range (IQR).

persist. Demographic skews for other ad categories are also shown in Table 6.5.

Table 6.5 Coefficients of two regression models, modeling the relationship between exposure to Healthcare and Opportunity ads, and participant demographics. Analysis setup similar to Table 6.4.
$p < 0.001^{***}$; $p < 0.01^{**}$, $p < 0.05^{*}$, $p < 0.1^{+}$.

Variable	**Estimate (β)** [95% CI]	
	Healthcare	**Opportunity**
Intercept	0.089^{***} [0.06, 0.11]	0.045^{**} [0.01, 0.08]
Gender: Woman	0.007 [-0.01, 0.03]	0.022^{+} [0, 0.05]
Race: Black	-0.024^{+} [-0.05, 0]	$\mathbf{0.038}^{*}$ [0.01, 0.07]
Race: Asian	-0.017 [-0.05, 0.01]	0.028 [-0.01, 0.07]
Education: college and above	-0.004 [-0.03, 0.02]	$\mathbf{0.034}^{*}$ [0.01, 0.06]
Ethnicity: Hispanic	0.011 [-0.03, 0.05]	-0.007 [-0.06, 0.04]
Age: 42 and above	0.018^{+} [0, 0.04]	-0.021 [-0.05, 0.01]

6.3.3 Who is responsible for skews?

With a better understanding of which participants have increased exposure to problematic ads, we next identify the reasons behind these differences. As discussed in Section 2.4.2, whether a particular user sees an ad on Facebook is affected by two main factors: (a) the user has to be among the audience targeted by the advertiser; (b) Facebook's ad delivery optimization considers the ad relevant to the user, which contributes to it winning an auction [87]. Thus, one can expect that when the advertiser targets a larger audience, the delivery optimization has more influence in selecting the actual audience. With this intuition, we start by investigating audience size across our ad categories.

As described in Section 6.1.3, we query Facebook's APIs to obtain audience sizes for each of our collected ads— Figure 6.14 shows the distributions of these audience sizes broken down by ad category. Observing Problematic categories, we find that the median target audience sizes for Sensitive: Financial (153.9M) and Clickbait (168.2M) ads are larger than for Neutral ads (117.9M); a pairwise Kruskal-Wallis [143] test rejected the null hypothesis that the medians are equal ($p = 0.001$ for both tests). This implies that Facebook exercises more control for picking the audience subset for these categories. On the other hand, median audience sizes for Potentially Prohibited (82.6M) and Sensitive: Other (49.9M) ads are significantly smaller

Table 6.6 Coefficients of linear regression models, modeling the relationship between exposure to Problematic ads and participants' demographics and privacy behavior, similar to Table 6.4. $p < 0.001^{***}$; $p < 0.01^{**}$, $p < 0.05^{*}$, $p < 0.1^{+}$.

Variable	**Estimate** (β) [95% CI]					
	Problematic	**Pot. Prohibited**	**Deceptive**	**Clickbait**	**Sensitive: Financial**	**Sensitive: Other**
Intercept	0.106* [0.02, 0.19]	0.011* [0, 0.02]	-0.001 [-0.03, 0.03]	0.026 [-0.01, 0.06]	0.056 [-0.01, 0.13]	0.015 [-0.02, 0.05]
Gender: Woman	**-0.066***** [-0.09, -0.04]	-0.002 [0, 0]	-0.005 [-0.01, 0]	-0.007 [-0.02, 0]	**-0.047***** [-0.07, -0.03]	-0.005 [-0.02, 0.01]
Race: Black	0.028^{+} [0, 0.06]	-0.001 [0, 0]	0.006 [0, 0.02]	**0.013*** [0, 0.02]	0.007 [-0.02, 0.03]	0.003 [-0.01, 0.02]
Race: Asian	-0.001 [-0.04, 0.04]	0.001 [0, 0.01]	-0.002 [-0.02, 0.01]	0.004 [-0.01, 0.02]	-0.007 [-0.04, 0.03]	0.003 [-0.01, 0.02]
Ethnicity: Hispanic	0.037 [-0.01, 0.09]	**-0.007*** [-0.01, 0]	0.006 [-0.01, 0.02]	-0.01 [-0.03, 0.01]	**0.048*** [0.01, 0.09]	0 [-0.02, 0.02]
Education: college and above	0.009 [-0.02, 0.04]	-0.001 [0, 0]	0.004 [-0.01, 0.01]	**0.011*** [0, 0.02]	-0.004 [-0.03, 0.02]	0 [-0.01, 0.01]
Age: 42 and above	**0.052***** [0.02, 0.08]	**-0.003*** [-0.01, 0]	**0.012*** [0, 0.02]	**0.017**** [0.01, 0.03]	0.018 [-0.01, 0.04]	0.009 [0, 0.02]
Privacy Settings Awareness (1–5)	0.003 [-0.02, 0.02]	0 [0, 0]	0.002 [0, 0.01]	-0.003 [-0.01, 0]	0.003 [-0.01, 0.02]	0.001 [-0.01, 0.01]

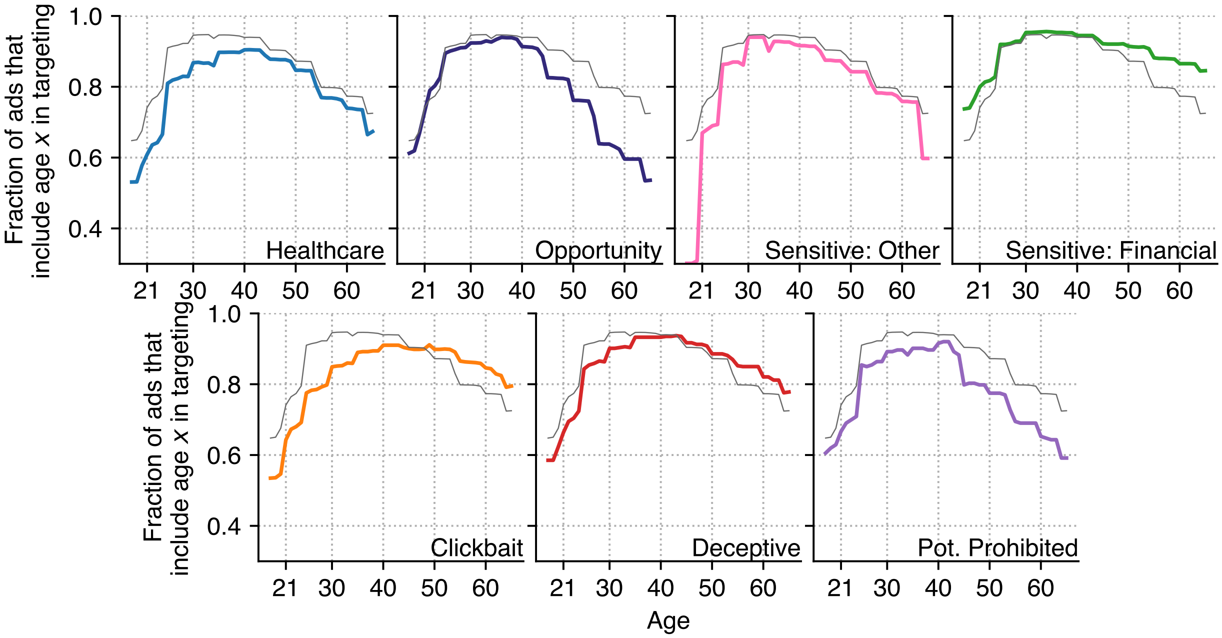


Figure 6.15: Fraction of ads that include given age ranges in their targeting. The thin line in each panel shows the fraction among Neutral ads for easier comparison.

than Neutral ($p = 0.006$ and $p < 0.001$, respectively), indicating that advertisers for these ads more precisely specify the audiences they want to reach. We also note that audience sizes for Opportunity (36.8M) and Healthcare (83.4M), considered non-problematic in this study, are actually smaller than Neutral ($p < 0.001$).

Next, we investigate what targeting options advertisers use to scope these various audiences. We find that the most used targeting option is age: nearly half the ads use some form of age targeting (49.7%). Around a quarter of ads use Custom Audiences [73] (25.6%) and platform-inferred user interests (26.9%); On the other hand, advertisers for 21.2% of the ads in our dataset don't change the targeting criteria at all, and use the default targeting of all U.S. adults (267 million users). Finally, we find that only 12.1% ads in our data specifically target by gender; a vast majority use the default option of targeting all genders. Note that these percentages do not sum up to 100% because each ad can be targeted using multiple targeting criteria. Below, we detail how age, custom audiences, interests and default targeting are used in our data.

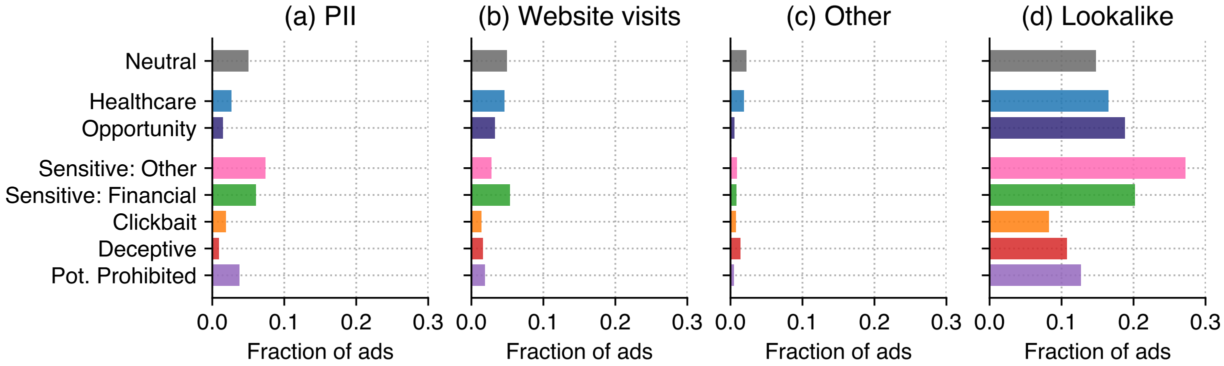


Figure 6.16: Prevalence of four types of Custom Audiences. Advertisers rely on automated tools such as (d) Lookalike Audiences more than manual audience configuration.

Table 6.7 Coefficients of linear regression models, with 95% confidence intervals, modeling relationship between exposure to problematic ads *due to platform optimization*, and participants' demographics. Dependent variable (columns): fraction of category, out of total ad diet of ads with default/no advertiser targeting. Independent variable (rows): participant demographics. $p < 0.001^{***}$; $p < 0.01^{**}$, $p < 0.05^{*}$.

Variable	**Estimate (β)** [95% CI]					
	Problematic	**Pot. Prohibited**	**Deceptive**	**Clickbait**	**Sensitive: Financial**	**Sensitive: Other**
Intercept	0.191*** [0.13, 0.26]	0.013*** [0.01, 0.02]	0.014* [0, 0.03]	0.023 [-0.01, 0.05]	0.133*** [0.08, 0.18]	0.009* [0, 0.02]
Gender: Woman	**-0.059*** [-0.11, -0.01]	**-0.006*** [-0.01, 0]	-0.007 [-0.02, 0]	-0.003 [-0.03, 0.02]	**-0.046*** [-0.09, 0]	0.004 [0, 0.01]
Race: Black	0.01 [-0.05, 0.07]	0.002 [0, 0.01]	0.007 [-0.01, 0.02]	0.011 [-0.02, 0.04]	-0.007 [-0.06, 0.04]	-0.003 [-0.01, 0]
race: Asian	-0.019 [-0.1, 0.06]	-0.005 [-0.01, 0]	-0.003 [-0.02, 0.01]	-0.007 [-0.04, 0.03]	-0.003 [-0.07, 0.06]	0 [-0.01, 0.01]
Ethnicity: Hispanic	0.017 [-0.08, 0.12]	-0.009 [-0.02, 0]	**0.028**** [0.01, 0.05]	-0.021 [-0.06, 0.02]	0.027 [-0.05, 0.11]	-0.008 [-0.02, 0]
Education: college and above	-0.033 [-0.09, 0.02]	-0.002 [-0.01, 0]	0 [-0.01, 0.01]	0.005 [-0.02, 0.03]	-0.036 [-0.08, 0.01]	-0.001 [-0.01, 0.01]
Age: Gen-X and older	**0.077**** [0.02, 0.13]	-0.003 [-0.01, 0]	0.011 [0, 0.02]	**0.041**** [0.02, 0.06]	0.034 [-0.01, 0.08]	-0.005 [-0.01, 0]

Age. Figure 6.15 shows the fraction of ads that include users of a given age in their targeting; fractions of all ages are presented together as a line, which can be perceived as a function of age. Each panel shows this function for a different ad category, and also features the function for Neutral ads for easier comparison. A category-specific line above the (gray) Neutral line signifies that the age group was more often targeted with ads of that category compared to Neutral ads. Focusing on Problematic categories, ads for Sensitive:

Other often exclude users aged 18-21. This can be explained by the prevalence of ads for alcoholic beverages in this category, selling of which to individuals below 21 is illegal in the US. Sensitive: Financial, Clickbait and Deceptive ads include older audiences at a higher rate than Neutral ads, which could explain why Deceptive and Clickbait skews towards older users in our panel. Similarly, Potentially Prohibited ads also exclude users over the age of 45. These differences provide evidence that advertisers actively use the platform's age targeting features to find older users to show clickbait and scam content to. This is notable, since prior work suggests that older users may be more susceptible to such content [178].

Custom Audiences. We make an distinction between custom audiences where the advertiser provides Facebook with a list of particular individuals to target using their PII (e.g., phone number, email), and Lookalike Audiences [86] that Facebook creates by finding users similar to those that the advertiser provides. The distinction is crucial because of the difference in control: the advertiser exercises complete control over who to include in the first group; however, they have little influence over the characteristics of the lookalike audiences. Figure 6.16 shows the prevalence of different types of custom audiences per ad category. We observe that lookalike audiences are used more often than PII custom audiences for all categories. We also note that as many as a quarter of Sensitive: Other ads were targeted using Lookalike Audiences. This suggests that while advertisers use the platform's tool to find vulnerable audiences (e.g., Figure 6.15), they often outsource this role to the platform, especially when targeting for sensitive themes like weight loss or gambling.

Interests. Precise targeting by inferred interests is one of the features that distinguishes online behavioral advertising from traditional advertising models. A total of 6,028 unique interests were used to target our participants, including highly specific and sensitive inferences pertaining to health ("Multiple sclerosis awareness", "Fibromyalgia awareness"), sexuality ("LGBT community", "Gay Love"), religion ("Evangelicalism", "Judaism"), and others. It is perhaps surprising that a majority of ads in our dataset (73.1%) do not actually use this functionality. Table 6.8 shows the most commonly targeted interests for each ad category.

Default Targeting. Finally, we investigate the delivery of ads that used the default targeting (i.e., the advertiser included all U.S. adults in their target audience). This allow us to observe the behavior of the delivery optimization in cases where the skew can not be

Table 6.8 Most popular targeting interests by category. We see that a majority of ads are not targeted by interests.

Ad Category	Targeted Interest (Prevalence)
Neutral	**None** (**72.3%**), Online shopping (1.3%), Health & wellness (0.8%), Family (0.7%), Physical fitness (0.7%), Yoga (0.6%)
Opportunity	**None** (**67.9%**), Employment (2.7%), Education (2.4%), Higher education (2.3%), Career (1.7%), Technology (1.6%)
Healthcare	**None** (**76.8%**), Health & wellness (2.4%), Clinical trial (2.2%), Physical fitness (1.7%), Physical exercise (1.5%), Medicine (1.0%)
Clickbait	**None** (**79.8%**), Online shopping (1.3%), Personal finance (1.0%), Amazon.com (0.9%), Home improvement (0.8%), Investment (0.8%)
Sensitive: Financial	**None** (**76.4%**), Personal finance (5.1%), Investment (3.4%), Online banking (3.1%), Credit cards (2.9%), Financial services (2.0%)
Sensitive: Other	**None** (**75.3%**), Gambling (2.9%), Alcoholic beverages (2.4%), Bars (2.1%), Beer (1.9%), Vodka (1.4%)
Pot. Prohibited	**None** (**81.3%**), Health & wellness (2.7%), Meditation (1.8%), Physical fitness (1.4%), Credit cards (1.4%), House Hunting (1.4%)
Deceptive	**None** (**68.1%**), Online shopping (4.0%), Shopping (2.1%), Amazon.com (1.5%), Clothing (1.5%), Digital marketing (1.5%)

attributed to the advertiser's actions. To identify skews in delivery, we run a series of linear models, shown in Table 6.7, to examine the relation between fraction of problematic ads in ad diets and participant demographics, similar to Section 6.3.2. In contrast to that analysis, however, we subset our data to only include ads that have default targeting from the advertiser. Therefore, for each participant, we model, say, the fraction of Clickbait they saw that had default targeting, out of all of their default-targeted ads. Consequently, we capture purely skews that arise due to the platform's optimization, since the advertiser specified the broadest possible targeting, and Facebook had to make its judgment of a relevant audience. Again, the first row (intercept) shows the fraction of ad diets for participants who are non-Hispanic white, younger, and without a college education; all significant coefficients mark biases in comparison to that baseline.

Table 6.7 shows that (similar to Table 6.4), the effect for older participants seeing a 7.7% higher fraction of Problematic ads (CI: 2-13%), and women seeing 5.9% fewer of them (CI: 1-11%), persists, even without advertiser targeting. Specifically, older participants' ad diets (additively) contain 4.1% (CI: 2-6%) more Clickbait than the younger participants. We also observe a novel effect of Hispanic participants seeing 2.8% more Deceptive ads (CI: 1-5%).

This implies that while their overall ad diets might not contain a significantly higher fraction of scams (Table 6.4)—delivery optimization independently skews these ads towards Hispanic participants. In further nuance, the effect of women seeing fewer Problematic ads can be explained by their ad diets comprising of 4.6% fewer Sensitive: Financial ads (CI: 0-9%), and 0.6% fewer Potentially Prohibited ads (CI: 0-1%) compared to participants who don't identify as women. These differences provide evidence that in addition to an advertiser's targeting—or regardless of it—Facebook's delivery optimization algorithms are also responsible for skewing the delivery of Problematic ads.

6.4 Discussion

Our study presents three main contributions. *First*, gathering insights from a diverse group of Facebook users, we identify a collection of Problematic categories of ads that were significantly more disliked, and determine participants' reasons for disliking these ads—they often mistrust these ads and recognize their deceptive nature. *Second*, we observe that while these ads make up a small fraction (12% on average) of our participants' ad diets, a subset of our panel are disproportionately exposed to them. *Third*, using a combination of techniques, we demonstrate that some of these skews in ad distribution persist without targeting from advertisers, implying that the platform's algorithms are responsible for at least some of the skews we observe.

While our observations are limited to our panel, our study validates anecdotal evidence [123, 166] that clickbait and scam advertising is shown to older users more often. We show that these differences exist both due to advertisers' targeting and due to the platform's delivery optimization—which together may create a feedback loop [152]. We also identify instances where the overall outcomes are different than delivery optimization's biases: Black participants see a higher fraction of Clickbait ads (Table 6.4), but only when targeted by advertisers. On the other hand, Hispanic participants have higher exposure to Deceptive ads (Table 6.7), but only within ads that are essentially untargetted by advertisers, suggesting this effect is due to the ad platform.

Further, we find that financial ads are shown more often to participants who identify as men, both as a system-level outcome, and when controlling for ad targeting. As annotators, we observe that Sensitive: Financial ads are quite diverse—ranging from problematic offers like high APR loans to possibly useful financial tools such as savings accounts. Thus, men in our panel are exposed to problematic financial products, as well as financial opportunities, more often.

Finally, our analysis of targeting practices shows that advertisers often cede control to the platform's optimizations – as evidenced by the popular use of lookalike audiences (Figure 6.16) and the low usage of targeting interests (Table 6.8). This implies that advertisers are aware of the usefulness of the platform's personalization, and malicious actors could rely on these capabilities to target Problematic advertising.

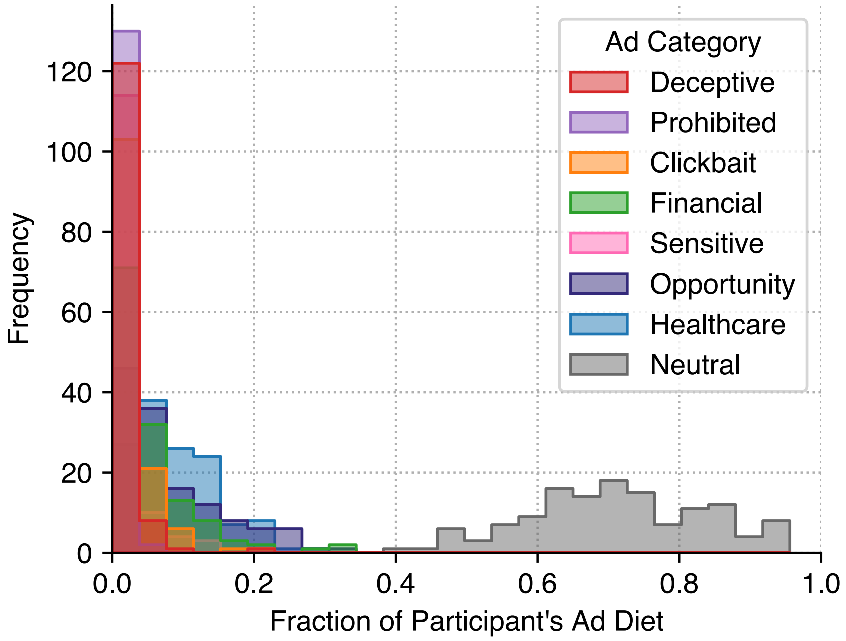


Figure 6.17: Fractions of exposure to different ad categories, out of participants' overall ad diet. Problematic ads are personalized similar to others in our codebook.

Taken together, our results offer concrete insights into user experiences with problematic advertising and raise questions about the power of platforms in delivering these ads to users.

Limitations. Our ad categories were created through pilot data collection and backed by review of platform policies and literature, including work that also examined user sentiments towards problematic advertising [259]. Still, categorizing ads into just seven categories diminishes some nuance within groups. We analyze a subset of our total collected ads that we were able to annotate manually (one-third of our overall collected data); therefore, we are not able to provide insight into the complete ad diets of our participants. To minimize any selection biases in our analyzed subset, we randomly sampled ads from participants each month for annotating and surveying, but recognize important data could be missed by not assessing the complete ad diets of participants.

Further, our observations are only about participants' desktop browsing experiences. While we suspect that similar ads would be present on the mobile Facebook app due to the diversity of Facebook's ad placement options, we do not have direct access to that data. We also do not have access to budgets of the ads that we observe, and therefore are not able to disambiguate whether certain advertisers are simply paying more money to Facebook, resulting in skews. However, to control for these differences, we compare fractions of ad categories out of the ad diets that we observe for each participant (e.g., in Section 6.3.2). This ensures that we compare only within participants' desktop experiences, and in the same budget-class of advertisers that were reaching them.

Additionally, we do not have access to participants' complete ad preferences, and the frequency with which they change these settings. This limits our ability to control for participant actions such as removing ads from an advertiser, or removing a specific interest. Prior work estimates that 10-19% of users tweak their ad settings [112, 127], either from the ad preferences page or from the contextual menu next to ads. We attempt to account for such variance by factoring participants' awareness of privacy settings in Table 6.5, and find that disparate exposure to Problematic ads for older and minority participants persists.

Finally, our work currently does not provide insight on advertising's contextual harms [174]; for instance, while we take an interest in sensitive ads with subject matters like gambling, we do not investigate their distribution among those with gambling addictions. Rather, we try to find commonalities in our panel's opinions through mixed-effects regression models, and then build our analysis on top of that data. We leave further exploration of contextually problematic ads, such as Gak et al. [100], to future work.

Recommendations. To limit users' exposure to problematic ads, we propose changes on two levels. *First*, we advocate for a more fine-grained and user-informed understanding of problematic ads, and other broader harms of advertising [7]. Currently, platforms recognize ads such as Deceptive, Clickbait and Potentially Prohibited as problematic, and typically include language scrutinizing them in their advertising guidelines [71, 76]. However, sensitive ads that present harms for users with addictions or other mental illness are less moderated. Yet, they are still widely disliked across our diverse set of participants. We advocate for a more refined understanding of ads with sensitive themes, and more scrutiny and moderation from platforms towards these ads. For a more nuanced understanding of problematic ads, our work, along with [259] and [100] provide a start.

Second, we argue for more controls not just on moderation, but on optimization as well.

Our results demonstrate that once problematic ads circumvent a platform's review process, the platform then optimizes them towards users similar to other personalized content (e.g. Figure 6.17). To avoid this systematic personalizing of problematic ads, platforms need policies on their delivery optimization in addition to their policies on content moderation. This would require platforms to constrain the optimization of problematic content for users. For instance, Facebook currently states that it demotes clickbait in content ranking [76], yet a demotion does not stop such content from inevitably reaching and harming some users. There is perhaps a need for an "optimization vacuum" so that problematic content, even after evading moderation, cannot reach users.

We advocate for platforms to take emerging works on user experiences with problematic ads into account, and for a more urgent call for platforms to not only moderate the content users see, but also have mechanisms to suppress the delivery of problematic content, instead of optimizing for it.

Chapter 7

Concluding Discussion

The thesis of this dissertation is that *personalization in online platforms can have adverse effects for users in sensitive content domains, and algorithm auditing methods can be used to measure these effects without privileged access to the platforms.* I have demonstrated this thesis to be true across a variety of content domains within Facebook's advertising system. In Chapter 4, I build methods to measure how personalization skews the delivery of job and housing ads along users' race and gender, even when the advertiser targets inclusively. Chapter 5 presents methods to measure how personalization limits the delivery of political advertising during elections, due to opaque judgments about the relevance of such content to targeted users. Finally, Chapter 6 shows through observational data how personalization is responsible for exposing vulnerable users to problematic ad types such as clickbait and scams. In this chapter, I revisit the details of these contributions, highlight the impact this work has had, and provide insights for future work in this research area.

7.1 Summary of Contributions

The highest level contributions of this work include **(a)** experimental and observational *methods* to measure the effect of personalization in targeted advertising; **(b)** application of these methods to *domains* where personalization could harm users, resulting in **(c)** concrete *measurements* of the harms of personalization in Facebook's advertising system. Below, I summarize these three axes of contributions in detail.

Methods. Throughout this dissertation, we build a series of methods to measure the adverse effects of personalization in the delivery of Facebook ads. In summary:

1. We introduce an experimental method to run pairs of ads that target the same audience, are run at the same time, and with the same bidding strategy and budget (Section 4.1.4). We leverage this method throughout our work to isolate the role of personalization in the differential delivery of two or more ads. Further, this approach can be extended to any ad platform that allows ad targeting with personally identifiable information (PII).

2. We build a novel method that exploits the location of an ad's audience to measure ad delivery along user attributes for which Facebook does not provide reporting (Section 4.1.5). We employ this method to obtain audience breakdowns by race, political party affiliation, as well as users' political donations. This approach can also be extended to any ad platform that allows PII targeting and reports geographical breakdown of ad delivery to advertisers.

3. We propose two novel methods to control for human interactions with the personalization mechanisms, such as moderator input or user reactions to our ads. This includes an approach to modify the alpha channel of an ad's image (Section 4.2.3), as well as a method to serve different HTML content to Facebook's crawlers compared to the users (Section 5.1.3). Both of these methods allow creating ads that look identical to humans, but contain content for Facebook's systems to estimate user relevance, thereby completely isolating the role of content in personalization.

4. We also build methods to disentangle the role of personalization in an observational setting by exploiting Facebook's ad transparency mechanisms. Specifically, we use the "Why am I seeing this?" feature to obtain the advertiser's original targeting specification, and use its specificity to reason about the role of targeting vs. personalization for thousands of real-world ads collected from users (Section 6.3.3).

Domains. We apply these methods to understand the impact of personalization in *four* distinct domains of sensitive content: (i) job ads, (ii) housing ads, (iii) political ads, (iv) problematic ads as reported by users. Personalizing content in each of these domains carries a different form of harm:

1. For jobs and housing, which we study in Chapter 4, limiting the delivery of such opportunities by users' gender or race is considered discriminatory in many jurisdictions, including United States [1–3, 231] and the European Union [28].

2. For political ads, while direct regulation for online platforms does not exist, similar regulation from a different domain, such as the FCC's Equal Time rule [4, 91] enforces that TV and broadcast media make air time available to political candidates on equivalent terms. Such regulation specifically exists to limit broadcast licensees from unduly influencing the outcome of elections. Online platforms can similarly adversely affect the outcomes of elections by introducing price differentiation and limiting the reach of political ads, as we see in Chapter 5.

3. Finally, studying user experiences of ads provides insight into content that might not be regulated, but has a presence on the platform, and is problematic according to the users themselves. Overexposing a minority of users to deceptive ads, as we see in Chapter 6, can harm them by leading to loss of finances or personal information, even when the majority of users might not be experiencing such problematic content.

Measurements. Through a series of large-scale measurement studies, we concretely quantify how users are adversely impacted by personalization in Facebook's advertising system. In summary:

1. For job advertising, we find drastic skews by the race and gender of users, despite targeting the exact same set of users, running ads at the same time of day, and specifying the same budget and optimization strategy. For example, job ads for janitor roles skew on average to an audience that is nearly 75% female, and nearly 60% Black; in contrast, ads for lumber jobs skew to an audience over 90% male and nearly 75% white. We hypothesize that these skews likely affect real job ads on Facebook, and therefore are systematically excluding users from seeing valuable job opportunities.

2. For housing ads, using our methodology for measuring racial breakdowns, we find that certain ads for home buying opportunities are delivered to over 85% white users. In contrast, certain opportunities to rent deliver to only 35% white users, and instead primarily to Black users.

3. For political content, we measure that Facebook's personalization skews—or rather limits—the delivery of election campaign ads based on its judgment of the users' political alignment. We find that these skews are replicated across dozens of geographies throughout the U.S. Moreover, forcing the platform to show an ad to a purportedly non-aligned audience leads to a price premium ranging from 50–300%, even when the ads look identical to the users themselves.

4. In our study of individual user experiences, we find that multiple forms of problematic ads exist on Facebook's platform—including clickbait, deceptive offers, as well as sensitive themes like weight loss or gambling. We show that such ads form a small fraction (12%) in the average participant's experience, but is significantly more concentrated for older and male participants. Further, these skews persist even within ads which have the broadest possible targeting from the advertiser. Therefore, personalization is able to find users who might be vulnerable [177] and over-expose them to problematic content.

7.1.1 Broader Impact

The work in this dissertation has created awareness about the role of algorithms in the discriminatory outcomes of advertising, and therefore in other applications of personalization. This awareness has been communicated by publications such as The *Washington Post* [224], *The Economist* [89], and *MIT Technology Review* [117], and others [25, 65, 160]. The scientific merit of this work has also been recognized within the research community: our work at *CSCW 2019* (Chapter 4) was a recipient of an *Honorable Mention* as well as the *Diversity and Inclusion Award*. I was also awarded an *Outstanding Graduate Student in Research Award* at Northeastern University in recognition of my work's impact.

I have been invited to Congress to brief the Financial Services Committee of the U.S. House of Representatives, and to present this work at the U.S. Federal Trade Commission (at *PrivacyCon* 2020). The work here has also been presented at the European Parliament for proposed legislation on the transparency of political advertising in the EU [209]. Finally, this work identified behaviors that led the U.S. Department of Justice (DOJ) to charge Meta with violations of the Fair Housing Act (FHA)—the first time that DOJ has *ever* charged an algorithm with FHA violations [236]. This independent investigation by the DOJ is in line with the results shown in Chapter 4. DOJ and Meta subsequently announced a settlement in 2022, where Meta agreed to deploy a novel Variance Reduction System (VRS) [167, 230], which specifically reduces racial disparities for housing ads in the ad delivery phase.

7.2 Future Work

The insights gained from this dissertation open up opportunities for multiple continuing and new directions of research.

Mitigating the harms of personalization. Similar to the current literature where fairness goals in machine learning are achieved through constraints on optimization [34, 217], I propose designing constraints corresponding to the set of user values obtained through surveys (e.g. in Chapter 6) and integrating them into the objective of the personalization task. As an example, Celis et al.'s [34] framework provides a natural fit for such a goal, where each sensitive advertising domain can constitute a content *group*, and users' sentiments towards the domain can be used to judge the strength of the constraints. Other frameworks such as Mehrotra et al.'s [165] also allow constraints to satisfy secondary stakeholders in a personalization system, which in this case are the users. I argue that encoding the perceptions of real users into personalization can help ensure that we are not solving towards a contrived fairness goal, but something that is eventually valued by users themselves [7].

Alternative, recommender system can rely on recently proposed modeling approaches, such as using using multi-vector representations, which specifically capture the diverse multiple interests of each user [111]. This approach has been shown to improve the *reachability* of items in a personalization setup, i.e. users can be recommended more diverse domains of content instead of a narrow subset, as we see in this dissertation.

User autonomy in ad controls. Ad platforms often present user controls or ad preferences as defense mechanisms against the adverse effects of personalization. If there are types of content that the user does not want to be exposed to, they can simply remove those interests or advertisers from their preferences [41]. However, prior work has shown that users are often not aware of these controls and therefore don't exercise them [112, 127]. Recent work also documents users feeling a lack of autonomy about the data inferred about them, and an inability to control these inferences [255]. Further, case studies question the efficacy of ad controls overall [97], showing that even when users modify their ad settings, alternative, correlated inferences about their interests continue to be made. Given these developments in prior work, it is valuable for future work to investigate whether users are indeed able to modify their ad controls and gain autonomy of their experiences. A concrete measurement showing the effectiveness of ad preferences would help with accountability of advertising platforms.

Variation across countries. Much of the work in this dissertation focuses on discrimination, politics, as well as user demographics in the context of the United States. It is valuable to extend this work to other geographies particularly in the Global South, where notions of race, gender, and discrimination are different. Our methods are extensible to other ad platforms, and can be used to measure other user demographics as long as PII data is available.

Furthermore, future work should investigate gaps in content quality and moderation standards as well. We anecdotally note from our study [10] that a few participants who signed up from Nigeria saw noticeably worse ads, including phishing attempts. Since our study focuses on U.S. participants only, we filter out data from these international participants. However, we leave a thorough investigation of ad quality variance and different forms of problematic ads in other countries to future work.

Recommendations of organic content. While our focus in this work is on ads, there is increasingly a need to understand personalization within organic content recommendations. On Facebook, users are frequently presented "Suggested" content by the platform; these posts are not chosen by a specific advertiser, but rather by Facebook's own judgment of relevance. On TikTok as well, "For You" constitutes an entirely separate feed of videos. Organic content does not have a targeting phase, and therefore, does not have transparency explanations, making it more challenging to study. However, the distribution of such content can be studied in an observational setting by working with users, as we do in Chapter 6. We propose future work to investigate which content is chosen to be recommended, and how it leads to varying user experiences.

Bibliography

[1] *12 CFR § 202.4 (b) – Discouragement.* https://www.law.cornell.edu/cfr/text/12/202.4.

[2] *24 CFR § 100.75 – Discriminatory advertisements, statements and notices.* https://www.law.cornell.edu/cfr/text/24/100.75.

[3] *29 USC § 623 – Prohibition of age discrimination.* https://www.law.cornell.edu/uscode/text/29/623.

[4] *47 USC §315 – Candidates for public office.* https://www.law.cornell.edu/uscode/text/47/315.

[5] Abubakar Abid, Maheen Farooqi, and James Zou. "Persistent Anti-Muslim Bias in Large Language Models". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 298–306.

[6] Alan Agresti and Brent A Coull. "Approximate is better than "exact" for interval estimation of binomial proportions". In: *The American Statistician* 52.2 (1998), pp. 119–126.

[7] Muhammad Ali. "Measuring and mitigating bias and harm in personalized advertising". In: *Proceedings of the 15th ACM Conference on Recommender Systems*. 2021, pp. 869–872.

[8] Muhammad Ali et al. "Discrimination through Optimization: How Facebook's Ad Delivery can Lead to Biased Outcomes". In: *Proceedings of ACM Conference on Computer-Supported Cooperative Work (CSCW)*. 2019.

[9] Muhammad Ali et al. "Ad Delivery Algorithms: the Hidden Arbiters of Political Messaging". In: *Proceedings of ACM Conference on Web Search and Data Mining (WSDM)*. 2021.

[10] Muhammad Ali et al. "Problematic Advertising and its Disparate Exposure on Facebook". In: *Proceedings of the 32nd USENIX Security Symposium*. USENIX Association. 2023.

[11] Laura Edelson et al. *An Analysis of United States Online Political Advertising Transparency*. https://arxiv.org/abs/1902.04385.

[12] *An Update on Senator Kyl's Review of Potential Anti-Conservative Bias*. https://newsroom.fb.com/news/2019/08/update-on-potential-anti-conservative-bias/.

[13] BuzzFeedNews. *Analysis of ActBlue contributions to presidential candidates, January-June 2020*. https://github.com/BuzzFeedNews/2019-08-actblue-donations.

[14] Athanasios Andreou et al. "Investigating ad transparency mechanisms in social media: a case study of facebook's explanations". In: *Proc. of Network and Distributed System Security (NDSS) Symposium*. 2018.

[15] Athanasios Andreou et al. "Measuring the facebook advertising ecosystem". In: *Proc. of Network and Distributed System Security (NDSS) Symposium*. 2019.

[16] Julia Angwin and Terry Parris Jr. *Facebook Lets Advertisers Exclude Users by Race*. ProPublica, https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race/. Oct. 2016.

[17] Jennifer Stevens Aubrey. "Exposure to sexually objectifying media and body self-perceptions among college women: an examination of the selective exposure hypothesis and the role of moderating variables". In: *Sex Roles* 55.3 (2006), pp. 159–172.

[18] Theophilus Azungah. "Qualitative research: deductive and inductive approaches to data analysis". In: *Qualitative research journal* (2018).

[19] Christopher A Bail et al. "Exposure to opposing views on social media can increase political polarization". In: *Proceedings of the National Academy of Sciences* 115.37 (2018), pp. 9216–9221.

[20] Stephanie Alice Baker, Matthew Wade, and Michael James Walsh. "The challenges of responding to misinformation during a pandemic: content moderation and the limitations of the concept of harm". In: *Media International Australia* 177.1 (2020), pp. 103–107.

[21] Eytan Bakshy, Solomon Messing, and Lada A Adamic. "Exposure to ideologically diverse news and opinion on facebook". In: *Science* 348.6239 (2015), pp. 1130–1132.

[22] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. http://www.fairmlbook.org. fairmlbook.org, 2019.

[23] Muhammad Ahmad Bashir et al. "Tracing information flows between ad exchanges using retargeted ads". In: *25th USENIX Security Symposium (USENIX Security 16)*. 2016, pp. 481–496.

[24] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1 (1995), pp. 289–300.

[25] Sam Biddle. *FACEBOOK'S AD ALGORITHM IS A RACE AND GENDER STEREOTYPING MACHINE, NEW STUDY SUGGESTS*. https://theintercept.com/2019/04/03/facebook-ad-algorithm-race-gender/. 2019.

[26] Jesús Bobadilla et al. "Recommender systems survey". In: *Knowledge-based systems* 46 (2013), pp. 109–132.

[27] Tolga Bolukbasi et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings". In: vol. 29. 2016.

[28] Silvia Borelli. "Article 23 EU Charter of Fundamental Rights: Equality between women and men". In: *International and European Labour Law*. 2018, pp. 209–214.

[29] Sherryl Browne Graves. "Television and prejudice reduction: when does television as a vicarious experience make a difference?" In: *Journal of Social Issues* 55.4 (1999), pp. 707–727.

[30] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 77–91.

[31] Twitter Business. *About Twitter Ads approval.* https://business.twitter.com/en/help/ads-policies/introduction-to-twitter-ads/about-twitter-ads-approval.html.

[32] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186.

[33] *CBS, Inc. v. FCC.* https://supreme.justia.com/cases/federal/us/453/367/.

[34] L Elisa Celis et al. "Controlling polarization in personalization: an algorithmic framework". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 2019, pp. 160–169.

[35] Meta Business Center. *About data sampling.* https://www.facebook.com/business/help/1691983057707189?id=354406972049255&helpref=faq_content, Accessed: September 4, 2023.

[36] Meta Business Help Center. *About Ad Delivery.* https://www.facebook.com/business/help/1000688343301256.

[37] Meta Business Help Center. *About Ad Principles.* https://www.facebook.com/business/about/ad-principles.

[38] Meta Business Help Center. *About advertising objectives.* https://www.facebook.com/business/help/517257078367892.

[39] Meta Business Help Center. *Ads About Social Issues, Elections or Politics.* https://www.facebook.com/business/help/208949576550051.

[40] Meta Business Help Center. *Creating Customer Lists with Mobile Advertiser IDs.* https://developers.facebook.com/docs/app-ads/targeting/mobile-advertiser-ids/.

[41] Meta Business Help Center. *Facebook Help: Ad Preferences.* https://www.facebook.com/help/109378269482053/.

[42] Meta Business Help Center. *Help: Choosing a Special Ad Category.* https://www.facebook.com/business/help/298000447747885.

[43] Meta Business Help Center. *How to avoid posting clickbait on Facebook*. https://www.facebook.com/business/help/503640323442584?id=208060977200861.

[44] Meta Business Help Center. *Marketing API*. https://developers.facebook.com/docs/marketing-apis/.

[45] Meta Business Help Center. *Showing Relevance Scores for Ads on Facebook*. https://www.facebook.com/business/news/relevance-score.

[46] *Certify Compliance to Facebook's Non-Discrimination Policy*. https://www.facebook.com/business/help/136164207100893.

[47] Yanto Chandra and Liang Shang. "Inductive coding". In: *Qualitative research using R: A systematic approach*. Springer, 2019, pp. 91–106.

[48] Le Chen et al. "Investigating the Impact of Gender on Rank in Resume Search Engines". In: *Proc. of CHI*. Montreal, Canada, Apr. 2018.

[49] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, 2013.

[50] Paul Covington, Jay Adams, and Emre Sargin. "Deep neural networks for youtube recommendations". In: *Proceedings of the 10th ACM conference on recommender systems*. 2016, pp. 191–198.

[51] Geff Cumming and Sue Finch. "Inference by eye: confidence intervals and how to read pictures of data." In: *American Psychologist* 60.2 (2005), p. 170.

[52] Amit Datta, Michael Carl Tschantz, and Anupam Datta. "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination". In: *Proc. of PETS*. 2015.

[53] Amit Datta et al. "Discrimination in online advertising: a multidisciplinary inquiry". In: *Conference on Fairness, Accountability and Transparency*. PMLR. 2018, pp. 20–34.

[54] James Davidson et al. "The YouTube video recommendation system". In: *Proceedings of the fourth ACM conference on Recommender systems*. 2010, pp. 293–296.

[55] Sarah Dean, Sarah Rich, and Benjamin Recht. “Recommendations and user agency: the reachability of collaboratively-filtered information”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 436–445.

[56] Nicholas Diakopoulos et al. “I vote for—how search informs our choice of candidate”. In: *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple*. Ed. by M. Moore and D. Tambini. Oxford University Press, 2018.

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

[58] Michael Dimock. *Defining generations: Where Millennials end and Generation Z begins.* https://www.pewresearch.org/fact-tank/2019/01/17/where-millennials-end-and-generation-z-begins/.

[59] *Doing More to Protect Against Discrimination in Housing, Employment and Credit Advertising*. https://newsroom.fb.com/news/2019/03/protecting-against-discrimination-in-ads/.

[60] *Doing More to Protect Against Discrimination in Housing, Employment and Credit Advertising*. https://newsroom.fb.com/news/2019/03/protecting-against-discrimination-in-ads/.

[61] Cynthia Dwork and Christina Ilvento. “Fairness Under Composition”. In: *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*. Vol. 124. Leibniz International Proceedings in Informatics (LIPIcs). 2018, 33:1–33:20. ISBN: 978-3-95977-095-8. DOI: 10.4230/LIPIcs.ITCS.2019.33. URL: http://drops.dagstuhl.de/opus/volltexte/2018/10126.

[62] Cynthia Dwork et al. “Fairness through awareness”. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ACM. 2012, pp. 214–226.

[63] Dean Eckles, Brett R Gordon, and Garrett A Johnson. “Field studies of psychologically targeted ads face threats to internal validity”. In: *Proceedings of the National Academy of Sciences* 115.23 (2018), E5254–E5255.

[64] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. “Internet advertising and the generalized second-price auction: selling billions of dollars worth of keywords”. In: *American Economic Review* 97.1 (2007), pp. 242–259.

[65] Gilad Edelman. *How Facebook's political ad system is designed to polarize*. https://www.wired.com/story/facebook-political-ad-system-designed-polarize/. 2019.

[66] Laura Edelson, Tobias Lauinger, and Damon McCoy. “A security analysis of the facebook ad library”. In: *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2020, pp. 661–678.

[67] Michael D. Ekstrand et al. “Fairness in information access systems”. In: *Foundations and Trends in Information Retrieval* 16.1-2 (2022), pp. 1–177. DOI: 10.1561/1500000079. URL: https://doi.org/10.1561%2F1500000079.

[68] *eyeWnder_Experiment*. http://www.eyewnder.webredirect.org/.

[69] Facebook. *Ad Standards, Online Pharmacies*. https://transparency.fb.com/policies/ad-standards/content-specific-restrictions/online-pharmacies.

[70] Facebook. *Advertising Standards, Deceptive Content*. https://transparency.fb.com/policies/ad-standards/deceptive-content/.

[71] Facebook. *Advertising Standards, Unacceptable Business Practices*. https://transparency.fb.com/policies/ad-standards/deceptive-content/unacceptable-business-practices.

[72] Facebook. *Advertising Standards, Unrealistic Outcomes*. https://transparency.fb.com/policies/ad-standards/deceptive-content/unrealistic-outcomes.

[73] Facebook. *Bout Custom Audiences*. https://www.facebook.com/business/help/744354708981227?id=2469097953376494.

[74] Facebook. *Facebook Advertising Standards*. https://transparency.fb.com/policies/ad-standards/.

[75] Facebook. *Facebook Terms of Service, Jobs Policies*. https://www.facebook.com/policies_center/Jobs/.

[76] Facebook. *Types of Content We Demote*. https://transparency.fb.com/features/approach-to-ranking/types-of-content-we-demote.

[77] *Facebook Ad Library*. https://www.facebook.com/ads/library/.

[78] *Facebook Ad Platform*. https://www.facebook.com/business.

[79] Meta Business Help Center. *Facebook Ads Manager*. https://www.facebook.com/business/help/200000840044554.

[80] *Facebook Advertising Policies, Discriminatory Practices*. https://www.facebook.com/policies/ads/prohibited_content/discriminatory_practices.

[81] *Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says*. https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html.

[82] *Facebook Motion to Dismiss in Onuoha v. Facebook*. https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.34.0.pdf.

[83] *Why Am I Seeing This? We Have an Answer for You*. https://about.fb.com/news/2019/03/why-am-i-seeing-this/.

[84] *Facebook Political Ad Collector*. https://projects.propublica.org/facebook-ads/.

[85] *Facebook: About Lookalike Audiences*. https://www.facebook.com/business/help/164749007013531.

[86] Facebook. *Facebook: About Lookalike Audiences*. https://www.facebook.com/business/help/164749007013531?id=401668390442328.

[87] Meta Business Help Center. *Facebook: About the delivery system: Ad auctions*. https://www.facebook.com/business/help/430291176997542.

[88] *Facebook's Ad Archive API is Inadequate*. https://blog.mozilla.org/blog/2019/04/29/facebooks-ad-archive-api-is-inadequate/.

[89] *Facebook's ad system seems to discriminate by race and gender*. https://www.economist.com/business/2019/04/04/facebooks-ad-system-seems-to-discriminate-by-race-and-gender. 2019.

[90] Irfan Faizullabhoy and Aleksandra Korolova. "Facebook's advertising platform: new attack vectors and the need for interventions". In: (Mar. 2018). https://arxiv.org/abs/1803.10099, Workshop on Technology and Consumer Protection (ConPro).

[91] *FCC Political Programming Rules*. https://www.fcc.gov/sites/default/files/political_programming_fact_sheet.pdf, Accessed: September 6, 2023.

[92] *Federal Election Commission — Individual Contributions*. https://www.fec.gov/data/receipts/individual-contributions.

[93] Michael Feldman et al. "Certifying and removing disparate impact". In: *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 259–268.

[94] Seth Flaxman, Sharad Goel, and Justin M. Rao. "Filter Bubbles, Echo Chambers, and Online News Consumption". In: *Public Opinion Quarterly* 80.S1 (Mar. 2016), pp. 298–320.

[95] Roger Allan Ford. "Data scams". In: *Houston Law Review* 57 (2019), p. 111.

[96] Mozilla Foundation. *HUD Sues Facebook Over Housing Discrimination and Says the Company's Algorithms Have Made the Problem Worse*. https://foundation.mozilla.org/en/youtube/findings/; Accessed: Aug 22, 2023.

[97] Panoptykon Foundation. *Algorithms of trauma: new case study shows that Facebook doesn't give users real control over disturbing surveillance ads*. https://en.panoptykon.org/algorithms-of-trauma. 2021.

[98] Erika Franklin Fowler et al. *Political Advertising Online and Offline*. 2021. DOI: 10.1017/S0003055420000696.

[99] FTC. *Challenging Deceptive Advertising and Marketing*. URL: https://www.ftc.gov/reports/annual-report-standard/ftc-2013/challenging-deceptive-advertising-and-marketing.

[100] Liza Gak, Seyi Olojo, and Niloufar Salehi. "The distressing ads that persist: uncovering the harms of targeted weight-loss ads among users with histories of disordered eating". In: *Proceedings of ACM Conference on Computer-Supported Cooperative Work (CSCW)*. New York, NY, USA: Association for Computing Machinery, Nov. 2022.

[101] Yingqiang Ge et al. "Understanding echo chambers in e-commerce recommender systems". In: *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval.* 2020, pp. 2261–2270.

[102] Andrew Gelman, Jennifer Hill, and Masanao Yajima. "Why we (usually) don't have to worry about multiple comparisons". In: *Journal of Research on Educational Effectiveness* 5.2 (2012), pp. 189–211.

[103] Avijit Ghosh, Giridhari Venkatadri, and Alan Mislove. "Analyzing facebook political advertisers' targeting". In: *Workshop on Technology and Consumer Protection (ConPro).* 2019.

[104] Emily Glazer. *Facebook Weighs Steps to Curb Narrowly Targeted Political Ads.* https://www.wsj.com/articles/facebook-discussing-potential-changes-to-political-ad-policy-11574352887. Nov. 2019.

[105] Cristos Goodrow. *On YouTube's recommendation system.* https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/. 2021.

[106] Google. *Google Ads policies.* URL: https://support.google.com/adspolicy/answer/6008942?hl=en.

[107] Google. *Google: About similar audiences on the Display Network.* https://support.google.com/google-ads/answer/2676774?hl=en.

[108] Google. *Google: An update on our political ads policy.* https://www.blog.google/technology/ads/update-our-political-ads-policy.

[109] Ben Green and Yiling Chen. "Disparate interactions: an algorithm-in-the-loop analysis of fairness in risk assessments". In: *Conference on Fairness, Accountability and Transparency.* 2019.

[110] Joshua Green and Sasha Issenberg. *Inside the Trump Bunker, With Days to Go.* https://www.bloomberg.com/news/articles/2016-10-27/inside-the-trump-bunker-with-12-days-to-go. 2016.

[111] Wenshuo Guo et al. "The stereotyping problem in collaboratively filtered recommender systems". In: *Equity and Access in Algorithms, Mechanisms, and Optimization.* 2021, pp. 1–10.

[112] Hana Habib et al. “Identifying user needs for advertising controls on Facebook”. In: *Proceedings of ACM Conference on Computer-Supported Cooperative Work (CSCW)*. ACM New York, NY, USA, 2022.

[113] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. “Burst of the filter bubble? effects of personalization on the diversity of google news”. In: *Digital journalism* 6.3 (2018), pp. 330–343.

[114] Aniko Hannak et al. “Measuring personalization of web search”. In: *Proc. of The Web Conference*. 2013.

[115] Aniko Hannak et al. “Measuring price discrimination and steering on e-commerce web sites”. In: *Proc. of the Internet Measurement Conference*. 2014.

[116] Aniko Hannak et al. “Bias in online freelance marketplaces: evidence from taskrabbit and fiverr”. In: *ACM Conference on Computer Supported Cooperative Work*. 2017.

[117] Karen Hao. *Facebook’s ad-serving algorithm discriminates by gender and race*. https://www.technologyreview.com/2019/04/05/1175/facebook-algorithm-discriminates-ai-bias/. 2019.

[118] Muhammad Haroon et al. “YouTube, the great radicalizer? Auditing and mitigating ideological biases in YouTube recommendations”. In: *arXiv preprint arXiv:2203.10666* (2022).

[119] Google Ads Help. *About Customer Match*. https://support.google.com/adwords/answer/6379332?hl=en.

[120] Jacob B Hirsh, Sonia K Kang, and Galen V Bodenhausen. “Personalized persuasion: tailoring persuasive appeals to recipients’ personality traits”. In: *Psychological science* 23.6 (2012), pp. 578–581.

[121] Homa Hosseinmardi et al. “Examining the consumption of radical content on youtube”. In: *Proceedings of the National Academy of Sciences* 118.32 (2021), e2101967118.

[122] Anna Marie Houlis. *Part 238, Guides Against Bait Advertising*. URL: https://www.sheknows.com/health-and-wellness/articles/2002320/targeted-advertising-trauma/.

[123] Craig Silverman. *How A Massive Facebook Scam Siphoned Millions Of Dollars From Unsuspecting Boomers*. https://www.buzzfeednews.com/article/craigsilverman/facebook-subscription-trap-free-trial-scam-ads-inc. 2019.

[124] Megan Graham and Jennifer Elias. *How Google's $150 billion advertising business works*. https://www.cnbc.com/2021/05/18/how-does-google-make-money-advertising-business-breakdown-.html. 2021.

[125] *Engineering at Meta: How machine learning powers Facebook's News Feed ranking algorithm*. https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking/. 2021.

[126] *HUD Sues Facebook Over Housing Discrimination and Says the Company's Algorithms Have Made the Problem Worse*. https://www.propublica.org/article/hud-sues-facebook-housing-discrimination-advertising-algorithms.

[127] Jane Im et al. "Less is not more: improving findability and actionability of privacy controls for online behavioral advertising". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 2023, pp. 1–33.

[128] *Improving Enforcement and Promoting Diversity: Updates to Ads Policies and Tools*. http://newsroom.fb.com/news/2017/02/improving-enforcement-and-promoting-diversity-updates-to-ads-policies-and-tools/.

[129] GDI Global Disinformation Index. *The Quarter Billion Dollar Question: How is Disinformation Gaming Ad Tech?* URL: https://www.disinformationindex.org/research/2019-9-1-the-quarter-billion-dollar-question-how-is-disinformation-gaming-ad-tech/.

[130] Ray Jiang et al. "Degenerate feedback loops in recommender systems". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 383–390.

[131] Matthew Joseph et al. "Fairness in learning: classic and contextual bandits". In: (2016).

[132] Chris Kanich et al. "Spamalytics: an empirical analysis of spam marketing conversion". In: *Proceedings of the 15th ACM Conference on Computer and Communications Security*. 2008, pp. 3–14.

[133] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. "Unequal representation and gender stereotypes in image search results for occupations". In: *Proc. of CHI*. 2015.

[134] Larry Kim. *The Most Expensive Keywords in Google AdWords*. http://www.wordstream.com/blog/ws/2011/07/18/most-expensive-google-adwords-keywords/. 2011.

[135] Michael P. Kim et al. *Preference-Informed Fairness*. https://arxiv.org/abs/1904.01793. 2019.

[136] Sara Kingsley et al. "Auditing digital platforms for discrimination in economic opportunity advertising". In: *arXiv preprint arXiv:2008.09656* (2020).

[137] Kate Konger. *Twitter Will Ban All Political Ads, C.E.O. Jack Dorsey Says*. https://www.nytimes.com/2019/10/30/technology/twitter-political-ads-ban.html. 2019.

[138] Yehuda Koren, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems". In: *Computer, IEEE Computer Society* 42.8 (2009), pp. 30–37.

[139] Aleksandra Korolova. "Privacy violations using microtargeted ads: a case study". In: *2010 IEEE International Conference on Data Mining Workshops* 3.1 (2011), pp. 27–49.

[140] Aleksandra Korolova. *Facebook's Illusion of Control over Location-Related Ad Targeting*. Medium. https://medium.com/@korolova/facebooks-illusion-of-control-over-location-related-ad-targeting-de7f865aee78. Dec. 2018.

[141] Michal Kosinski, David Stillwell, and Thore Graepel. "Private traits and attributes are predictable from digital records of human behavior". In: *Proceedings of the National Academy of Sciences* 110.15 (2013), pp. 5802–5805.

[142] Daniel Kreiss and Shannon C. Mcgregor. "The "Arbiters of What Our Voters See": Facebook and Google's Struggle with Policy, Process, and Enforcement around Political Advertising". In: *Political Communication* 0.0 (2019), pp. 1–24.

[143] William H Kruskal and W Allen Wallis. "Use of ranks in one-criterion variance analysis". In: *Journal of the American Statistical Association* 47.260 (1952), pp. 583–621.

[144] Juhi Kulshrestha et al. “Quantifying search bias: investigating sources of bias for political searches in social media”. In: *ACM Conference on Computer Supported Cooperative Work*. 2017.

[145] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. “Interpretable decision sets: a joint framework for description and prediction”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1675–1684.

[146] Michelle Lam et al. “Sociotechnical audits: broadening the algorithm auditing lens to investigate targeted advertising”. In: *Proceedings of ACM Conference on Computer-Supported Cooperative Work (CSCW)*. 2023.

[147] Anja Lambrecht and Catherine E. Tucker. *Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2852260. 2018.

[148] J Richard Landis and Gary G Koch. “The measurement of observer agreement for categorical data”. In: *biometrics* (1977), pp. 159–174.

[149] Victor Le Pochat et al. “An Audit of Facebook’s Political Ad Policy Enforcement”. In: *Proceedings of the 31st USENIX Security Symposium*. USENIX Association. 2022.

[150] Mathias Lecuyer et al. “XRay: Enhancing the Web’s Transparency with Differential Correlation”. In: *Proc. of USENIX Security Symposium*. 2014.

[151] Mathias Lecuyer et al. “Sunlight: fine-grained targeting detection at scale with statistical confidence”. In: *Proc. of Conference on Computer and Communications Security (CSS)*. 2015.

[152] Liu Leqi and Sarah Dean. “Engineering a safer recommender system”. In: *Workshop on Responsible Decision Making in Dynamic Environments*. 2022.

[153] Zhou Li et al. “Knowing your enemy: understanding and detecting malicious web advertising”. In: *Proceedings of the 2012 ACM conference on Computer and communications security*. 2012, pp. 674–686.

[154] *LinkedIn Marketing Solutions: Matched Audiences*. https://business.linkedin.com/marketing-solutions/ad-targeting/matched-audiences.

[155] Yabing Liu et al. “Measurement and Analysis of OSN Ad Auctions”. In: *Proc. of Conference on Online Social Networks (COSN)*. 2014.

[156] Zhuang Liu et al. “Contrastive learning for recommender system”. In: *arXiv preprint arXiv:2101.01317* (2021).

[157] Julia Angwin et al. *Machine Bias*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[158] Leonardo Madio and Martin Quinn. “Content moderation and advertising in social media platforms”. In: *Social Science Research Network* (2021).

[159] Saranya Maneeroj and Atsuhiro Takasu. “Hybrid recommender system using latent features”. In: *2009 International Conference on Advanced Information Networking and Applications Workshops*. IEEE. 2009, pp. 661–666.

[160] Louise Matsakis. *Facebook’s ad system might be hard-coded for discrimination*. https://www.wired.com/story/facebooks-ad-system-discrimination/. 2019.

[161] Sandra C Matz et al. “Psychological targeting as an effective approach to digital mass persuasion”. In: *Proceedings of the national academy of sciences* 114.48 (2017), pp. 12714–12719.

[162] Sandra C Matz et al. “Reply to eckles et al.: facebook’s optimization algorithms are highly unlikely to explain the effects of psychological targeting”. In: *Proceedings of the National Academy of Sciences* 115.23 (2018), E5256–E5257.

[163] James McInerney et al. “Explore, exploit, and explain: personalizing explainable recommendations with bandits”. In: *Proceedings of the 12th ACM Conference on Recommender Systems*. 2018, pp. 31–39.

[164] Stacy Jo Dixon. *Media usage in an Internet minute as of April 2022*. https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/, Accessed: September 10, 2023.

[165] Rishabh Mehrotra et al. “Towards a fair marketplace: counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems”. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*. 2018, pp. 2243–2251.

[166] Jeremy Merrill and Hanna Kozlowska. *A precious-metals scheme used fear and Facebook to trick older conservatives out of their savings.* https://qz.com/1749695/facebook-ads-targeted-fox-news-fans-for-shady-silver-coin-scheme. 2019.

[167] Meta. *Meta: An Update on Our Ads Fairness Efforts.* https://about.fb.com/news/2023/01/an-update-on-our-ads-fairness-efforts/. 2023.

[168] Meta. *Advertising Policies.* https://www.facebook.com/policies_center/ads.

[169] Meta. *Advertising Policies.* URL: https://www.facebook.com/policies_center/ads.

[170] Meta. *New Updates to Reduce Clickbait Headlines.* URL: https://about.fb.com/news/2017/05/news-feed-fyi-new-updates-to-reduce-clickbait-headlines/.

[171] *Meta Business Help Center: Age and Gender.* https://www.facebook.com/business/help/151999381652364, Accessed: September 1, 2023.

[172] *Meta Pixel: Measure, Optimize and Retarget Ads.* https://www.facebook.com/business/tools/meta-pixel.

[173] Danaë Metaxa et al. “Auditing algorithms: understanding algorithmic systems from the outside in”. In: *Foundations and Trends in Human–Computer Interaction* 14.4 (2021), pp. 272–344.

[174] Silvia Milano et al. “Epistemic fragmentation poses a threat to the governance of online targeting”. In: *Nature Machine Intelligence* 3.6 (2021), pp. 466–472.

[175] Youngme Moon. “Personalization and personality: some effects of customizing message style based on consumer personality”. In: *Journal of Consumer Psychology* 12.4 (2002), pp. 313–325.

[176] *More than 200 Researchers Sign Letter Supporting Knight Institute’s Proposal to Allow Independent Research of Facebook’s Platform.* https://knightcolumbia.org/news/more-200-researchers-sign-letter-supporting-knight-institutes-proposal-allow-independent/.

[177] Kevin Munger. “All the news that’s fit to click: the economics of clickbait media”. In: *Political Communication* 37.3 (2020), pp. 376–397.

[178] Kevin Munger et al. “The (null) effects of clickbait headlines on polarization, trust, and learning”. In: *Public Opinion Quarterly* 84.1 (2020), pp. 49–73.

[179] Arvind Narayanan. *Understanding Social Media Recommendation Algorithms*. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms, Accessed: Aug 23, 2023. 2023.

[180] Milad Nasr and Michael Carl Tschantz. “Bidding strategies with gender nondiscrimination constraints for online ad auctions”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 337–347.

[181] Maxim Naumov et al. “Deep learning recommendation model for personalization and recommendation systems”. In: *CoRR* abs/1906.00091 (2019). URL: https://arxiv.org/abs/1906.00091.

[182] *Neilson DMA® Regions*. https://www.nielsen.com/intl-campaigns/us/dma-maps.html.

[183] Terry Nelms et al. “Towards measuring and mitigating social engineering software download attacks”. In: *25th USENIX Security Symposium (USENIX Security 16)*. 2016, pp. 773–789.

[184] Nico Neumann, Catherine E. Tucker, and Timothy Whitfield. “How effective is black-box digital consumer profiling and audience delivery?: evidence from field studies”. In: *Social Science Research Network Working Paper Series* (2018).

[185] *New Marketing API Requirements for all Advertising Campaigns*. https://developers.facebook.com/blog/post/2019/08/15/new-marketing-api-requirements-for-all-advertising-campaigns/.

[186] *New Targeting Tools Make Pinterest Ads Even More Effective*. https://business.pinterest.com/en/blog/new-targeting-tools-make-pinterest-ads-even-more-effective.

[187] Rob Nixon. *Slow Violence and the Environmentalism of the Poor*. Harvard University Press, 2011.

[188] Stacy Jo Dixon. *Number of daily active Facebook users worldwide as of Q2 2023.* https://www.statista.com/statistics/346167/facebook-global-dau/, Accessed: August 31, 2023.

[189] NYU. *NYU Cybersecurity for Democracy, Social Media Monitor Extension.* https://github.com/CybersecurityForDemocracy/social-media-collector.

[190] Office of the High Commissioner for Human Rights. *Moderating online content: fighting harm or silencing dissent?* URL: https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent.

[191] Eli Pariser. *The Filter Bubble: What the Internet Is Hiding from You.* London, UK: Penguin Press, 2011.

[192] Javier Parra-Arnau, Jagdish Prasad Achara, and Claude Castelluccia. "MyAdChoices: Bringing Transparency and Control to Online Advertising". In: *ACM Transactions on the Web (TWEB)* 11 (2017).

[193] Kenny Peng et al. *Reconciling the accuracy-diversity trade-off in recommendations.* 2023. arXiv: 2307.15142 [cs.IR].

[194] Supavich Fone Pengnate. "Measuring emotional arousal in clickbait: eye-tracking approach". In: *Americas Conference on Information Systems.* 2016.

[195] Supavich Fone Pengnate, Jeffrey Chen, and Alex Young. "Effects of clickbait headlines on user responses: an empirical investigation". In: *Journal of International Technology and Information Management* 30.3 (2021), pp. 1–18.

[196] *Pinterest: Audience targeting.* https://help.pinterest.com/en/business/article/audience-targeting.

[197] Heidi D Posavac, Steven S Posavac, and Emil J Posavac. "Exposure to media images of female attractiveness and concern with body weight among young women". In: *Sex Roles* 38.3 (1998), pp. 187–201.

[198] W James Potter. "Cultivation theory and research: a conceptual critique". In: *Human Communication Research* 19.4 (1993), pp. 564–601.

[199] Prolific. *Audience checking tool.* URL: https://app.prolific.co/audience-checker.

[200] Vaibhav Rastogi et al. "Are these ads safe: detecting hidden attacks through the mobile app-web interfaces." In: *Proc. of Network and Distributed System Security (NDSS) Symposium.* 2016.

[201] Elissa M Redmiles, Neha Chachra, and Brian Waismeyer. "Examining the demand for spam: who clicks?" In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* 2018, pp. 1–10.

[202] *Report for Receipts and Disbursements, ActBlue Mid-Year Report, Filing FEC-1344765.* https://docquery.fec.gov/cgi-bin/forms/C00401224/1344765/.

[203] Filipe N. Ribeiro et al. "On microtargeting socially divisive ads: a case study of russia-linked ad campaigns on facebook". In: *Conference on Fairness, Accountability and Transparency.* ACM. 2019, pp. 140–149.

[204] Manoel Horta Ribeiro, Veniamin Veselovsky, and Robert West. "The amplification paradox in recommender systems". In: *Proceedings of the International AAAI Conference on Web and Social Media.* Vol. 17. 2023, pp. 1138–1142.

[205] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning". In: *arXiv preprint arXiv:1606.05386* (2016).

[206] Ronald E. Robertson et al. "Auditing partisan audience bias within google search". In: *Proc. of CHI.* 2018.

[207] Diego Saez-Trumper et al. "Beyond cpm and cpc: determining the value of users on osns". In: *Proc. of Conference on Online Social Networks (COSN).* 2014.

[208] Christian Sandvig et al. "Auditing algorithms: research methods for detecting discrimination on internet platforms". In: 2014.

[209] Piotr Sapiezynski. *Oral Testimony before the European Parliament's Committee on Internal Market and Consumer Protection (IMCO).* URL: https://sapiezynski.com/papers/sapiezynski2022imco.pdf. July 2022.

[210] Piotr Sapiezynski et al. “Quantifying the impact of user attention on fair group representation in ranked lists”. In: *Companion proceedings of the 2019 World Wide Web conference*. 2019.

[211] J Ben Schafer, Joseph Konstan, and John Riedl. “Recommender systems in e-commerce”. In: *Proceedings of the 1st ACM conference on Electronic commerce*. 1999, pp. 158–166.

[212] Kate Scott. “You won’t believe what’s in this paper! clickbait, relevance and the curiosity gap”. In: *Journal of pragmatics* 175 (2021), pp. 53–66.

[213] D Sculley et al. “Detecting adversarial advertisements in the wild”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 274–282.

[214] YouTube NBC News. https://www.youtube.com/watch?v=n2H8wx1aBiQ, Accessed: August 29, 2023. 2015.

[215] Eugene F Shaw. “Agenda-setting and mass communication theory”. In: *Gazette (Leiden, Netherlands)* 25.2 (1979), pp. 96–105.

[216] *Sheryl Sandberg delivered a passionate, defiant defense of Facebook’s business*. https://www.cnbc.com/2018/04/26/facebooks-sheryl-sandbergs-brilliant-defense-of-the-ad-business.html.

[217] Ashudeep Singh et al. “Building healthy recommendation sequences for everyone: a safe reinforcement learning approach”. In: (2020).

[218] Stacy L Smith and Edward Donnerstein. “Harmful effects of exposure to media violence: learning of aggression, emotional desensitization, and fear”. In: *Human Aggression*. Elsevier, 1998, pp. 167–202.

[219] *Social Media Fact Sheet*. https://www.pewresearch.org/internet/fact-sheet/social-media/.

[220] *Sorrell v. IMS Health Inc.* https://supreme.justia.com/cases/federal/us/564/552/.

[221] Vera Sosnovik and Oana Goga. "Understanding the complexity of detecting political ads". In: *Proceedings of the Web Conference 2021*. 2021, pp. 2002–2013.

[222] Till Speicher et al. "On the potential for discrimination in online targeted advertising". In: *Conference on Fairness, Accountability and Transparency*. 2018.

[223] Till Speicher et al. "Potential for discrimination in online targeted advertising". In: *Conference on Fairness, Accountability and Transparency*. PMLR. 2018, pp. 5–19.

[224] Isaac Stanley-Becker. *Facebook's ad tools subsidize partisanship, research shows.* https://www.washingtonpost.com/technology/2019/12/10/facebooks-ad-delivery-system-drives-partisanship-even-if-campaigns-dont-want-it-new-research-shows/. 2019.

[225] Latanya Sweeney. "Discrimination in online ad delivery". In: *Communications of the ACM* 56.5 (2013), pp. 44–54.

[226] Jenny Tang, Eleanor Birrell, and Ada Lerner. "How well do my results generalize now? the external validity of online privacy and security surveys". In: *arXiv preprint arXiv:2202.14036* (2022).

[227] Liang Tang et al. "Automatic ad format selection via contextual bandits". In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2013, pp. 1587–1594.

[228] David R Thomas. "A general inductive approach for qualitative data analysis". In: (2003).

[229] *TikTok Business Help Center: About Custom Audiences.* https://ads.tiktok.com/help/article/custom-audiences.

[230] Aditya Srinivas Timmaraju et al. "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*. 2023, pp. 4937–4947.

[231] *Title VII of the Civil Rights Act of 1964.* https://www.law.cornell.edu/wex/title_vii.

[232] *Trump campaign drops more than $1 million on Facebook ads in counter impeachment push*. https://abcnews.go.com/Politics/trump-campaign-drops-million-facebook-ads-counter-impeachment/story?id=65962584.

[233] Zeynep Tufekci. "Youtube, the great radicalizer". In: *The New York Times* 10.3 (2018), p. 2018.

[234] *Twitter Business: Intro to Custom Audiences*. https://business.twitter.com/en/help/campaign-setup/campaign-targeting/custom-audiences.html.

[235] *United States v. Internet Research Agency*. https://www.justice.gov/file/1035477/download.

[236] *United States vs. Meta Platforms Settlement, Case 1:22-CV-05187*. https://www.justice.gov/opa/press-release/file/1514031/download.

[237] Upturn. *Leveling the Platform: Real Transparency for Paid Messages on Facebook*. https://www.upturn.org/reports/2018/facebook-ads/. 2018.

[238] *Upturn Amicus Brief in Onuoha v. Facebook*. https://www.courtlistener.com/recap/gov.uscourts.cand.304918/gov.uscourts.cand.304918.76.1.pdf.

[239] *Van Buren v. United States, 141 S. Ct. 1648, 593 U.S., 210 L. Ed. 2d 26*. 2021.

[240] Patricia Van den Berg et al. "Is dieting advice from magazines helpful or harmful? Five-year associations with weight-control behaviors and psychological outcomes in adolescents". In: *Pediatrics* 119.1 (2007), e30–e37.

[241] Giridhari Venkatadri and Alan Mislove. "On the potential for discrimination via composition". In: *Proceedings of the ACM Internet Measurement Conference*. 2020, pp. 333–344.

[242] Giridhari Venkatadri, Alan Mislove, and Krishna P. Gummadi. "Treads: transparency-enhancing ads". In: *Proc. of HotNets*. 2018.

[243] Giridhari Venkatadri et al. "Privacy risks with Facebook's pii-based targeting: auditing a data broker's advertising interface". In: 2018.

[244] Giridhari Venkatadri et al. "Auditing offline data brokers via facebook's advertising platform". In: *Proc. of The Web Conference*. 2019.

[245] Giridhari Venkatadri et al. “Investigating sources of pii used in Facebook’s targeted advertising”. In: *Proc. of PETS*. 2019.

[246] Neil Vidmar and Milton Rokeach. “Archie Bunker’s bigotry: a study in selective perception and exposure”. In: *Journal of Communication* 24.1 (1974), pp. 36–47.

[247] Melanie Wakefield et al. “Role of the media in influencing trajectories of youth smoking”. In: *Addiction* 98 (2003), pp. 79–103.

[248] Tian Wang, Yuri M Brovman, and Sriganesh Madhvanath. “Personalized embedding-based e-commerce recommendations at ebay”. In: *arXiv preprint arXiv:2102.06156* (2021).

[249] Ellen L. Weintraub. *Don’t abolish political ads on social media. Stop microtargeting.* https://www.washingtonpost.com/opinions/2019/11/01/dont-abolish-political-ads-social-media-stop-microtargeting/. Nov. 2019.

[250] *What it means when your ad is pending review*. https://www.facebook.com/business/help/204798856225114.

[251] *Wikipedia: Gini coefficient*. https://en.wikipedia.org/wiki/Gini_coefficient.

[252] *Wikipedia: Skewness*. https://en.wikipedia.org/wiki/Skewness.

[253] Craig E. Wills and Can Tatar. “Understanding what they do with what they know”. In: 2012.

[254] Abby K Wood and Ann M Ravel. “Fool me once: regulating fake news and other online advertising”. In: *Southern California Law Review* 91 (2017), p. 1223.

[255] Yuxi Wu et al. “The slow violence of surveillance capitalism: how online behavioral advertising harms people”. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT ’23. Chicago, IL, USA: Association for Computing Machinery, 2023, pp. 1826–1837. ISBN: 9798400701924. DOI: 10.1145/3593013.3594119. URL: https://doi.org/10.1145/3593013.3594119.

[256] Guang-Xin Xie and David M Boush. “How susceptible are consumers to deceptive advertising claims? a retrospective look at the experimental research literature”. In: *The Marketing Review* 11.3 (2011), pp. 293–314.

[257] Guang-Xin Xie, Robert Madrigal, and David M Boush. “Disentangling the effects of perceived deception and anticipated harm on consumer responses to deceptive advertising”. In: *Journal of Business Ethics* 129.2 (2015), pp. 281–293.

[258] Apostolis Zarras et al. “The dark alleys of madison avenue: understanding malicious advertisements”. In: *Proceedings of the 2014 conference on internet measurement conference*. 2014, pp. 373–380.

[259] Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. “Bad news: clickbait and deceptive ads on news and misinformation websites”. In: *Workshop on Technology and Consumer Protection*. 2020.

[260] Eric Zeng, Tadayoshi Kohno, and Franziska Roesner. “What makes a “bad” ad? user perceptions of problematic online advertising”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–24.

[261] Eric Zeng et al. “Polls, clickbait, and commemorative $2 bills: problematic political advertising on news and media websites around the 2020 US elections”. In: *Proceedings of the 21st ACM Internet Measurement Conference*. 2021, pp. 507–525.

[262] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. “The impact of YouTube recommendation system on video views”. In: *Proc. of the Internet Measurement Conference*. 2010, pp. 404–410.

[263] Zhihui Zhou, Lilin Zhang, and Ning Yang. “Contrastive collaborative filtering for cold-start item recommendation”. In: *arXiv preprint arXiv:2302.02151* (2023).

[264] Shoshana Zuboff. *The Age of Surveillance Capitalism*. PublicAffairs, 2019.