What Twitter Knows: Characterizing Ad Targeting Practices, User Perceptions, and Ad Explanations Through Users' Own Twitter Data

Miranda Wei[△], Madison Stamos, Sophie Veys, Nathan Reitinger[★], Justin Goodman[★], Margot Herman, Dorota Filipczuk[†], Ben Weinshel, Michelle L. Mazurek[★], Blase Ur
University of Chicago, [★]University of Maryland, [†]University of Southampton, [△]University of Chicago and University of Washington

Abstract

Although targeted advertising has drawn significant attention from privacy researchers, many critical empirical questions remain. In particular, only a few of the dozens of targeting mechanisms used by major advertising platforms are well understood, and studies examining users' perceptions of ad targeting often rely on hypothetical situations. Further, it is unclear how well existing transparency mechanisms, from data-access rights to ad explanations, actually serve the users they are intended for. To develop a deeper understanding of the current targeting advertising ecosystem, this paper engages 231 participants' own Twitter data, containing ads they were shown and the associated targeting criteria, for measurement and user study. We find many targeting mechanisms ignored by prior work — including advertiser-uploaded lists of specific users, lookalike audiences, and retargeting campaigns — are widely used on Twitter. Crucially, participants found these understudied practices among the most privacy invasive. Participants also found ad explanations designed for this study more useful, more comprehensible, and overall more preferable than Twitter's current ad explanations. Our findings underscore the benefits of data access, characterize unstudied facets of targeted advertising, and identify potential directions for improving transparency in targeted advertising.

1 Introduction

Social media companies derive a significant fraction of their revenue from advertising. This advertising is typically highly targeted, drawing on data the company has collected about the user, either directly or indirectly. Prior work suggests that while users may find well-targeted ads useful, they also find them "creepy" [40, 42, 58, 61, 70]. Further, users sometimes find targeted ads potentially embarrassing [3], and they may (justifiably) fear discrimination [4, 15, 21, 47, 57, 59, 60, 73]. In addition, there are questions about the accuracy of categorizations assigned to users [7, 12, 21, 71, 72]. Above all, users currently have a limited understanding of the scope and mechanics of targeted advertising [17, 22, 48, 50, 54, 70, 77].

Many researchers have studied targeted advertising, largely focusing on coarse demographic or interest-based targeting. However, advertising platforms like Twitter [63] and Google [27] offer dozens of targeting mechanisms that are far more precise and leverage data provided by users (e.g., Twitter accounts followed), data inferred by the platform (e.g., potential future purchases), and data provided by advertisers (e.g., PII-indexed lists of current customers). Further, because the detailed contents and provenance of information in users' advertising profiles are rarely available, prior work focuses heavily on abstract opinions about hypothetical scenarios.

We leverage data subjects' right of access to data collected about them (recently strengthened by laws like GDPR and CCPA) to take a more comprehensive and ecologically valid look at targeted advertising. Upon request, Twitter will provide a user with highly granular data about their account, including all ads displayed to the user in the last 90 days alongside the criteria advertisers used to target those ads, all interests associated with that account, and all advertisers who targeted ads to that account.

In this work, we ask: What are the discrete targeting mechanisms offered to advertisers on Twitter, and how are they used to target Twitter users? What do Twitter users think of these practices and existing transparency mechanisms? A total of 231 Twitter users downloaded their advertising-related data from Twitter, shared it with us, and completed an online user study incorporating this data. Through this method, we analyzed Twitter's targeting ecosystem, measured participants' reactions to different types of ad targeting, and ran a survey-based experiment on potential ad explanations.

We make three main contributions. First, we used our 231 participants' files to characterize the current Twitter ad-targeting ecosystem. Participants received ads targeted based on 30 different *targeting types*, or classes of attributes through which advertisers can select an ad's recipients. These types ranged from those commonly discussed in the literature (e.g., INTERESTS, AGE, GENDER) to others that have received far less attention (e.g., AUDIENCE LOOKALIKES, advertiser-uploaded LISTS of specific users, and RETARGETING CAMPAIGNS). Some partic-

ipants' files contained over 4,000 distinct keywords, 1,000 follower lookalikes, and 200 behaviors. Participants' files also revealed they had been targeted ads in ways that might be seen as violating Twitter's policies restricting use of sensitive attributes. Participants were targeted using advertiser-provided LISTS of users with advertiser-provided names containing "DLX_Nissan_AfricanAmericans," "Christian Audience to Exclude," "Rising Hispanics | Email Openers," and more. They were targeted using KEYWORDS like "#transgender" and "mexican american," as well as conversation topics like the names of UK political parties. These findings underscore how data access rights facilitate transparency about targeting, as well as the value of such transparency.

Second, we investigated participants' perceptions of the fairness, accuracy, and desirability of 16 commonly observed targeting types. Different from past work using hypothetical situations, we asked participants about specific examples that had actually been used to target ads to them in the past 90 days. Whereas much of the literature highlights users' negative perceptions of interest-based targeting [42, 61], we found that over two-thirds of participants agreed targeting based on INTEREST was fair, the third most of the 16 types. In contrast, fewer than half of participants agreed that it was fair to target using understudied types like FOLLOWER LOOKALIKE TARGETING, TAILORED AUDIENCE LISTS, EVENTS, and BEHAVIORS. Many targeting types ignored by prior work were the ones viewed least favorably by participants, emphasizing the importance of expanding the literature's treatment of ad-targeting mechanisms.

Third, we probe a fuller design space of specificity, readability, and comprehensiveness for *ad explanations*. Although ad explanations are often touted as a key part of privacy transparency [24], we find that existing ad explanations are incomplete and participants desire greater detail about how ads were targeted to them. Compared to Twitter's current explanation, participants rated explanations we created to be significantly more useful, helpful in understanding targeting, and similar to what they wanted in future explanations.

Our approach provides a far richer understanding of the Twitter ad ecosystem, users' perceptions of ad targeting, and ad explanation design than was previously available. Our results emphasize the benefits of advertising transparency in surfacing potential harms associated with increasingly accurate and complex inferences. Our findings also underscore the need for a more ethical approach to ad targeting that can maintain the trust of users whose data is collected and used.

2 Related Work

We review prior work on techniques for targeted advertising, associated transparency mechanisms, and user perceptions.

2.1 Targeted Advertising Techniques

Web tracking dates back to 1996 [38]. The online ad ecosystem has only become more sophisticated and complex since. Companies like Google, Facebook, Bluekai, and many others track users' browsing activity across the Internet, creating profiles for the purpose of sending users targeted advertising. Commercial web pages contain an increasing number of trackers [52], and much more data is being aggregated about users [13]. Many studies have examined tools to block tracking and targeted ads, finding that tracking companies can still observe some of a user's online activities [2, 10, 11, 19, 30].

Social media platforms have rich data for developing extensive user profiles [7, 12, 57], augmenting website visits with user-provided personal information and interactions with platform content [7]. This data has included sensitive categories like 'ethnic affinity' [8] and wealth. Even seemingly neutral attributes can be used to target marginalized groups [57].

To date, studies about user perceptions of ad-targeting mechanisms have primarily focused on profiles of users' demographics and inferred interests (e.g., yoga, travel) regardless of whether the studies were conducted using users' own ad-interest profiles [12, 20, 50] or hypothetical scenarios [17, 36]. Furthermore, most studies about advertising on social media have focused on Facebook [7, 25, 57, 71]. While some recent papers have begun to examine a few of the dozens of other targeting mechanisms available [7, 72], our study leverages data access requests to characterize the broad set of targeting types in the Twitter ecosystem much more comprehensively than prior work in terms of both the mechanisms considered and the depth of a given user's data examined.

Newer techniques for targeting ads go beyond collecting user data in several ways that may be less familiar to both users and researchers. For example, since 2013, Facebook [23] and Twitter [9] have offered "custom" or "tailored" audience targeting, which combine online user data with offline data. Advertisers upload users' personally identifiable information (PII), such as their phone numbers and email addresses gathered from previous transactions or interactions, in order to link to users' Facebook profiles. This offline data can also include data supplied by data brokers [72], often pitched to advertisers as "partner audiences" [32], or even PII from voter and criminal records [7]. These features can be exploited by advertisers to target ads to a single person [25], or evade restrictions about showing ads to people in sensitive groups [57].

Another newer form of targeting is lookalike-audience targeting, which relies on inferences about users relative to other users. For example, on Facebook, advertisers can reach new users with similar profiles as their existing audience [39]. This feature can be exploited, as a biased input group will lead to an output group that contains similar biases [57]. Services are increasingly implementing lookalike targeting [56]. To our knowledge, we are the first to study user perceptions of these lesser-known forms of targeting with real-world data.

2.2 Transparency Mechanisms

Ad and analytics companies increasingly offer transparency tools [16, 29]. These include ad preference managers [12], which allow users to see the interest profiles that platforms have created for them, and *ad explanations*, or descriptions of why a particular advertiser displayed a particular ad to a user [7]. Nevertheless, a disparity remains between information available to advertisers and information visible to users [50, 58]. Although researchers have documented advertisers' use of a multitude of attributes, including sensitive ones, they rarely appear in user-facing content [7, 15, 50, 74, 75]. Facebook's ad preferences are vague and incomplete [7], notably leaving out information from data brokers [72].

To shed light on the black box of advertising, researchers have developed "reverse engineering" tools that can extract some information about targeting mechanisms, associated explanations, and inferences that have been made. Techniques include measuring the ads users see [6,7,10,11,15,34,35,75], purchasing ads in controlled experiments [4,71,72], and scraping companies' ad-creation interface [25,57,71,72,74], ad-interest profiles [7,12,15,16,60,75], and ad explanations [6,7]. Unfortunately, these excellent tools are limited by the difficulty of scaling them (as they require making many requests per user) and by companies continually making changes to their interfaces, perhaps in part to thwart such tools [43].

2.3 Perceptions of Targeting & Transparency

Users do not understand advertising data collection and targeting processes [7,17,20,45,54]. They instead rely on imprecise mental models [58] or folk models [22,77]. While some users like more relevant content [40] and understand that ads support free content on the web [42], many others believe tracking browser activity is invasive [42, 53]. Users are concerned about discrimination [47] or bias [21], inaccurate inferences, and companies inferring sensitive attributes such as health or financial status [50, 70]. Studies have shown that when users learn about mechanisms of targeted advertising, their feelings towards personalization become more negative [53, 58, 61].

To an increasing extent, studies have looked into the design and wording of transparency tools [5, 37, 74]. Unfortunately, these tools are meant to provide clarity but can be confusing due to misleading icons [36] or overly complicated language [37, 54]. Improving the design of transparency tools is important because vague ad explanations decrease users' trust in personalized advertising, while transparency increases participants' likelihood to use that service [20] and to appreciate personalization [54, 70]. Users want to know the specific reasons for why they saw an ad [17] and want more control over their information by being able to edit their interest profiles [31, 41]. Users continually express concern about their privacy [18, 28] but cannot make informed decisions if information about how their data is used is not transparent [58].

Ad explanations are a particularly widespread form of transparency [7, 17]. Sadly, prior work has found current explanations incomplete [7,71,72] and companion ad-interest profiles to be both incomplete [15] and inaccurate [12,16]. While studies have examined existing ad explanations [7, 20, 71, 72] or engaged in speculative design of new explanations [20], surprisingly little work has sought to quantitatively test improved explanations. We build on this work by quantitatively comparing social media platforms' current ad explanations with new explanations we designed based on prior user research [17,20]. Emphasizing ecological validity, we test these explanations using ads that had actually been shown to participants while explaining the true reasons those ads had been targeted to them, leveraging the participant's own Twitter data.

3 Method

To examine Twitter ad targeting data, we designed an online survey-based study with two parts. First, participants followed our instructions to request their data from Twitter. Upon receipt of this data a few days later, they uploaded the advertising-relevant subset of this data and completed a survey that instantly incorporated this data across two sections.

Section 1 of the survey elicited participants' reactions to different targeting types, such as follower lookalike targeting and interest targeting. We selected 16 commonly observed targeting types, many of which have not previously been explored in the literature. In Section 2, we conducted a within-subjects experiment measuring participants' reactions to six potential ad explanations, including three novel explanations we created by building on prior work [17, 20], as well as approximations of Twitter and Facebook's current ad explanations. We also asked participants about their general Twitter usage. We concluded with demographic questions. Our survey was iteratively developed through cognitive interviews with people familiar with privacy research, as well as pilot testing with people who were not. Below, we detail our method.

3.1 Study Recruitment

We recruited 447 participants from Prolific to request their Twitter data, paying $0.86 for this step. The median completion time was 7.3 minutes. We required participants be at least 18 years old, live in the US or UK, and have a 95%+ approval rating on Prolific. Additionally, participants had to use Twitter at least monthly and be willing to upload their Twitter ad data to our servers. During this step, we requested they paste into our interface the ad interest categories Twitter reported for them in their settings page. If a participant reported 10 or fewer interests (another indicator of infrequent usage), we did not invite them to the survey.

To give participants time to receive their data from Twitter, we waited several days before inviting them back. A total of

254 participants completed the survey. The median completion time for the 231 valid participants (see Section 4.1) was 31.5 minutes, and compensation was $7.00.

To protect participants' privacy, we automatically extracted and uploaded only the three Twitter files related to advertising: `ad-impressions.js`, `personalization.js`, and `twitter_advertiser_list.pdf`. The JavaScript file `ad-impressions.js` contained data associated with ads seen on Twitter in the preceding 90 days, including the advertiser's name and Twitter handle, targeting types and values, and a timestamp. An example of this JSON data is presented in our online Appendix A [1]. The file `twitter_advertiser_list.pdf` contained advertisers who included the participant in a tailored audience list, as well as lookalike audiences in which Twitter placed the participant.

3.2 Survey Section 1: Targeting Types

Our goal for the first section of the survey was to comparatively evaluate user awareness, perceptions, and reactions to the targeting types advertisers frequently use to target ads on Twitter. We wanted to include as many targeting types as possible, while ensuring that a given participant would be likely to have seen at least one ad targeted using that type. If we had included all 30 types, we would have only been able to show a few participants an ad relying on the more obscure types, and would likely not have had a sufficient number of participants to meaningfully carry out our statistical analyses. In our pilot data, only 16 targeting types appeared in the data of more than half of our pilot participants; therefore, we opted to use these 16 in the survey. The 16 targeting types were as follows: FOLLOWER LOOKALIKE; LOCATION; TAILORED AUDIENCE (LIST); KEYWORD; AGE; CONVERSATION TOPIC; INTEREST; TAILORED AUDIENCE (WEB); PLATFORM; LANGUAGE; BEHAVIOR; GENDER; MOVIES AND TV SHOWS; EVENT; RETARGETING CAMPAIGN ENGAGER; and MOBILE AUDIENCE. We refer to a specific attribute of a type as an *instance* of that type. For example, LANGUAGE targeting has instances like English and French, and EVENT targeting has instances including "2019 Women's World Cup" and "Back to School 2019." These targeting types are described in detail in Section 4.3; Twitter's definitions are given in our online Appendix B [1]. Using a mixed between- and within-subjects design, we showed each participant four randomly selected targeting types, chosen from however many of the 16 types were in that user's ad impressions file. Prior work has covered only a fraction of these 16 targeting types. Furthermore, asking questions about instances from participants' own Twitter data increased the ecological validity of our study compared to the hypothetical scenarios used in prior work.

For each targeting type, we repeated a battery of questions. First, we asked participants to define the targeting type in their own words. Next, we gave a definition of the term adapted from Twitter for Business help pages [63]. We then showed participants one specific instance of the targeting type, drawn from their Twitter data (e.g., for KEYWORD, "According to your Twitter data, you have searched for or Tweeted about cats"). Finally, we showed participants the five most and five least frequent instances of that targeting type in their Twitter data (if there were fewer than 10 instances, we showed all available), as well as an estimate of how many ads they had seen in the last three months that used that targeting type.

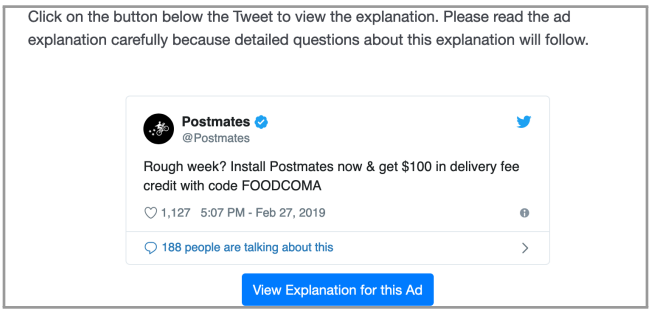


Figure 1: Example ad shown in Section 2 of the survey. Participants always saw the ad before the corresponding explanation.

At this point, the participant had seen a definition of the targeting type as well as several examples to aid their understanding. We then asked questions about participants' comfort with, perception of the fairness of, perceptions of the accuracy of, and desire to be targeted by the type. For these questions, we asked participants to rate their agreement on a 5-point Likert scale from strongly agree to strongly disagree. Hereafter, we say participants "agreed" with a statement as shorthand indicating participants who chose either "agree" or "strongly agree." Similarly, we use "disagreed" as shorthand for choosing "disagree" or "strongly disagree." We also asked participants to explain their choices in free-text responses to confirm that participants were understanding our constructs as intended. The text of all questions is shown in Appendix E [1].

3.3 Survey Section 2: Ad Explanations

Our goal for the second section of the survey was to characterize user reactions to the ad explanations companies like Twitter and Facebook currently provide on social media platforms, as well as to explore whether ideas proposed in past work (but not quantitatively tested on a large scale) could lead to improved explanations. To that end, we used participants' Twitter data to craft personalized ad explanations for ads that were actually displayed to them on Twitter within the last 90 days. We tested six different ad explanations.

Rather than trying to pinpoint the best design or content through extensive A/B testing, we instead constructed our explanations as initial design probes of prospective ad explanations that are more detailed than those currently used by major social media platforms. The explanations differed in several ways, allowing us to explore the design space. Our

within-subjects design invited comparison among the explanations, which helped participants to evaluate the them, as well as answer the final question of this section: “Please describe your ideal explanation for ads on Twitter."

To study reactions to widely deployed ad explanations, our first two explanations were modeled on those Twitter and Facebook currently use. They retained the same information and wording, but were recreated in our visual theme for consistency and to avoid bias from participants knowing their origin. The first was based on **Twitter**'s current ad explanation (Fig. 2a), which features most commonly, but not always, two of the many possible ad targeting types: INTEREST and LOCATION (most frequently at the level of country). Notably, ads themselves can be targeted to more granular locations and using many more targeting types; Twitter's current explanation does not present these facets to users. We also adapted one of **Facebook**'s current ad explanations (Fig. 2b), which uses a timeline to explain TAILORED AUDIENCE targeting and incorporates AGE and LOCATION. These explanations represent two major platforms' current practices.

Because current ad explanations are vague and incomplete [7, 72], we wanted to explore user reactions to potential ad explanations that are more comprehensive and also integrate design suggestions from prior work [17, 20]. We thus created two novel explanations, **Detailed Visual** (Fig. 2c) and **Detailed Text** (Fig. 2d), that showed a more comprehensive view of all the targeting types used, including lesser-known, yet commonly used, targeting types like FOLLOWER LOOKALIKE, MOBILE AUDIENCE, EVENT and TAILORED AUDIENCE. The distinction between our two conditions let us explore the communication medium. While we hypothesized that Detailed Visual would perform better than Detailed Text, we wanted to probe the trade-off between the comprehensiveness and comprehensibility of text-based explanations.

While ad explanations should be informative and intelligible, they should also nudge users to think about their choices regarding personalized advertising. We designed our third novel ad explanation, “**Creepy**" (Fig. 2e), to more strongly nudge participants toward privacy by including information likely to elicit privacy concerns. This explanation augmented our broader list of targeting types with information the participant leaks to advertisers, such as their device, browser, and IP address. This explanation also used stronger declarative language, such as “you are" instead of “you may."

Finally, we designed a generic **Control** explanation (Fig. 2f) that provided no targeting information. This explanation was designed to be vague and meaningless. Following other work [29, 76], Control provides a point of comparison.

Our ad explanations are the result of several iterations of design. After each iteration, we discussed whether the designs met our goal of creating a spectrum of possibilities for specificity, readability, and comprehensiveness. We then redesigned the explanations until we felt that they were satisfactory based on both pilot testing and group discussion.

Participants were shown ad explanations in randomized order. Each explanation was preceded by an ad from their data and customized with that ad's targeting criteria. We created a list of all ads a participant had been shown in the last 90 days and sorted this list in descending order of the number of targeting types used. To filter for highly targeted ads, we selected six ads from the beginning of this list. Participants who had fewer than six ads in their Twitter data saw explanations for all of them. After each explanation, we asked questions about whether the ad explanation was useful, increased their trust in advertisers, and more.

The six ad explanations collectively represent a spectrum of possible ad explanations in terms of specificity: Control represents a lower bound, Creepy represents an upper bound, and the others fall in between.

3.4 Analysis Method and Metrics

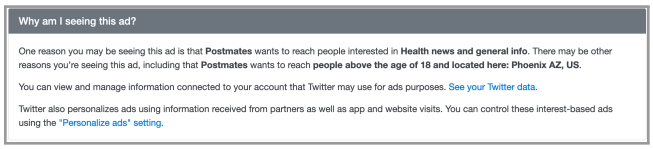
We performed quantitative and qualitative analyses of survey data. We provide descriptive statistics about Twitter data files.

Because each participant saw only up to four of the 16 targeting types in survey Section 1, we compared targeting types using mixed-effects logistic regression models. These are appropriate for sparse, within-subjects, ordinal data [49, 62]. Each model had one Likert question as the outcome and the targeting type and participant (random effect) as input variables. We used INTEREST targeting as our baseline because it is the most widely studied targeting type. INTEREST targeting is also commonly mentioned in companies' advertising disclosures and explanations, in contrast to most other targeting types we investigated (e.g., TAILORED AUDIENCE). Appendix I [1] contains our complete regression results.

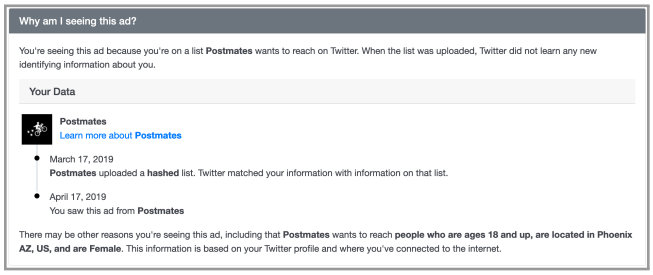
To investigate how targeting accuracy impacted participant perceptions, we also compared the accuracy of targeting type instances (self-reported by participants) to participants' responses to the other questions for that targeting type. To examine correlation between these pairs of Likert responses, we used Spearman's ρ, which is appropriate for ordinal data.

To compare a participant's Likert responses to the six different ad explanations they saw, we used Friedman's rank sum test (appropriate for ordinal within-subjects data) as an omnibus test. We then used Wilcoxon signed-rank tests to compare the other five explanations to the Twitter explanation, which we chose as our baseline because Twitter currently uses it to explain ads. We used the Holm method to correct p-values within each family of tests for multiple testing.

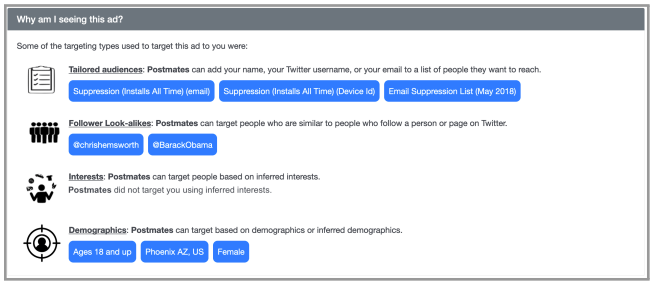
We qualitatively analyzed participants' free-response answers to five questions about targeting types and ad explanations through an open coding procedure for thematic analysis. One researcher made a codebook for each free-response question and coded participant responses. A second coder independently coded those responses using the codebook made by the first. The pair of coders for each question then met to discuss the codebook, verifying understandings of the codes and



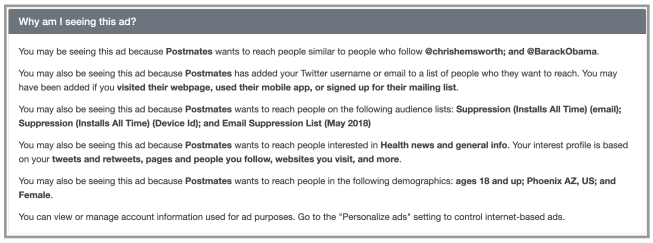
(a) **Twitter** ad explanation.



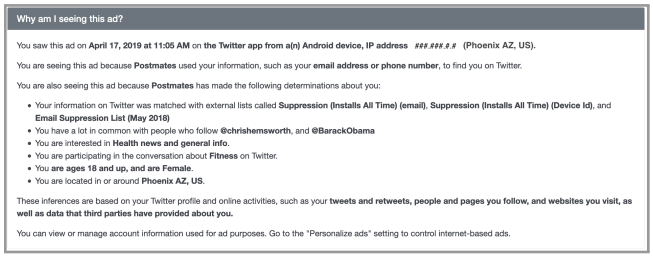
(b) **Facebook** ad explanation.



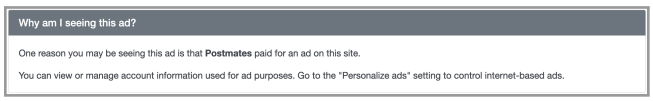
(c) **Detailed Visual** ad explanation.



(d) **Detailed Text** ad explanation.



(e) **Creepy** ad explanation.



(f) **Control** ad explanation.

Figure 2: The six ad explanations tested, using a hypothetical ad to demonstrate all facets of the explanations.

combining codes that were semantically similar. Inter-coder reliability measured with Cohen's κ ranged from 0.53 to 0.91 for these questions. Agreement > 0.4 is considered "moderate" and > 0.8 "almost perfect" [33]. To provide context, we report the fraction of participants that mentioned specific themes in these responses. However, a participant failing to mention something is not the same as disagreeing with it, so this prevalence data should not be considered generalizable. Accordingly, we do not apply hypothesis testing.

3.5 Ethics

This study was approved by our institutions' IRB. As social media data has potential for abuse, we implemented many measures to protect our participants' privacy. We did not collect any personally identifiable information from participants and only identified them using their Prolific ID numbers. Additionally, we only allowed participants to upload the three files necessary for the study from participants' Twitter data; all other data remained on the participant's computer. These three files did not contain personally identifiable information. In this paper, we have redacted potential identifiers found in targeting data by replacing numbers with #, letters with *, and dates with MM, DD, or YYYY as appropriate.

To avoid surprising participants who might be uncomfortable uploading social media data, we placed a notice in our study's recruitment text explaining that we would request such data. As some of our participants were from the UK and Prolific is located in the UK, we complied with GDPR.

3.6 Limitations

Like all user studies, ours should be interpreted in the context of its limitations. We used a convenience sample via Prolific that is not necessarily representative of the population, which lessens the generalizability of our results. However, prior work suggests that crowdsourcing for security and privacy survey results can be more representative of the US population than census-representative panels [51], and Prolific participants produce higher quality data than comparable platforms [46]. We may have experienced self-selection bias in that potential participants who are more privacy sensitive may have been unwilling to upload their Twitter data to our server. Nonetheless, we believe our participants provided a useful window into user reactions. While we did find that the average character count of free response questions decreased over the course of the survey ($\rho = -0.399$; $p < 0.01$ between question order and average character number), we were satisfied with the qualitative quality of our responses. Responses included in our analysis and results were on-topic and complete.

We were also limited by uncertainty in our interpretation of the Twitter data files at the time we ran the user study. Twitter gives users their data files without documentation defining the elements in these files. For instance, each ad in the data file contains a JSON field labeled "matchedTargetingCriteria" that contains a list of targeting types and instances. It was

initially ambiguous to us whether all instances listed had been matched to the participant, or whether this instead was a full list of targeting criteria specified by the advertiser regardless of whether each matched to the participant. The name of this field suggested the former interpretation. However, the presence of multiple instances that could be perceived as mutually exclusive (e.g., non-overlapping income brackets) and Twitter telling advertisers that some targeting types are "ORed" with each other (see online Appendix F, Figure 6 [1]) made us question our assumption. Members of the research team downloaded their own data and noticed that most "matchedTargetingCriteria" were consistent with their own characteristics. We made multiple requests for explanations of this data from Twitter, including via a GDPR request from an author who is an EU citizen (see online Appendix C [1]). We did not receive a meaningful response from Twitter for more than 4.5 months, by which point we had already run the user study with softer language in survey questions and ad explanations than we might otherwise have used. Ultimately, Twitter's final response reported that the instances shown under "matchedTargetingCriteria" indeed were all matched to the user, confirming our initial interpretation.

Because we wanted to elicit reactions to ad explanations for ads participants had actually been shown, our comparisons of ad explanations are limited by peculiarities in participants' ad impressions data. If an ad did not have a particular targeting type associated with it, then that targeting type was omitted from the explanation. The exception was Visual, which told participants whether or not each targeting type was used. Further, 38 participants' files contained data for fewer than six ads. In these cases, we showed participants explanations for all ads in their file. The targeting types and specific example instances randomly chosen for each participant had inherent variance. Some targeting types had more potential instances than others. Some instances undoubtedly seemed creepier or more invasive than others, even within the same targeting type. To account for these issues, we recruited several hundred participants and focused on comparisons among targeting types and explanations, interpreting our results accordingly. Additionally, the more detailed explanations were less readable, and participants may have been more likely to skim them. We performed a broad exploration of the design space in an effort to understand what features participants liked and disliked. There is a trade-off between readability and comprehensiveness that future work should address.

4 Results

In this section, we first characterize current ad-targeting practices by analyzing our 231 participants' Twitter data. We then report participants' reactions to targeting mechanisms as well as to six potential ad explanations from our online survey.

We observed 30 different targeting types in use, some with thousands of unique instances. Participants' perceptions of fairness, comfort, and desirability differed starkly by type, but comfort and desirability generally increased with the perceived accuracy of the targeting. Further, all three ad explanations we designed (based on the literature) outperformed explanations currently deployed on Twitter and Facebook.

4.1 Participants

We report on data from the 231 participants who uploaded their Twitter data, completed all parts of the study, and wrote on-topic answers to free-response prompts. Our participants had been on Twitter for between 1 month and 12.3 years, with an average of 6.6 years. Two-thirds of participants reported spending under an hour a day on Twitter. Among participants, 52.8% identified as female, 84.0% reported at least some college education, and 20.8% percent reported some background in computer science or IT. When asked early in the survey, participants only recognized an average of 1.6 companies (min: 0, max: 8) out of a random sample of 10 companies that had shown them ads in the past 90 days. Interestingly, more than 50 participants reported looking at their files before the survey. Although this may have biased participants regarding specific ads shown, this is unlikely given both the large number of files found in the original data download and the large size of the `ad-impressions.js` files containing per-ad data. Participants would have had to parse many blocks like the one in Appendix A [1] and particularly notice the specific ones we asked about.

4.2 Aggregate Overview of Targeting

Participants had an average of 1046.6 ads in their files (min: 1, max: 14,035); a full histogram of ad impressions is shown in Appendix H, Figure 8 [1]. Our 231 participants' data files collectively contained 240,651 ads that had been targeted with at least one targeting type. As detailed in Table 1, we observed 30 different targeting types, with 45,209 unique instances of those targeting types.

Usage of the different targeting types varied greatly, as shown in Figure 3 (left). The most commonly used types were LOCATION (99.2% of all ads) and AGE (72.3%). The least commonly used was FLEXIBLE AUDIENCE LOOKALIKES (0.2%). A single ad could be targeted using multiple instances of a given type, but LANGUAGE, AGE, and GENDER targeting always used one instance. In contrast, FOLLOWER LOOKALIKES and KEYWORDS often employed multiple instances: 6.0 and 4.9 instances on average per ad, respectively. The largest set we observed was 158 BEHAVIOR instances. Figure 3 (center) shows how often multiple instances were used to target a given ad.

For nine targeting types, we observed fewer than ten unique instances (e.g., male and female were the only two GENDER instances). In contrast, KEYWORDS (25,745), FOLLOWER LOOKALIKES (8,792), and TAILORED LISTS (2,338) had the most unique instances across participants. For many targeting types, the

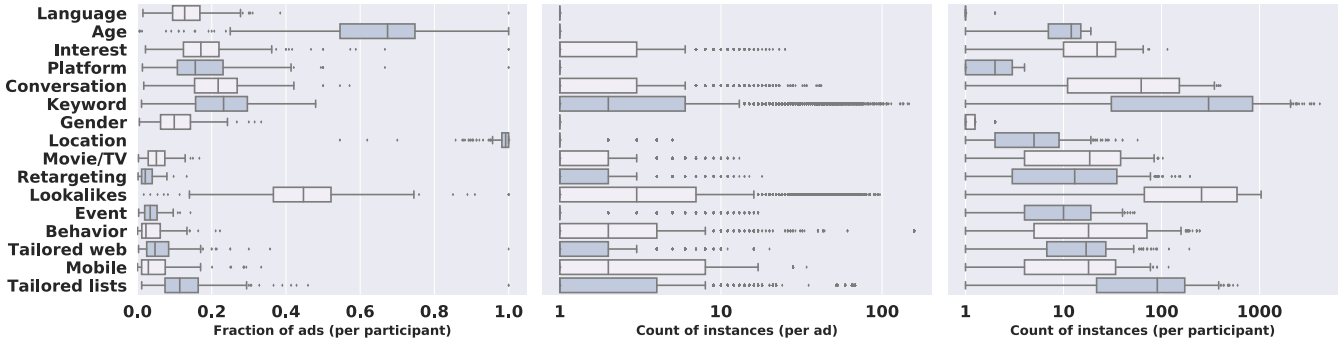


Figure 3: Summaries of our 231 participants' Twitter ad data. Left: The fraction of ads seen by each participant that included each targeting type. Center: Instances of each targeting type per ad. Right: Unique instances of each targeting type per participant.

median participant encountered dozens or hundreds of unique instances of that type, as shown in Figure 3 (right).

4.3 Detailed Usage of Targeting Types

Next, we detail how each targeting type was used to target ads to our participants. Based on the source of the data underlying each type, we grouped the targeting types into three clusters. The first two clusters — targeting types related to user demographics and targeting types related to user psychographics (behaviors and interests) — use information collected directly by Twitter. In contrast, the third cluster consists of targeting types using data provided by prospective advertisers.

4.3.1 Twitter Demographic Targeting Types

The first of our three clusters consists of demographic-based targeting. We include in this category characteristics about both a person and their device(s). Sometimes, users directly provide this information to Twitter (e.g., providing a birth date upon registration). In other cases, Twitter infers this data.

Advertisers commonly used demographics to target broad audiences. LANGUAGE was used frequently, with English being the most popularly targeted (208 participants). AGE targeting was also extremely common, yet also used coarsely (only 23 unique instances). "18 and up" was the most frequently targeted value; 83.11% of participants were targeted on this attribute. Many AGE instances overlapped (e.g., "18 and up", "18 to 24", "18 to 34," "18 to 49"). The five most frequently observed LOCATIONS were the US, UK, Los Angeles, London, and Chicago. We also observed locations as granular as ZIP codes (e.g., 44805 for Ashland, OH). Different ads for a single participant were sometimes targeted to multiple, non-overlapping locations, demonstrating that their Twitter location changed over time. GENDER targeting was much less frequently used than LANGUAGE, AGE, or LOCATION. Almost 70% of GENDER instances targeted women. The README.txt file accompanying data downloads says that Twitter infers a user's gender if they did not provide it; our analysis (and others [44]) support this assertion. We also found that this inference may change over time: 19.9% were targeted as male in some ads and female in others.

Twitter also collects data about users' devices for targeting [67]. PLATFORM was used to target ads to users of iOS (115 participants), desktop (115), and Android (98). In total 14,605 ads were targeted to iOS users, while 8,863 were targeted to Android users. The most frequently targeted DEVICE MODELS were the iPhone 8, Galaxy Note 9, iPhone 8 Plus, and iPhone 7. Participants were often associated with multiple instances (e.g., both Android Lollipop and Jelly Bean) or even targeted cross-OS (e.g., both Android Marshmallow and iOS 12.4). Twitter also offers targeting of Twitter users on a NEW DEVICE; 62.6% of the 243 instances we observed were to devices Twitter designated as 1 month old (as opposed to 2, 3, or 6 months). Advertisers also targeted by CARRIER, most commonly to T-Mobile (21 participants) and O2 (19).

4.3.2 Twitter Psychographic Targeting Types

We next discuss targeting types related to participants' psychographic attributes, which users provide via Twitter activity or which are inferred by Twitter's algorithms. Psychographic attributes relate to a user's lifestyle, behavioral, or attitudinal propensities [26]. Although "behavioral targeting" is commonly used in industry and research as an umbrella term for all forms of psychographic targeting, we describe the range of targeting based on user behaviors and attitudes as psychographic, in contrast to the specific BEHAVIOR targeting type offered by Twitter. While some participants may be aware of the inferences that could be made about them from their Twitter activity, many likely are not [73] .

Some of the most frequently used psychographic targeting types are based directly on users' Twitter activity. FOLLOWERS OF A USER ID, which targets all followers of the same Twitter account, was used 590,502 times in our data. Out of the five of the most commonly targeted values, four were related to news agencies: @WSJ, @nytimes, @TheEconomist, @washingtonpost, and @BillGates. KEYWORDS, which are selected by advertisers and approved by Twitter [65], was the most unique targeting type, with a total of 25,745 distinct

Targeting Type	Total Uses	# Unique Instances	Most Frequently Observed Instance
Source: Twitter (Demographic)			
Language*	350,121	4	English
Age*	173,917	23	18 and up
Platform*	32,351	4	iOS
Location*	31,984	566	United States
OS version	7,382	29	iOS 10.0 and above
Device model	2,747	36	iPhone 8
Carriers	1,442	11	T-Mobile UK
Gender*	1,327	2	Female
New device	236	4	1 month
WiFi-Only	108	1	WiFi-Only
Source: Twitter (Psychographic)			
Followers of a user ID	590,502	138	@nytimes
Follower lookalikes*	242,709	8,792	@netflix
Conversation topics*	128,005	2,113	Food
Keyword*	91,841	25,745	parenting
Behavior*	35,088	854	US - Household income: $30,000-$39,000
Interest*	25,284	206	Comedy
Movies and TV shows*	22,590	548	Love Island
Event*	17,778	198	2019 Women's World Cup
Retargeting campaign*	15,529	1,842	Retargeting campaign engager: ########
Retargeting engagement type	11,185	5	Retargeting engagement type: #
Retargeting user engager	2,184	218	Retargeting user engager: ##########
Retargeting lookalikes	229	66	Nielson Online - Website Visitors - Finance/In
Source: Advertiser			
Tailored audience (list)*	113,952	2,338	Lifetime Suppression [Installs] (Device Id)
Mobile audience*	21,631	478	Purchase Postmates - Local Restaurant Delivery
Tailored audience (web)*	18,016	550	Quote Finish
Tailored audience CRM lookalikes	1,179	22	Samba TV > Mediacom - Allergan - Botox Chronic
Flexible audience	382	12	iOS > Recently Installed (14days), No Checkout
Mobile lookalikes	141	23	Install New York Times Crossword IOS All
Flexible audience lookalike	7	2	All WBGs Android Purchase Events
Source: Unknown (as labeled by Twitter)			
Unknown	927	179	Unknown: ####

Table 1: Targeting types observed in our 231 participants' Twitter data. We report how many of the 240,651 ads were targeted by that type, as well as the number of unique instances of that type and the most frequently observed instance. We group targeting types by their source (advertisers or Twitter). * indicates targeting types also studied in the user survey.

instances. Keywords varied greatly in content and specificity, ranging from "technology" and "food" to "first home" (used by realtor.com) and "idiopathic thrombocytopenic purpura" (used by WEGO Health). We identified several keywords as potentially violating Twitter policies prohibiting targeting to sensitive categories "such as race, religion, politics, sex life, or health," [65, 69]. Examples include "ostomy", "Gay", and "latinas" (see Table 2 for more). Twitter infers CONVERSATION TOPIC instances based on users' Twitter activity (Tweets, clicks, etc.), allowing advertisers to target narrow populations: about a third of our unique CONVERSATION instances were in only one user's ad data. The top five topics, however, were broad: "technology," "food," "travel," "soccer," and "fashion."

Inferences made for INTERESTS targeting are one step more abstract; they are inferred from the accounts a user follows (and the content from those accounts) as well as their direct activities. The top five interests were similar to the top five conversation topics: "comedy," "tech news," "technology," "music festivals and concerts," and "soccer." Other targeted interests were more specific, such as "vintage cars" and "screenwriting."

Similarly to INTERESTS, the EVENT and MOVIES AND TV SHOWS targeting types appear to rely on both a user's direct activities and on inferences to label users as interested in offline events and entertainment. These targeting types most commonly reflected sports ("2019 Women's World Cup," 2,713 instances; "MLB Season 2019," 1,344 instances) and popular shows such as "Love Island," "Stranger Things," and "Game of Thrones."

Highly targeted psychographic targeting types are based on Twitter algorithms. FOLLOWER LOOKALIKES targeting is even more indirect: the targeted users are labeled as sharing interests or demographics with followers of a particular account, despite not actually following that account. Follower lookalikes is the second most individualized targeting type in our dataset (after keywords), with 8,792 distinct targeted values. A majority of these values (4,126) were associated with a single participant (e.g., one participant was targeted as a follower look-alike of @FDAOncology while 26 users were targeted as follower lookalikes of @SpeakerPelosi). However, a few well-known handles were frequently the focus of lookalikes: @netflix (used in targeting 5,199 ads), @espn (3,608), and @nytimes (3,440).

BEHAVIOR targeting, one specific targeting type offered by Twitter within the full range of psychographic targeting types, is based on inferences drawn from proprietary algorithms. Our most commonly observed instances were related to income or lifestyles (e.g., "US - Household income: $30,000 - $39,999," "US - Executive/C-suite," "US - Presence in household: yes ," "US - Fit moms"). Some were surprisingly specific: "Home insurance expiration month: 10 October," "US - Likely to switch cell phone providers," "Country Club Climbers - Suburban Empty Nesters: K59," and "US - Animal charity donors."

Finally, Twitter offers four retargeting types, based on previous user engagement with ads. There were 15,814 uses (1,812 unique instances) of RETARGETING CAMPAIGN targeting, which targets users who responded to an advertiser's prior campaign. The ambiguous naming of these instances ("Retargeting campaign engager: ########") makes them hard to interpret in detail. RETARGETING USER ENGAGER, used 707 times, is similarly vague. RETARGETING CUSTOM AUDIENCE LOOKALIKE TARGETING, which combines retargeting with Twitter's lookalike algorithms, was very rarely used in our data.

4.3.3 Advertiser Targeting Types

The final category of targeting types use advertiser-provided information. Instead of providing any targeting data, Twitter only facilitates matching to Twitter users via Twitter usernames, email addresses, or other identifiers. Notably, adver-

tiser targeting types are also the most covert from a user's perspective: while Twitter-provided data could potentially be deduced from the standard user interface (e.g., interests based on likes or Retweets), targeting types using advertiser-provided data are completely unrelated to Twitter activity.

TAILORED AUDIENCE (LISTS) match Twitter users to lists uploaded by advertisers. We found 113,952 instances of list targeting across 2,338 unique lists; companies using list targeting the most were Anker (22,426 instances), Postmates (11,986), Rockstar Games (8,494), and Twitter Surveys (3,131). Tailored lists often used words like 'Negative', 'Holdout', and 'Blacklist', which we hypothesize reference consumers who previously opted out of receiving targeted ads or content via other mediums. Advertisers may also use list targeting for targeting offline purchasers, as list names included the words 'Purchase' and 'Buyers.' Many lists use naming schemes that make it difficult or impossible to discern the contents of the lists (e.g. "#####_#_########", "###_MM_YY_*******_#####").

We identified several lists with names that suggest targeting on attributes prohibited by Twitter's policies (see Table 2), including financial status ("YYYY account status: balance due"), race ("***_Nissan_AfricanAmericans_YYYYMM"), religion ("Christian Audience to Exclude"), or sex life ("LGBT Suppression List") [66]. TAILORED AUDIENCE (WEB) also consists of advertiser-collected lists of website visitors, e.g., "Started New Credit Card Application" or "Registered but not Activated User on Cloud." This targeting type therefore connects users' potentially sensitive browsing activity to their Twitter accounts in ways that may violate Twitter's health advertising policies [64].

TAILORED AUDIENCE CRM LOOKALIKE targeting combines advertiser lists with the lookalike algorithm to find Twitter users who may be similar to known current or potential customers. We observed this mechanism being used in incredibly specific ways, such as to find users similar to "QSR Ice Cream Frozen Yogurt Frequent Spender" or "Frozen_Snacks_Not_Frozen_Yogurt_Or_Ice_Cream _Used_in_last_6_months_Principal_Shoppers_Primary _Fla_Vor_Ice_###," both used by advertiser Dairy Queen.

Twitter also offers targeting types that enable cross-platform tracking. MOBILE AUDIENCE targets Twitter users who also use an advertiser-owned mobile app (i.e., "people who have taken a specific action in your app, such as installs or sign-ups" [68]). Instances reflect the user's status with the app, app name, and mobile platform, e.g., "Install Gemini: Buy Bitcoin Instantly ANDROID All" and "Install Lumen - Over 50 Dating IOS All". MOBILE AUDIENCE LOOKALIKE targeting, which combines the prior mechanism with the lookalike algorithm, was rarely used. FLEXIBLE AUDIENCE targeting allows advertisers to combine tailored audiences (lists, web, or mobile) using AND, OR, and NOT operations. We observed seven ads using this type, all from one advertiser.

Targeting Value	Policy	Advertiser(s)
Keywords		
ostomy	Health	ConvaTec Stoma UK
unemployment	Financial	Giant Eagle Jobs
Gay	Sex Life	H&M United Kingdom
mexican american	Race	Just Mercy, Doctor Sleep, The Kitchen Movie
#AfricanAmerican	Race	sephora
#native	Race	sephora
hispanics	Race	sephora
latinas	Race	sephora
mexican	Race	sephora
-Racist	Religion	xbox
Conversation Topics		
Liberal Democrats (UK)	Politics	Channel 5, Irina von Wiese MEP
Tailored Audience (List)		
YYYY account status: balance due (*translated from Mandarin Chinese*)	Financial	Anker
segment_Control | Rising Hispanics | Email Openers_########	Race	Big Lots
segment_Control | Rising Hispanics | Non-Opener_########	Race	Big Lots
***_Nissan_AfricanAmericans_YYYYMM	Race	Nissan
Christian Audience to Exclude	Religion	nycHealthy
LGBT Suppression List	Sex Life	nycHealthy
ASL Marketing > Hispanic Millennials - ##########	Race	Verizon
Tailored Audience (Web)		
Website Retargeting - Tagrisso.com (*a site about lung cancer therapy*)	Health	Test Lung Cancer

Table 2: Examples of targeted ads that could be seen as violating Twitter's keyword targeting policy (see Appendix F, Figure 7 [1]) or Twitter's privacy policy: "...our ads policies prohibit advertisers from targeting ads based on categories that we consider sensitive or are prohibited by law, such as race, religion, politics, sex life, or health" [69].

Finally, for the curiously-named targeting type UNKNOWN, 25 participants were associated with a single instance ("Unknown: ####"), all related to the advertiser "Twitter Surveys."

4.4 Participant Reactions to Targeting Types

One key benefit of our study design is that we could ask participants questions about advertising criteria actually used in ads they saw. Participants answered questions about up to four randomly selected targeting types, filtered by those present in their uploaded data. Advertisers used certain targeting types more often than others, meaning different numbers of participants saw each type (see Appendix G, Table 4 [1]).

4.4.1 Fairness, Comfort, Desirability, and Accuracy

Participants perceived LANGUAGE, AGE, and INTEREST targeting to be the most fair, with 86.3%, 72.0%, and 69.0% agreeing respectively (Figure 4). Overall, few participants thought any given targeting type was unfair to use: no type had more than 50% of participants disagree that its use would be fair (Figure 4, General: Fair). TAILORED AUDIENCE (LIST), which was perceived as least fair overall, was still roughly evenly split

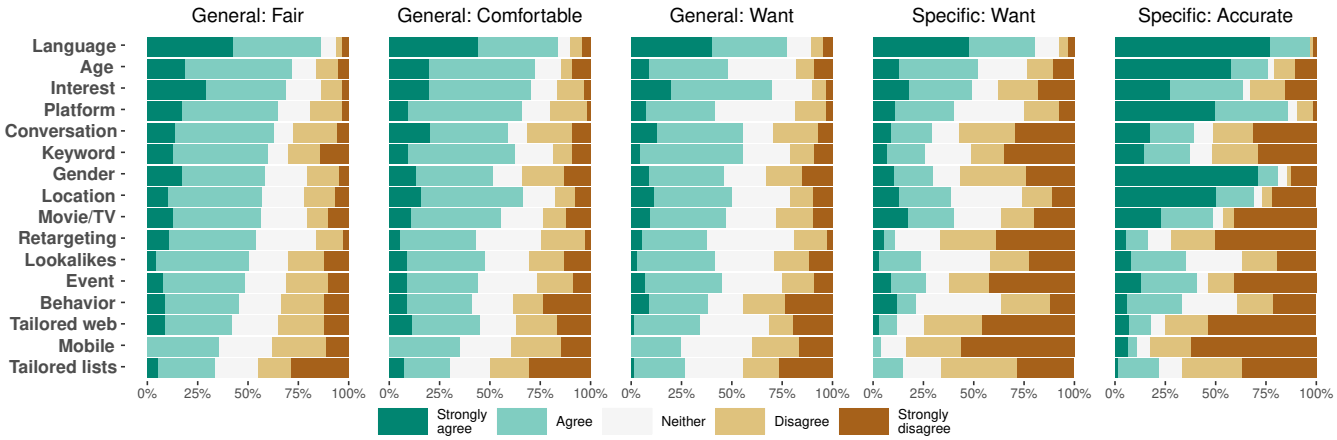


Figure 4: Participants' level of agreement to questions about targeting types in *general* and *specific* instances.

between participants agreeing and disagreeing. Compared to the regression baseline (INTEREST), participants were significantly more likely to find LANGUAGE targeting fair ($OR = 4.48$, $p < 0.001$). RETARGETING CAMPAIGN, AGE, and PLATFORM targeting were not statistically different from INTEREST ($\alpha = 0.05$). Participants found all other targeting types significantly less fair than INTEREST ($OR = 0.0607 - 0.401$, all $p < 0.05$).

To dig deeper into perceptions of fairness, we asked participants to elaborate on their Likert-scale answers in a free-response question, gathering a total of 898 responses. Participants had varying conceptions of the meaning of fairness. Some equated fairness with utility, some equated fairness with comfort, and some equated fairness with accuracy of the information. Across all targeting types, the most common rationale used to judge fairness were that targeting is useful to the user in some way (24.8%). For instance, participants mentioned that they preferred to see relevant rather than random ads if they had to see ads at all, and that advertising allows them to access Twitter for free. 14.6% said that targeting was fair because the advertiser benefited in some way, namely by increased effectiveness of advertising. These two rationales centered on deriving benefits, either for advertisers or users, but failed to consider the privacy or data autonomy of the participant. Others considered that Twitter is a public platform. "Twitter is pretty much a public arena, if I were shouting about various topics in a town square, people would infer my interests from that, and potentially attempt to profit from them" (P191). Participants' rationales seemed to assume that personalized targeting types like these *must* be used for advertising. Only a few suggested profiting off of users' private information was fundamentally unfair.

Perceptions of comfort largely aligned with perceptions of fairness, with small exceptions. For example, participants rated GENDER and KEYWORD targeting as more fair than LOCATION targeting, but were curiously more comfortable with LOCATION than GENDER and KEYWORD (Figure 4, General: Comfortable). Some participants' comments suggested discomfort may relate to whether participants understood how data about them was obtained. P184 commented, "I'm not sure how they would know my income level. Disturbing."

We were also curious about participants' desire for advertising that used each targeting type and found general affirmation, with some strong opposition to specific instances. We told participants to assume the number of ads they would see would stay the same and asked them to consider how much they would want to see ads targeted with a given type, for both a specific instance of that type and for type generally. As an example, 53.8% of participants who saw an instance of EVENT targeting disagreed that it described them accurately and 65.0% disagreed that they would want to see advertising based on that specific example. However, only 25.0% disagreed that they would want to see ads utilizing event targeting in general.

In the general case, participants were significantly more likely to want ads that used LANGUAGE targeting than the regression-baseline INTEREST ($OR = 3.3$, $p = 0.004$). All other targeting types were significantly less wanted than INTEREST ($OR = 0.1 - 0.4$, all $p < 0.05$).

Participants found specific instances of some demographic targeting types to be very accurate, but other psychographic types to be very inaccurate. More than half of participants strongly agreed that a specific instances of LANGUAGE, AGE, PLATFORM, GENDER, LOCATION targeting was accurate for them, while more than half strongly disagreed that RETARGETING, TAILORED WEB, and MOBILE targeting was accurate (Figure 4, Specific: Accurate). Participants were more likely to agree that specific instances of PLATFORM, LANGUAGE, GENDER, and AGE targeting described them accurately compared to a specific instance of INTEREST ($OR = 2.9 - 9.7$, all $p < 0.01$). Specific instances of MOVIES AND TV SHOWS, LOCATION, and BEHAVIOR targeting were not significantly different from INTEREST in agreed accuracy ($\alpha = 0.05$), while all remaining significant targeting types were less likely to be rated as accurate ($OR = 0.1 - 0.5$, all $p < 0.05$). As we found in their initial free-

Response	ρ	p
General: Fair	0.332	**<.001**
General: Comfortable	0.366	**<.001**
General: Want	0.341	**<.001**
Specific: Comfortable	0.614	**<.001**
Specific: Want	0.732	**<.001**

Table 3: Spearman's ρ correlation between participants' agreement with Specific: Accurate ("*Specific instance* describes me accurately") and their other Likert-scale responses.

response reactions to uses of a particular targeting type in their data, if participants perceived an instance of targeting to be accurate, it was generally well-received. Participants seemed to enjoy seeing information being accurately reflected about themselves, as P189 described about CONVERSATION targeting: "I am okay with this. It's cool how accurate it is."

As shown in Table 3, the accuracy of a specific instance of a targeting type was significantly correlated with all of our other measurements of participants' perceptions. That is, when participants disagreed that a specific instance of a targeting type described them accurately, they were also significantly less likely to be comfortable with that instance being used ($\rho = 0.614, p < 0.001$) and to want to see more ads based on that instance ($\rho = 0.732, p < 0.001$). We found similar correlations for perceptions of the use of a targeting type generally. It is possible that inaccuracy leads to perceptions of discomfort and unwantedness; it is also possible that when people see ads they find undesirable, they are less likely to believe the associated targeting is accurate.

Even if a majority of people are comfortable with certain targeting in the abstract, it is important to understand, and potentially design for, those who feel less comfortable. To explore this, we looked for participants who consistently disagreed with questions about fairness, comfort, and desirability. In particular, for each of the questions presented in Figure 4 besides Specific: Accurate, we generated a median response for each participant of the up to four targeting types they were asked questions about. From this, we found only 23 of our 231 participants disagreed or strongly disagreed as their median response for all 4 questions.

4.4.2 Targeting Types: Awareness and Reactions

We were also interested in participants' familiarity with, or misconceptions of, the various targeting types. Before participants were given any information about a targeting type, we showed them the term Twitter uses to describe that type [63] and asked them to indicate their current understanding or best guess of what that term meant in the context of online advertising. Nearly all participants had accurate mental models of LOCATION, AGE, GENDER, and KEYWORD targeting, likely because these types are fairly well-known and straightforward. Further, 93% of participants asked about INTEREST correctly defined it, suggesting it is also relatively straightforward. In fact, some participants confused other targeting types with INTEREST targeting: "I have never heard this term before. I'm guessing that they target ads based on your followers' interests as well?" (P161 on FOLLOWER LOOKALIKE targeting).

TAILORED AUDIENCE (LIST), BEHAVIOR, and MOBILE AUDIENCE targeting were the least well understood, with 96.4%, 97.0%, and 100% of participants, respectively, providing an incorrect or only partially correct definition. The first two rely on offline data being connected with participants' online accounts, but most participants incorrectly defined the term only based on online activities. MOBILE AUDIENCE targeting was misunderstood due to different interpretations of "mobile" (e.g., P122 guessed, "advertising based on your phone network?") or other mobile details. The correct answer relates to the user's interactions with mobile apps. Participants also frequently believed a targeting type meant advertising that type of thing (e.g., an event) as opposed to leveraging user data about that thing for targeting ads (e.g., targeting a product only to users who attended an event).

While 63.6% of participants who were asked to define LANGUAGE targeting correctly referenced the user's primary language, many of the 28.8% who incorrectly defined it posed a more involved, and potentially privacy-invasive, definition: "I suppose that language targeting would be communicating in a way that is targeted to how that specific person communicates. For example, as a millennial I would want to see language that is similar to how I speak rather than how someone who is my parents age would speak" (P76). PLATFORM targeting was similarly misunderstood, with some participants believing that this was the practice of targeting by social media *platform* use or even political *platform*: "It looks at my list of people I follow and sends me ads based on what they believe my political stance is" (P147). We also found evidence, across targeting types, of the belief that advertising is based on surreptitious recordings of phone audio. For example, P231 said of CONVERSATION targeting: "Given what I know about how phone microphones are always on, I would guess that it's when ads pop up based on what I've said in a conversation."

4.5 Participant Responses to Ad Explanations

We examined reactions to our ad explanations among the 193 participants who saw all six variants. Our approximation of Twitter's current explanation served as our primary basis of comparison. We also report qualitative opinions about what was memorable, perceived to be missing, or would be ideal.

4.5.1 Comparing Our Ad Explanations to Twitter's

Overall, participants found explanations containing more detail to be more useful, as shown in Figure 5. Unsurprisingly, Control was the least useful explanation; only 31.3% of participants agreed it was useful. This is significantly less than our Twitter baseline, where 48.8% agreed ($V = 6344.5$,

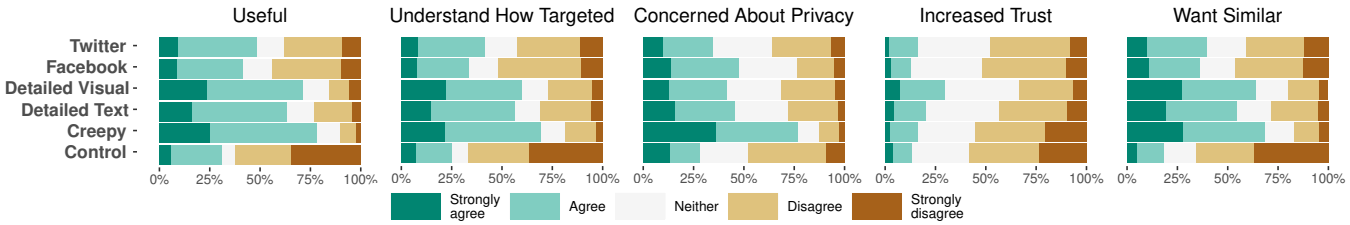


Figure 5: Participants' level of agreement to questions about ad explanations.

$p < 0.001$). The Facebook explanation was comparable to the Twitter explanation (41.4% agreed; $V = 3520.0$, $p = 0.154$). In contrast, the three explanations we designed were rated as significantly more useful than Twitter's ($V = 1352.5$–2336.0, all $p < 0.001$). Specifically, 63.6%, 71.2% and 78.6% of participants respectively agreed the Detailed Text, Detailed Visual, and Creepy explanations were useful.

The usefulness ratings closely resembled responses to whether the explanation provided "enough information to understand how the ad was chosen for me." Again, Twitter performed better than only Control (V: 5906.0, $p < 0.001$), and did not significantly differ from Facebook ($V = 4261.0$, $p = 0.091$). Participants agreed our explanations—Detailed Text, Detailed Visual, Creepy—were most helpful in understanding how they were targeted; all three significantly differed from Twitter ($V = 1928.0$–2878.0, all $p \leq 0.001$).

We saw a different trend for privacy concern: 77.2% of participants agreed Creepy made them "more concerned about my online privacy," compared to 34.8% for Twitter, and just 28.2% for the Control. Privacy concern for Creepy was significantly higher than for Twitter ($V = 989.5$, $p < 0.001$). Both Facebook and Detailed Text also exhibited significantly more concern than Twitter ($V = 1821.0$, 2835.0; $p = 0.002$, 0.015), but to a lesser extent. Respondents reported comparable privacy concern for the Twitter explanation as for Detailed Visual and Control ($V = 2029.5$, 3751.0, $p = 0.080$,0.064).

Transparency and usefulness generally did not translate to increased trust in an advertiser. In fact, only a minority of participants agreed that they trusted the advertiser more as a result of any provided ad explanation. Only the Detailed Visual explanation increased trust significantly relative to Twitter ($V = 1695.5$, $p < 0.001$).

A majority of participants agreed they would "want an ad explanation similar to this one for all ads I see on Twitter" for our Creepy (68.8%), Detailed Visual (64.4%), and Detailed Text (54.9%) versions. Agreement for these was significantly larger ($V = 1798.5$–2132.0, all $p < 0.001$) than the 39.8% who wanted Twitter-like. Participants significantly preferred Twitter to the Control ($V = 6831.5$, $p < 0.001$), but not to Facebook ($V = 4249.0$, $p = 0.339$).

4.5.2 Qualitative Responses to Ad Explanations

Participants want detail and indicators of non-use. We asked participants what they found most memorable about each ad explanation. For Control, Facebook, and Twitter, most memorable was how little detail they gave about how participants were targeted (30.7%, 21.6%, and 13.3% of participants, respectively). By comparison, 16.3% (Detailed Text), 7.9% (Visual), and 3.1% (Creepy) of participants noted a lack of detail as the most memorable part. Conversely, 81.7% found the amount of detail in Creepy to be the most memorable part, followed by 61.2% for Visual. These findings may be because Creepy included the most information and Detailed Visual indicated which targeting types were *not* used.

Ambiguity was perceived as missing information. We also asked participants what information, if any, they thought was missing from each ad explanation. We wanted to help participants identify what information could be missing, so our within-subjects design featured randomly-shown variants that demonstrated information that could be included. In line with the quantitative results for usefulness, our Detailed Visual, Detailed Text, and Creepy explanations performed best, with 61.2%, 58.9%, and 53.0% of participants, respectively, answering nothing was missing. Conversely, Facebook, Control, and Twitter performed less well, with 69.2%, 67.3%, and 52.4%, respectively, of participants stating that some information was missing or unclear. For Detailed Text and Detailed Visual, among the most commonly noted missing information related to our use of "may" and "might" about which criteria actually were matched the participant. This was necessitated by the ambiguity of the Twitter files (prior to receiving a clarification from Twitter; see Section 3.6 for details). For Facebook, the most commonly missing information was associated with the hashed tailored audience list: several wrote that they did not know what a hashed list was. P125 wrote, "The nature of the list mentioned should be clarified in some detail. It's unfair to be put on a list without access to what the list is and who compiled it and who has access to it."

Describing their ideal Twitter ad explanation, 46.8% of participants wanted to see the specific actions (e.g., what they Tweeted or clicked on) or demographics that caused them to see a given ad. 34.2% wanted to know more about how the advertiser obtained their information. They also wanted clear language (19.0%) and settings for controlling ads (13.4%).

5 Discussion

We study Twitter's targeted advertising mechanisms, which categorize users by demographic and psychographic attributes, as determined from information provided by the user, provided by advertisers, or inferred by the platform. While prior work has surfaced and studied user reactions to ad targeting as a whole [20, 70], or specific mechanisms like inferred interests [17], our work details advertisers' use of 30 unique targeting types and investigates user perceptions into 16 of them. These distinct types, including FOLLOWER LOOKALIKES and TAILORED AUDIENCES, are rarely studied by the academic community, but frequently used by advertisers (see Table 1). Our participants expressed greater discomfort with some of these less studied targeting types, highlighting a need for future work.

We complement existing work on Facebook ad transparency by investigating ad explanations on a different platform, Twitter, and using participants' own Twitter data to evaluate them. Strengthening prior qualitative work [20], we quantitatively find that our participants preferred ad explanations with richer information than currently provided by Facebook and Twitter. We also find significant user confusion with "hashed" lists, a term introduced to ad explanations by Facebook in 2019 [55] to explain how platforms match user data to information on advertiser-uploaded lists for TAILORED AUDIENCE targeting (called custom audiences on Facebook).

Can sensitive targeting be prohibited in practice? We find several instances of ad targeting that appear to violate Twitter's stated policy prohibiting targeting on sensitive attributes. Such targeting is often considered distasteful and in some cases may even be illegal. We observed these instances most commonly in targeting types where advertisers provide critical information: KEYWORDS (where advertisers can provide any keyword of choice, subject to Twitter acceptance) and variations of TAILORED AUDIENCES, where the advertiser provides the list of users to target. Potentially discriminatory keywords are a problem that Twitter could theoretically solve given a sufficiently accurate detection algorithm or (more likely) manual review. TAILORED AUDIENCES, however, are more pernicious. Advertisers can use any criteria to generate a list. We were only able to identify potentially problematic cases because the list name, which is under advertiser control, happened to be meaningfully descriptive. It would be trivial for an advertiser to name a list generically to skirt scrutiny, calling into question whether Twitter's policy on sensitive attributes has (or can have) any real force in practice. It also raises larger concerns about regulating potentially illegal or discriminatory practices as long as tailored audiences remain available.

More accuracy, fewer problems? Similarly to prior work, we found that the perceived inaccuracy of targeting instances correlates with users having less desire for such targeting to be used for them [14, 17]. This has potentially dangerous implications. If accuracy reduces discomfort, this may appear to justify increasing invasions of privacy to obtain evermore-precise labels for users. However, participants' free-text responses indicate an upper bound where increasing accuracy is no longer comfortable. For example, P220 noted that a specific instance of LOCATION targeting was "very accurate, ... but I don't really like how they are able to do that without my knowledge and even show me ad content related to my location, because I choose not to put my specific location on my Twitter account in any way for a reason." Future work should investigate how and when accuracy crosses the line from useful to creepy.

Transparency: A long way to go. This work also contributes a deeper understanding of ad explanations, amid substantial ongoing work on transparency as perhaps the only way for the general public to scrutinize the associated costs and benefits. Participants found our ad explanations, which provide more details, significantly more useful, understandable, and desirable than currently deployed ad explanations from Twitter and Facebook. However, our results also highlight a significant challenge for transparency: platform and advertiser incentives. Some of our proposed explanations, despite being more useful, decreased participant trust in the advertiser, which clearly presents a conflict of interest. This conflict may explain why currently deployed explanations are less complete or informative than they could be.

Finally, our results suggest it is insufficient to simply require data processing companies to make information available. While the option to download advertising data is a strong first step, key aspects of the ad ecosystem — such as the origins of most targeting information — remain opaque. In addition, even as researchers with significant expertise, we struggled to understand the data Twitter provided (see Section 3.6). This creates doubt that casual users can meaningfully understand and evaluate the information they receive. However, our participants indicated in free response answers that they found the transparency information provided in our study useful and that it illuminated aspects of tracking they had not previously understood, making it clear that comprehensible transparency has value. We therefore argue that transparency regulations should mandate that raw data files be accompanied by clear descriptions of their contents, and researchers should develop tools and visualizations to make this raw data meaningful to users who want to explore it.

Acknowledgments

We gratefully acknowledge support from the Data Transparency Lab and Mozilla, as well as from a UMIACS contract under the partnership between the University of Maryland and DoD. The views expressed are our own.

References

[1] Online appendix. https://www.blaseur.com/papers/usenix20twitterappendix.pdf.

[2] Gunes Acar, Christian Eubank, Steven Englehardt, Marc Juarez, Arvind Narayanan, and Claudia Diaz. The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. In *Proc. CCS*, 2014.

[3] Lalit Agarwal, Nisheeth Shrivastava, Sharad Jaiswal, and Saurabh Panjwani. Do Not Embarrass: Re-Examining User Concerns for Online Tracking and Advertising. In *Proc. SOUPS*, 2013.

[4] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination Through Optimization: How Facebook's Ad Delivery Can Lead to Skewed Outcomes. In *Proc. CSCW*, 2019.

[5] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Cranor, and Yuvraj Agarwal. Your Location has been Shared 5,398 Times! A Field Study on Mobile App Privacy Nudging. In *Proc. CHI*, 2015.

[6] Athanasios Andreou, Márcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. Measuring the Facebook Advertising Ecosystem. In *Proc. NDSS*, 2019.

[7] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. In *Proc. NDSS*, 2018.

[8] Julia Angwin and Terry Parris. Facebook Lets Advertisers Exclude Users by Race. ProPublica, October 28, 2016.

[9] @ashrivas. More Relevant Ads with Tailored Audiences. Twitter Blog, December 2013. https://blog.twitter.com/marketing/en_us/a/2013/more-relevant-ads-with-tailored-audiences.html.

[10] Rebecca Balebako, Pedro Leon, Richard Shay, Blase Ur, Yang Wang, and Lorrie Faith Cranor. Measuring the Effectiveness of Privacy Tools for Limiting Behavioral Advertising. In *Proc. W2SP*, 2012.

[11] Muhammad Ahmad Bashir, Sajjad Arshad, and William Robertson. Tracing Information Flows Between Ad Exchanges Using Retargeted Ads. In *Proc. USENIX Security*, 2016.

[12] Muhammad Ahmad Bashir, Umar Farooq, Maryam Shahid, Muhammad Fareed Zaffar, and Christo Wilson. Quantity vs. Quality: Evaluating User Interest Profiles Using Ad Preference Managers. In *Proc. NDSS*, 2019.

[13] Muhammad Ahmad Bashir and Christo Wilson. Diffusion of User Tracking Data in the Online Advertising Ecosystem. In *Proc. PETS*, 2018.

[14] Rena Coen, Emily Paul, Pavel Vanegas, Alethea Lange, and G.S. Hans. A User-Centered Perspective on Algorithmic Personalization. Master's thesis, Berkeley School of Information, https://www.ischool.berkeley.edu/projects/2016/user-centeredperspective-algorithmic-personalization, 2016.

[15] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings. In *Proc. PETS*, 2015.

[16] Martin Degeling and Jan Nierhoff. Tracking and Tricking a Profiler: Automated Measuring and Influencing of Bluekai's Interest Profiling. In *Proc. WPES*, 2018.

[17] Claire Dolin, Ben Weinshel, Shawn Shan, Chang Min Hahn, Euirim Choi, Michelle L. Mazurek, and Blase Ur. Unpacking Privacy Perceptions of Data-Driven Inferences for Online Targeting and Personalization. In *Proc. CHI*, 2018.

[18] Serge Egelman, Adrienne Porter Felt, and David Wagner. Choice Architecture and Smartphone Privacy: There's A Price for That. In *Workshop on the Economics of Information Security*, 2012.

[19] Steven Englehardt and Arvind Narayanan. Online Tracking: A 1-Million-Site Measurement and Analysis. In *Proc. CCS*, 2016.

[20] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. Communicating Algorithmic Process in Online Behavioral Advertising. In *Proc. CHI*, 2018.

[21] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. "Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In *Proc. AAAI*, 2017.

[22] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. User Attitudes towards Algorithmic Opacity and Transparency in Online Reviewing Platforms. In *Proc. CHI*, 2019.

[23] Facebook. Coming Soon: New Ways to Reach People Who've Visited Your Website or Mobile App. Facebook Business, October 15, 2013. https://www.facebook.com/business/news/custom-audiences.

[24] Facebook. Introducing New Requirements for Custom Audience Targeting. Facebook Business, June 2018. https://www.facebook.com/business/news/introducing-new-requirements-for-custom-audience-targeting.

[25] Irfan Faizullabhoy and Aleksandra Korolova. Facebook's Advertising Platform: New Attack Vectors and the Need for Interventions. In *Proc. ConPro*, 2018.

[26] Faye W. Gilbert and William E. Warran. Psychographic Constructs and Demographic Segments. *Psychology and Marketing*, 12:223–237, 1995.

[27] Google. About Audience Targeting. Google Ads Help, 2020. https://support.google.com/google-ads/answer/2497941?hl=en.

[28] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy Attitudes of Mechanical Turk Workers and the U.S. Public. In *Proc. USENIX Security*, 2014.

[29] Saranga Komanduri, Richard Shay, Greg Norcie, and Blase Ur. Adchoices? Compliance with Online Behavioral Advertising Notice and Choice Requirements. *IS: A Journal of Law and Policy for the Information Society*, 7:603, 2011.

[30] Georgios Kontaxis and Monica Chew. Tracking Protection in Firefox for Privacy and Performance. In *Proc. W2SP*, 2015.

[31] Balachander Krishnamurthy, Delfina Malandrino, and Craig E. Wills. Measuring Privacy Loss and the Impact of Privacy Protection in Web Browsing. In *Proc. SOUPS*, 2007.

[32] @KyleB. Introducing Partner Audiences. Twitter Blog, March 5, 2015. https://blog.twitter.com/marketing/en_us/a/2015/introducing-partner-audiences.html.

[33] J. Richard Landis and Gary G. Koch. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174, 1977.

[34] Mathias Lécuyer, Guillaume Ducoffe, Francis Lan, Andrei Papancea, Theofilos Petsios, Riley Spahn, Augustin Chaintreau, and Roxana Geambasu. XRay: Enhancing the Web's Transparency with Differential Correlation. In *Proc. USENIX Security*, 2014.

[35] Mathias Lecuyer, Riley Spahn, Yannis Spiliopolous, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. Sunlight: Fine-grained Targeting Detection at Scale with Statistical Confidence. In *Proc. CCS*, 2015.

[36] Pedro Leon, Justin Cranshaw, Lorrie Faith Cranor, Jim Graves, Manoj Hastak, Blase Ur, and Guzi Xu. What do Online Behavioral Advertising Privacy Disclosures Communicate to Users? In *Proc. WPES*, 2012.

[37] Pedro Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Cranor. Why Johnny Can't Opt out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising. In *Proc. CHI*, 2012.

[38] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016. In *Proc. USENIX Security*, 2016.

[39] Qiang Ma, Eeshan Wagh, Jiayi Wen, Zhen Xia, Robert Ormandi, and Datong Chen. Score Look-Alike Audiences. In *Proc. ICDMW*, 2016.

[40] Aleksandar Matic, Martin Pielot, and Nuria Oliver. "OMG! How did it know that?" Reactions to Highly-Personalized Ads. In *Proc. UMAP*, 2017.

[41] Jonathan R Mayer and John C Mitchell. Third-party Web Tracking: Policy and Technology. In *Proc. IEEE S&P*, 2012.

[42] Aleecia M. McDonald and Lorrie Faith Cranor. Americans' Attitudes About Internet Behavioral Advertising Practices. In *Proc. WPES*, 2010.

[43] Jeremy B. Merrill and Ariana Tobin. Facebook Moves to Block Ad Transparency Tools — Including Ours. ProPublica, January 28, 2019. https://www.propublica.org/article/facebook-blocks-ad-transparency-tools.

[44] Katie Notopoulos. Twitter Has Been Guessing Your Gender And People Are Pissed. Buzzfeed News, May 2017. https://www.buzzfeednews.com/article/katienotopoulos/twitter-has-been-guessing-your-gender-and-people-are-pissed.

[45] Jason R. C. Nurse and Oliver Buckley. Behind the Scenes: a Cross-Country Study into Third-Party Website Referencing and the Online Advertising Ecosystem. *Human-centric Computing and Information Sciences*, 7(1):40, 2017.

[46] Eyal Peer, Laura Brandimarte, Sonam Somat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. In *Journal of Experimental Social Psychology*, 2017.

[47] Angelisa C. Plane, Elissa M. Redmiles, Michelle L. Mazurek, and Michael Carl Tschantz. Exploring User Perceptions of Discrimination in Online Targeted Advertising. In *Proc. USENIX Security*, 2017.

[48] Emilee Rader and Rebecca Gray. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proc. CHI*, 2015.

[49] J. H. Randall. The Analysis of Sensory Data by Generalized Linear Model. *Biometrical Journal*, 1989.

[50] Ashwini Rao, Florian Schaub, and Norman Sadeh. What Do They Know About Me? Contents and Concerns of Online Behavioral Profiles. In *Proc. ASE BigData*, 2014.

[51] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *Proc. IEEE S&P*, 2019.

[52] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. Detecting and Defending Against Third-Party Tracking on the Web. In *Proc. USENIX Security*, 2012.

[53] Sonam Samat, Alessandro Acquisti, and Linda Babcock. Raise the Curtains: The Effect of Awareness About Targeting on Consumer Attitudes and Purchase Intentions. In *Proc. SOUPS*, 2017.

[54] Florian Schaub, Aditya Marella, Pranshu Kalvani, Blase Ur, Chao Pan, Emily Forney, and Lorrie Faith Cranor. Watching Them Watching Me: Browser Extensions' Impact on User Privacy Awareness and Concern. In *Proc. USEC*, 2016.

[55] Ramya Sethuraman. Why Am I Seeing This? We Have an Answer for You. Facebook Blog, March 2019. https://about.fb.com/news/2019/03/why-am-i-seeing-this/.

[56] Matt Southern. LinkedIn Now Lets Marketers Target Ads to 'Lookalike Audiences'. Search Engine Journal, March 20, 2019. https://www.searchenginejournal.com/linkedin-now-lets-marketers-target-ads-to-lookalike-audiences/299547/.

[57] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna Gummadi, Patrick Loiseau, and Alan Mislove. Potential for Discrimination in Online Targeted Advertising. In *Proc. FAT*, 2018.

[58] Darren Stevenson. *Data, Trust, and Transparency in Personalized Advertising*. PhD thesis, University of Michigan, 2016.

[59] Latanya Sweeney. Discrimination in Online Ad Delivery. *CACM*, 56(5):44–54, 2013.

[60] Michael Carl Tschantz, Serge Egelman, Jaeyoung Choi, Nicholas Weaver, and Gerald Friedland. The Accuracy of the Demographic Inferences Shown on Google's Ad Settings. In *Proc. WPES*, 2018.

[61] Joseph Turow, Jennifer King, Chris Jay Hoofnagle, Amy Bleakley, and Michael Hennessy. Americans Reject Tailored Advertising and Three Activities that Enable It. Technical report, Annenberg School for Communication, 2009.

[62] Gerhard Tutz and Wolfgang Hennevogl. Random effects in ordinal regression models. *Computational Statistics and Data Analysis*, 1996.

[63] Twitter. Ad Targeting Best Practices for Twitter. Twitter Business, 2019. https://business.twitter.com/en/targeting.html.

[64] Twitter. Healthcare. Twitter Business, 2019. https://business.twitter.com/en/help/ads-policies/restricted-content-policies/health-and-pharmaceutical-products-and-services.html.

[65] Twitter. Keyword targeting. Twitter Business, 2019. https://business.twitter.com/en/help/campaign-setup/campaign-targeting/keyword-targeting.html.

[66] Twitter. Policies for conversion tracking and tailored audiences. Twitter Business, 2019. https://business.twitter.com/en/help/ads-policies/other-policy-requirements/policies-for-conversion-tracking-and-tailored-audiences.html.

[67] Twitter. Target based on how people access Twitter. Twitter Business, 2019. https://business.twitter.com/en/targeting/device-targeting.html.

[68] Twitter. Intro to Tailored Audiences. Twitter Business, 2020. https://business.twitter.com/en/help/campaign-setup/campaign-targeting/tailored-audiences.html.

[69] Twitter. Twitter Privacy Policy. https://twitter.com/en/privacy, 2020. Accessed February 13, 2020.

[70] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In *Proc. SOUPS*, 2012.

[71] Giridhari Venkatadri, Athanasios Andreou, Yabing Liu, Alan Mislove, Krishna Gummadi, Patrick Loiseau, and Oana Goga. Privacy Risks with Facebook's PII-based Targeting: Auditing a Data Broker's Advertising Interface. In *Proc. IEEE S&P*, 2018.

[72] Giridhari Venkatadri, Piotr Sapiezynski, Elissa Redmiles, Alan Mislove, Oana Goga, Michelle Mazurek, and Krishna Gummadi. Auditing Offline Data Brokers via Facebook's Advertising Platform. In *Proc. WWW*, 2019.

[73] Jeffrey Warshaw, Nina Taft, and Allison Woodruff. Intuitions, Analytics, and Killing Ants: Inference Literacy of High School-educated Adults in the US. In *Proc. SOUPS*, 2016.

[74] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L. Mazurek, and Blase Ur. Oh, the Places You've Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing. In *Proc. CCS*, 2019.

[75] Craig E. Wills and Can Tatar. Understanding What They Do with What They Know. In *Proc. WPES*, 2012.

[76] Yuxi Wu, Panya Gupta, Miranda Wei, Yasemin Acar, Sascha Fahl, and Blase Ur. Your Secrets are Safe: How Browers' Explanations Impact Misconceptions About Private Browsing Mode. In *Proc. WWW*, 2018.

[77] Yaxing Yao, Davide Lo Re, and Yang Wang. Folk Models of Online Behavioral Advertising. In *Proc. CSCW*, 2017.

A Definitions of Targeting Types

In this section, we provide the terminology and definitions for the 16 targeting types we investigated in Part 1 of our user study. We took all terminology and definitions that we showed participants verbatim from Twitter Business's help pages (`https://business.twitter.com/en/targeting.html` and pages linked from it).

- **Age targeting** allows advertisers to target people by age buckets, such as 18+ years old or 18-24 years old.
- **Behavior targeting** allows advertisers to target people based on inferred behavior, such as shopping and lifestyle habits or income.
- **Conversation topic targeting** allows advertisers to target people based on topics they have engaged with (e.g., Tweeted, clicked, Retweeted, replied, liked, viewed) on Twitter.
- **Event targeting** allows advertisers to target people based on events they are interested in or have engaged with (e.g., Tweeted, clicked, Retweeted, replied, liked, viewed) on Twitter.
- **Follower lookalike targeting** allows advertisers to target people who don't necessarily follow a given account, but have similar interests or demographics to the account's actual followers.
- **Gender targeting** allows advertisers to target people based on their self-reported or inferred gender.
- **Interest targeting** allows advertisers to target people based on inferred interests, as determined by who they follow on Twitter and their Tweets, Retweets, and clicks.
- **Keyword targeting** allows advertisers to target people based on words or phrases they have Tweeted about or searched for on Twitter.
- **Language targeting** allows advertisers to target people who use a certain language on Twitter.
- **Location targeting** allows advertisers to target people based on region, city, metro or zip code.
- **Mobile audience targeting** allows advertisers to target people who use their mobile app.
- **Movie and TV show targeting** allows advertisers to target people based on movies and TV shows they have watched or are likely to watch.
- **Platform targeting** allows advertisers to target people who use a certain platform, such as iOS or Desktop, to access Twitter.
- **Retargeting campaign engager targeting** allows advertisers to target people based on prior engagement with (e.g., Tweeting, clicking, Retweeting, replying, liking, or viewing) their company.
- **Tailored audience (list) targeting** allows advertisers to reach specific people on Twitter by uploading lists, which contain personal information (email addresses, phone numbers, or Twitter handles) that are matched to Twitter users' accounts.
- **Tailored audience (web) targeting** allows advertisers to target people who have visited their website.

B Raw Data Example


```
[ {
  "ad" : {
    "adsUserData" : {
      "adImpressions" : {
        "impressions" : [ {
          "deviceInfo" : {
            "osType" : "Ios",
            "deviceId" : "##############",
            "deviceType" : "iPhone X"
          },
          "displayLocation" : "SearchTweets",
          "promotedTweetInfo" : {
            "tweetId" : "####################",
            "tweetText" : "RT @SpotifyBrands: Young people's digital lives are subtly shifting
                           culture. Discover more with our global trends report.",
            "urls" : [ ],
            "mediaUrls" : [ ]
          },
          "advertiserInfo" : {
            "advertiserName" : "Spotify",
            "screenName" : "@Spotify"
          },
          "matchedTargetingCriteria" : [ {
            "targetingType" : "Events",
            "targetingValue" : "Back to School 2019"
          }, {
            "targetingType" : "Age",
            "targetingValue" : "18 to 49"
          }, {
            "targetingType" : "Locations",
            "targetingValue" : "United States"
          }, {
            "targetingType" : "Platforms",
            "targetingValue" : "iOS"
          } ],
          "impressionTime" : "YYYY-MM-DD HH:MM:SS"
        } ]
      }
    }
  }
}, {
  "ad" : { ...
```

C Requests For Clarification of the Data Files We Made to Twitter Under GDPR

In this section, we describe the communications between a member of our research team and Twitter through which we attempted to confirm and clarify the meaning of fields present in users' data files downloaded from Twitter. This individual invoked their GDPR rights as an EU citizen as the basis for these requests and communications.

C.1 Initial Request

A member of the research team (an EU citizen) submitted the following request and series of follow-ups to both of the following channels:

- We sent the request below to Twitter Support via the Twitter Privacy Inquiries online form indicated in Twitter's Privacy Policy (`https://help.twitter.com/forms/privacy`). We sent this request on **June 26, 2019**.
- After we received an unsatisfactory response to the initial request (below, "Twitter's Initial Response") on July 7, 2019, we sent the same request to Twitter's Data Protection Officer via the online form indicated in Twitter's Privacy Policy (`https://twitter.ethicspointvp.com/custom/twitter/forms/data/form_data.asp`). We sent this request on **July 9, 2019**.
- We received a satisfactory response from Twitter Office of Data Protection on **November 15, 2019**.

The member of the research team included the following data they downloaded about their own Twitter account: the PDF file containing the list of "Similar audiences" and "Tailored audiences," as well as the "ad-impressions.js" file.

Hello,

I have downloaded my Twitter data from the settings page. The information displayed in this exported data or in various help center articles does not fully explain about how my personal data is being used for advertising purposes, and I ask for a number of clarifications:

1. I have downloaded my data, and in the zip file there is a file called "ad-impressions.js", which I have also submitted with this request as a courtesy. For each ad in this file, there is a field called "matchedTargetingCriteria", and I would like to understand how to interpret this information.
   (a) Does matchedTargetingCriteria represent the criteria that the advertiser chose to target this ad? If not, what does it represent?
   (b) If (a), does the matchedTargetingCriteria shown for a single ad represent ALL the criteria that the advertiser chose for that ad, or are there criteria that the advertiser chose when targeting this ad that are not reflected in the export?
   (c) Do all the matchedTargetingCriteria apply to me, e.g. if an ad is targeted to "Follower look-alikes: @Twitter, Keywords: Privacy, and Locations: United Kingdom", does that mean that I am seeing this ad because ALL of those criteria apply to me, or only a subset?
      i. If it is only a subset, how many and which of the advertiser's targeting criteria need to match a specific user to determine whether they receive the ad?
      ii. If it is only a subset, how can I determine which of the criteria used for targeting were matched to me?
   (d) According to `https://business.twitter.com/en/targeting/tailored-audiences.html`, there are three types of tailored audiences: lists, web, and mobile. I noticed that "Tailored audiences (lists)" and "Tailored audiences (web)" are reflected in the ad-impressions.js file, but are "Tailored audiences (mobile)" also used to target ads? Why does it not appear?
   (e) When an ad notes that targetingType "Tailored audiences (lists)" was used, does this only mean that I was on the list, or does this also include the "expanding the reach" feature (as explained here: `https://business.twitter.com/en/help/campaign-setup/campaign-targeting/tailored-audiences/TA-from-lists.html`) such that I was not on the list, but only similar to others that were included on the list?
2. The ad-impressions.js file says it contains "Promoted Tweets viewed by the account and associated metadata." This file appears to contain the ads shown in the past 90 days. Does Twitter retain data about ad impressions outside this window?
3. Additionally, I requested the list of "Similar audiences" and "Tailored audiences" from Twitter's settings page, which is sent as a PDF file via email that I have also attached to this request.
   (a) Is this file accurate for the entirety of a Twitter user's account, 90 days, or some other period of time?
   (b) Do the "Similar audiences" from this PDF file correspond to the same functionality as "expanding the reach" of a tailored audience, as described at `https://business.twitter.com/en/help/campaign-setup/campaign-targeting/tailored-audiences/TA-from-lists.html`?
   (c) When are the "Similar audiences" lists created? Does Twitter generate this automatically?

(d) Does the presence of an advertiser on the PDF's "Similar audience" list indicate that they explicitly chose to expand the reach of a tailored audience?

(e) Does an advertiser's inclusion on our "Similar audiences" list mean that I have seen an ad using that criteria? Or does it simply reflect all advertiser "audiences that are similar to tailored audiences" that I am included in?

Before reporting my concerns to the Information Commissioner's Office (ICO), I understand that I should give you the chance to respond. You can find guidance on your obligations under information rights legislation on the ICO's website (www.ico.org.uk) as well as information on their regulatory powers and the action they can take.

Please send a full response within one calendar month. If you cannot respond within that timescale, please tell me when you will be able to respond.

If there is anything you would like to discuss, please contact me on the following number [anonymized telephone number].

Yours faithfully,

[Name]

C.2 Twitter's Initial Response

Twitter responded with the following on **July 7, 2019**.

Hello,

We found a page in our help center that we think will help you out: (`https://help.twitter.com/en/managing-your-account/accessing-your-twitter-data` and `https://help.twitter.com/en/safety-and-security/privacy-controls-for-tailored-ads`)

If you've checked out that page and are still confused, write back to let us know more about where you're stuck. We'll do our best to help you out!

Thanks,

Twitter Support

C.3 Twitter's DPO's Response

As described at the beginning of this section, after we received this unsatisfactory response from Twitter Support ("Twitter's Initial Response") on **July 7, 2019**, we sent the same request to Twitter's Data Protection Officer on **July 9, 2019**. Under the GDPR, data controllers are obliged to respond within 30 days from receiving the request. After 31 days, on **August 9, 2019**, Twitter's DPO responded with the following.

Hello [Name],

Thank you for contacting us.

We are in the process of reviewing your inquiry. Due to its scope, however, we avail ourselves of the deadline extension of 60 days.

If you have any questions about this notice, please let us know.

Sincerely,

Twitter Office of Data Protection

C.4 Twitter's DPO's Detailed Response

On **November 15, 2019**, we received the following, detailed reply.

Hello [Name],

Thank you for your inquiry and patience.

With respect to the questions in your inquiry, we answer in turn below.

1. I have downloaded my data, and in the zip file there is a file called "ad-impressions.js", which I have also submitted with this request as a courtesy. For each ad in this file, there is a field called "matchedTargetingCriteria", and I would like to understand how to interpret this information.

   (a) Does matchedTargetingCriteria represent the criteria that the advertiser chose to target this ad? If not, what does it represent?

   When you download your Twitter data, an explanatory file is included. This file does indicate that it is the targeting criteria that is used to run the campaign.

(b) If (a), does the matchedTargetingCriteria shown for a single ad represent ALL the criteria that the advertiser chose for that ad, or are there criteria that the advertiser chose when targeting this ad that are not reflected in the export?

The information represents all of the targeting criteria for how the ad was served to a specific user, in this case @[Twitter handle].

(c) Do all the matchedTargetingCriteria apply to me, e.g. if an ad is targeted to "Follower look-alikes: @Twitter, Keywords: Privacy, and Locations: United Kingdom", does that mean that I am seeing this ad because ALL of those criteria apply to me, or only a subset?

Yes, a specific user will be targeted if all criteria match.

(i) If it is only a subset, how many and which of the advertiser's targeting criteria need to match a specific user to determine whether they receive the ad? (ii) If it is only a subset, how can I determine which of the criteria used for targeting were matched to me?

As mentioned above, they all apply to the user.

(d) According to https://business.twitter.com/en/targeting/tailored-audiences.html, there are three types of tailored audiences: lists, web, and mobile. I noticed that "Tailored audiences (lists)" and "Tailored audiences (web)" are reflected in the ad-impressions.js file, but are "Tailored audiences (mobile)" also used to target ads? Why does it not appear?

The list is a list of device IDs provided by the advertiser. A user may belong to "Tailored audiences (mobile)", but the response may not be human readable, so these are displayed under the targetingType "Unknown".

(e) When an ad notes that targetingType "Tailored audiences (lists)" was used, does this only mean that I was on the list, or does this also include the "expanding the reach" feature (as explained here: https://business.twitter.com/en/help/campaign-setup/campaign-targeting/tailored-audiences/TA-from-lists.html) such that I was not on the list, but only similar to others that were included on the list?

It means that the user is on a Tailored audiences list, not similar to others that were included on the list. It does not include the "expanding the reach" feature.

2. The ad-impressions.js file says it contains "Promoted Tweets viewed by the account and associated metadata." This file appears to contain the ads shown in the past 90 days. Does Twitter retain data about ad impressions outside this window?

   In accordance with our Privacy Policy, we do not retain data that is associated with a specific user ID past 90 days. We retain raw logs for 18 months, but they are stripped of the user ID after 90 days.

3. Additionally, I requested the list of "Similar audiences" and "Tailored audiences" from Twitter's settings page, which is sent as a PDF file via email that I have also attached to this request.

   (a) Is this file accurate for the entirety of a Twitter user's account, 90 days, or some other period of time?

   This is the latest snapshot which includes data no older than 7 days.

   (b) Do the "Similar audiences" from this PDF file correspond to the same functionality as "expanding the reach" of a tailored audience, as described at https://business.twitter.com/en/help/campaign-setup/campaign-targeting/tailored-audiences/TA-from-lists.html?

   Yes, they are the same.

   (c) When are the "Similar audiences" lists created? Does Twitter generate this automatically?

   The lists are automatically created within 24 hours after the audience list is uploaded by the advertiser.

   (d) Does the presence of an advertiser on the PDF's "Similar audience" list indicate that they explicitly chose to expand the reach of a tailored audience?

   Yes, we only expand it when advertisers ask us to do so.

   (e) Does an advertiser's inclusion on our "Similar audiences" list mean that I have seen an ad using that criteria? Or does it simply reflect all advertiser "audiences that are similar to tailored audiences" that I am included in?

   Yes, it reflects the advertiser audiences that are similar to tailored audiences in which you have been included.

Should you have any further questions, please let us know.
Sincerely,
Twitter Office of Data Protection

D Instructions Provided to Participants For Data Request (Part 1 of the study)

Below is the text we provided to explain to participants how to request their Twitter data. We included detailed and annotated screenshots highlighting each step of the process.

D.1 Consent Form

Study Title: Twitter Ad Transparency

DESCRIPTION: We are researchers at [redacted] doing research to better understand Twitter advertising transparency. In this survey, you will be asked about your experiences and opinions about Twitter. People who are age 18+ and live in the United States or United Kingdom are eligible to participate. Additionally, you must have an active Twitter account. Participation consists of two parts: first, a short 5-minute preliminary survey, and then the main survey, which should take about 35 minutes.

RISKS and BENEFITS: The risks to your participation in this online study are those associated with basic computer tasks, including boredom, fatigue, mild stress, or breach of confidentiality. The only benefit to you is the learning experience from participating in a research study. The benefit to society is the contribution to scientific knowledge.

COMPENSATION: Participants who complete all tasks will be compensated $7.86: $0.86 for Part 1 and $7.00 for Part 2.

CONFIDENTIALITY: No personally-identifiable information will be collected from you. Any reports and presentations about the findings from this study will not include your name or any other information that could identify you. In some cases, you might provide personal stories or beliefs that we might quote or paraphrase as part of our research findings – any personally identifying information will be removed to protect your privacy. We may share the data we collect in this study with other researchers doing future studies – if we share your data, we will not include information that could identify you.

SUBJECT'S RIGHTS: Your participation is voluntary. You may stop participating at any time by closing the browser window or the program to withdraw from the study.

[Additional content removed for anonymity]

Please indicate below, that you are at least 18 years old, have read and understand this consent form, and agree to participate in this online research study.

I am at least 18 years old. ◯ Yes ◯ No

I have read and understood this consent form. ◯ Yes ◯ No

I agree to participate in this research study. ◯ Yes ◯ No

D.2 Introduction and Instructions

Thank you for your participation in our study.

In Part 1 of this study (today), you will log into your Twitter account and request two data downloads. On the next page, you will be guided through the process of requesting your Twitter data.

NOTE: The information we collect in this study will not include your personal information. We will NOT ask for your Twitter username, messages, tweets, etc.

We are only interested in data about ads you have seen on Twitter. You will request your entire Twitter data archive today, but in Part 2 of this study, we will provide instructions for uploading only the data we need for our research.

There are two downloads that you need to request in this part of the study.

How to make the first request:

1) Log into Twitter: `https://twitter.com` (opens in a new tab).
2) Click on "More" at the bottom left. On narrower screens, it may only display the icon with three dots, without the word "More".
3) Click "Settings and privacy".
4) Click "Your Twitter data" at the bottom of the menu on the right side.
5) If prompted, enter your Twitter password.
6) Scroll to the bottom of the page. In the "Download your data" section, click "Request data" in the Twitter row.

If you do not see a button that says "Request data" where the red box appears above, this means you have already requested your data. Continue to the instructions below.

Twitter will email you when your download is ready. There is no need to do anything with this data until Part 2.

To verify that you have successfully requested your data, please copy the text immediately to the left of the "Retrieving data" button, and where the gray box appears in the screenshot below. Paste the text in the text box below.

No text where the gray box appears in the screenshot? You may have previously requested your Twitter data. Instead, please write out the two words on the button that appears instead of the "Retrieving data" button.

How to make the second request:

To make the second request, begin with the same first 5 steps.

1) Log into Twitter: `https://twitter.com` (opens in a new tab).
2) Click on "More" at the bottom left.
3) Click "Settings and privacy".
4) Click "Your Twitter data" at the bottom of the menu on the right side.
5) If prompted, enter your Twitter password.
6) Scroll to the bottom of the page. Now, click "Interests and ads data".
7) Click "Tailored Audiences".
8) Click "Request advertiser list".
9) On the pop-up, click "Request".

Twitter will email you when your download is ready. There is no need to do anything with this data until Part 2. That's it for the second request!

To verify that you have successfully requested your data, please copy the text shown where the gray box appears in the screenshot below. Paste the text in the text box below.

Thank you for making the data requests. For today's last task, please find the summary statistics shown in your Twitter settings on the "Interests and ads data" page.

Please enter the summary statistics into the fields below as numbered in the screenshot. (Fields 1, 2, 3, 4)

D.3 Conclusions

Thank you for completing Part 1 of our study.

It may take a few hours or days until your Twitter data is ready to download.

You will be invited back for Part 2 via Prolific in a few days. In Part 2, you will be given instructions on how to download your Twitter data and upload it to the study.

Part 2 will be a survey that takes 35 minutes to complete.

(Optional) Do you have any final thoughts or comments?

E Survey Instrument (Part 2 of the study)

This section provides the survey instrument for the main part of our user study.

E.1 Introduction and General Questions

Thank you for your participation in our study. This survey will take about 35 minutes.

This survey has 4 sections. The first section will ask a few general questions about your data and Twitter.

If a company writes in their privacy policy that "we do not sell your data," what does that mean to you? In your explanation, please include at least one example of a specific thing you think they would not be allowed to do.

Please rate your agreement with the following statement: I believe that Twitter sells my data. ◯ Strongly agree ◯ Agree ◯ Somewhat agree ◯ Neither agree nor disagree ◯ Somewhat disagree ◯ Disagree

E.2 Companies

This is the 2nd section (of 4).

In this section, we use the data you uploaded from your own Twitter account. You will be asked about advertisers and up to 4 different advertising methods on Twitter.

The list below shows some companies that showed you an ad on Twitter in the last 3 months.

Please select all of the companies, if any, you remember seeing ads from.

- None of the below
- *[Company 1]*
- *[Company 2]*
- *[Company 3]*
- *[Company 4]*
- *[Company 5]*
- *[Company 6]*
- *[Company 7]*
- *[Company 8]*
- *[Company 9]*
- *[Company 10]*

E.3 Section 1 (Targeting Types)

[We repeated this section 4 times for a random selection of 4 *[targeting types]* (e.g., "keywords") and associated specific *[instances]* of that type (e.g., "my cat is my best friend") from the participant's Twitter data. We first asked about the targeting type in the abstract, then about a specific instance from the participant's Twitter data, and then about the targeting type more generally with a selection of frequent and infrequent instances of that type from the participant's Twitter data.]

E.3.1 Abstract

What does the term **[targeting type]** in the context of online advertising mean to you?

If you have never heard this term before, please write your best guess.

This next section is about *[targeting type]*

[targeting type] definition of targeting type

Prior to this survey, I would have expected that advertisers currently target ads on Twitter using *[targeting type]*. ◯ Strongly agree ◯ Agree ◯ Somewhat agree ◯ Neither agree nor disagree ◯ Somewhat disagree ◯ Disagree

E.3.2 Specific

On this page, we will give you a specific example of *[targeting type]* from your Twitter data.

According to your Twitter data, you are interested in *[instance]*

Please rate your agreement with the following statements:

I can think of a reason why I Twitter would conclude that I am [interested in, located in or around, would be added to a list of mobile app users by, etc.] *[instance]*. ◯ Strongly agree ◯ Agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree

Being [interested in, a speaker of, in the age group, etc.] *[instance]* describes me accurately. ◯ Strongly agree ◯ Agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree

Assume the number of ads you see doesn't change.

I want some of the ads I see to be chosen for me based on being interested in *[instance]*. ◯ Strongly agree ◯ Agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree

I am comfortable with Twitter allowing advertisers to target me based on being interested in *[instance]*. ◯ Strongly agree ◯ Agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree

E.3.3 General

Overall, in the last three months, advertisers have targeted up to *[#]* ads using *[targeting type]*.

In two sentences, please describe your initial reaction to the data above.

This section will ask you to consider how you feel about advertisers using *[targeting type]* in general. Please rate your agreement with the following statements:

Assume the number of ads you see doesn't change.

I want some of the ads I see on Twitter to be chosen for me using *[targeting type]*. ◯ Strongly agree ◯ Agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree

I am comfortable with *[targeting type]* being used to choose ads for me. ◯ Strongly agree ◯ Agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree

I believe it is fair that Twitter allows advertisers to choose ads for me using *[targeting type]*. ◯ Strongly agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree

Please explain your answer to the previous question. If you believe it is fair, why? If you do not believe it is fair, why not?

E.4 Section 2 (Ad Explanations)

[We repeated this section 6 times for the six ad explanations in randomized order. Each was associated with an ad that the participant had been shown according to their Twitter data, alongside the matched targeting criteria for that ad.]

This is the 3rd section (of 4).

This section will ask for your opinions about potential explanations for why you received a particular ad on Twitter.

To your knowledge, does Twitter have a feature that explains why you received a particular ad? ◯ Yes ◯ No ◯ Don't Know

Imagine a Twitter feature that explains how a particular ad was chosen for you.

In this next section, you will see up to 6 different ads that Twitter has shown you before on its platform, each followed by a different explanation. Then, you will answer questions about what you thought of each ad explanation.

What was the most memorable part of this ad explanation?

What information, if any, did you feel was missing from this ad explanation?

I think this ad explanation shows me all of the information used to target the ad to me. ◯ Yes ◯ No ◯ Don't Know

I feel that this ad explanation was useful. ◯ Yes ◯ No ◯ Don't Know

I feel that this ad explanation gave me enough information to understand how the ad was chosen for me. ◯ Strongly agree ◯ Agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree

I would want an ad explanation similar to this one for all ads I see on Twitter. ◯ Strongly agree ◯ Agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree

Seeing this ad explanation made me more concerned about my online privacy. ◯ Strongly agree ◯ Agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree

Seeing this ad explanation increased my trust in the advertiser who displayed this ad. ◯ Strongly agree ◯ Agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree

Do you have any additional comments about this ad explanation?

E.5 General Opinions About Ad Explanations

If an ad explanation on Twitter did not include all reasons an ad was shown to you, which reasons would be most important for you to see?

For your reference, the previous ad explanations that you've seen in this study will appear below:

Please describe your ideal explanation for ads on Twitter. You are not limited to the things you have seen in this study. Feel free to think big!

E.6 Demographics

This is the 4th section (of 4). Almost done!

In this section, you will be asked about your Twitter usage and demographics.

Please rate your agreement with the following statement: I believe that Twitter sells my data. ◯ Strongly agree ◯ Agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree

Please explain your answer to the previous question. If you believe Twitter sells your data, why? If you believe Twitter does not sell your data, why not?

What month and year did you join Twitter?

On average, about how many hours do you spend on Twitter each day? ◯ Less than 1 hour ◯ 1-2 hours ◯ 2-4 hours ◯ 4-6 hours ◯ More than 6 hours

Have you ever gone to your Twitter account's settings to look at or make changes to your advertising preferences? ◯ Yes ◯ No ◯ Don't know

Did you look at any of the Twitter files you requested in Part 1 of this study before beginning Part 2? ◯ Yes ◯ No ◯ Don't Know

What is your gender? ◯ Woman ◯ Man ◯ Non-binary ◯ Prefer to self-describe ◯ Prefer not to say

What is your age? ◯ 18-24 ◯ 25-34 ◯ 35-44 ◯ 45-54 ◯ 55-64 ◯ 65 or older ◯ Prefer not to say

What is the highest degree or level of school you have completed? ◯ Some high school ◯ High school ◯ Some college ◯ Trade, technical, or vocational training ◯ Associate's degree ◯ Bachelor's degree ◯ Master's degree ◯ Professional degree ◯ Doctorate ◯ Prefer not to say

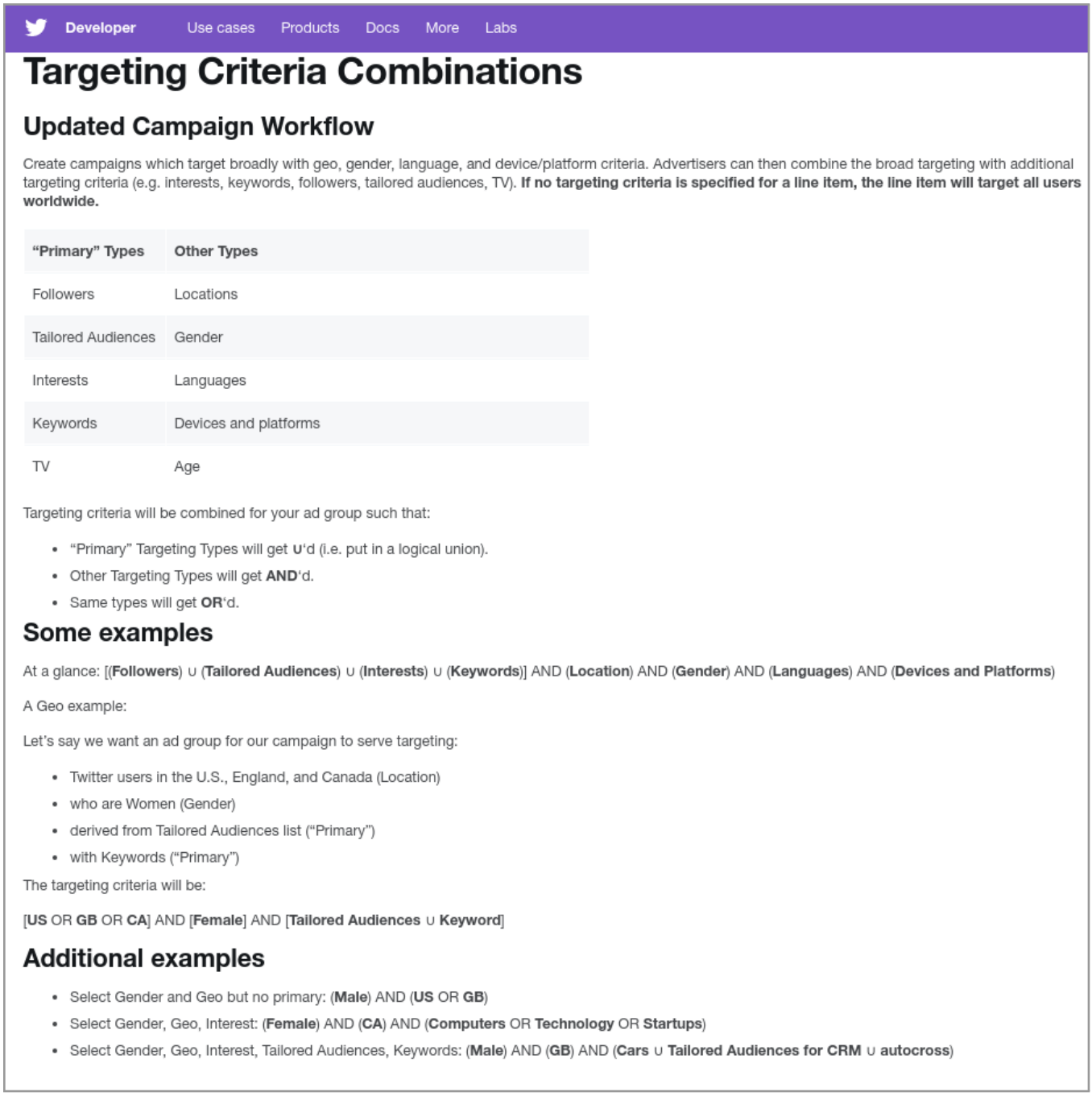
Which of the following best describes your educational background or job field? ◯ I have an education in, or work in, the field of computer science, engineering, or IT. ◯ I do not have an education in, or work in, the field of computer science, engineering, or IT. ◯ Prefer not to say

What is your annual household income? ◯ Less than $20,000 ◯ $20,000 to $49,999 ◯ $50,000 to $99,999 ◯ $100,000 to $249,999 ◯ Over $250,000 ◯ Prefer not to say

When people work on tasks, they are sometimes in situations that can be distracting. How distracted were you while completing this survey? ◯ Not distracted at all ◯ Somewhat distracted ◯ Very distracted

(Optional) Do you have any final thoughts or comments?

F Twitter's Explanation About Combining Targeting Criteria



Developer Use cases Products Docs More Labs

Targeting Criteria Combinations

Updated Campaign Workflow

Create campaigns which target broadly with geo, gender, language, and device/platform criteria. Advertisers can then combine the broad targeting with additional targeting criteria (e.g. interests, keywords, followers, tailored audiences, TV). **If no targeting criteria is specified for a line item, the line item will target all users worldwide.**

"Primary" Types	Other Types
Followers	Locations
Tailored Audiences	Gender
Interests	Languages
Keywords	Devices and platforms
TV	Age

Targeting criteria will be combined for your ad group such that:

- "Primary" Targeting Types will get **U**'d (i.e. put in a logical union).
- Other Targeting Types will get **AND**'d.
- Same types will get **OR**'d.

Some examples

At a glance: [(**Followers**) ∪ (**Tailored Audiences**) ∪ (**Interests**) ∪ (**Keywords**)] AND (**Location**) AND (**Gender**) AND (**Languages**) AND (**Devices and Platforms**)

A Geo example:

Let's say we want an ad group for our campaign to serve targeting:

- Twitter users in the U.S., England, and Canada (Location)
- who are Women (Gender)
- derived from Tailored Audiences list ("Primary")
- with Keywords ("Primary")

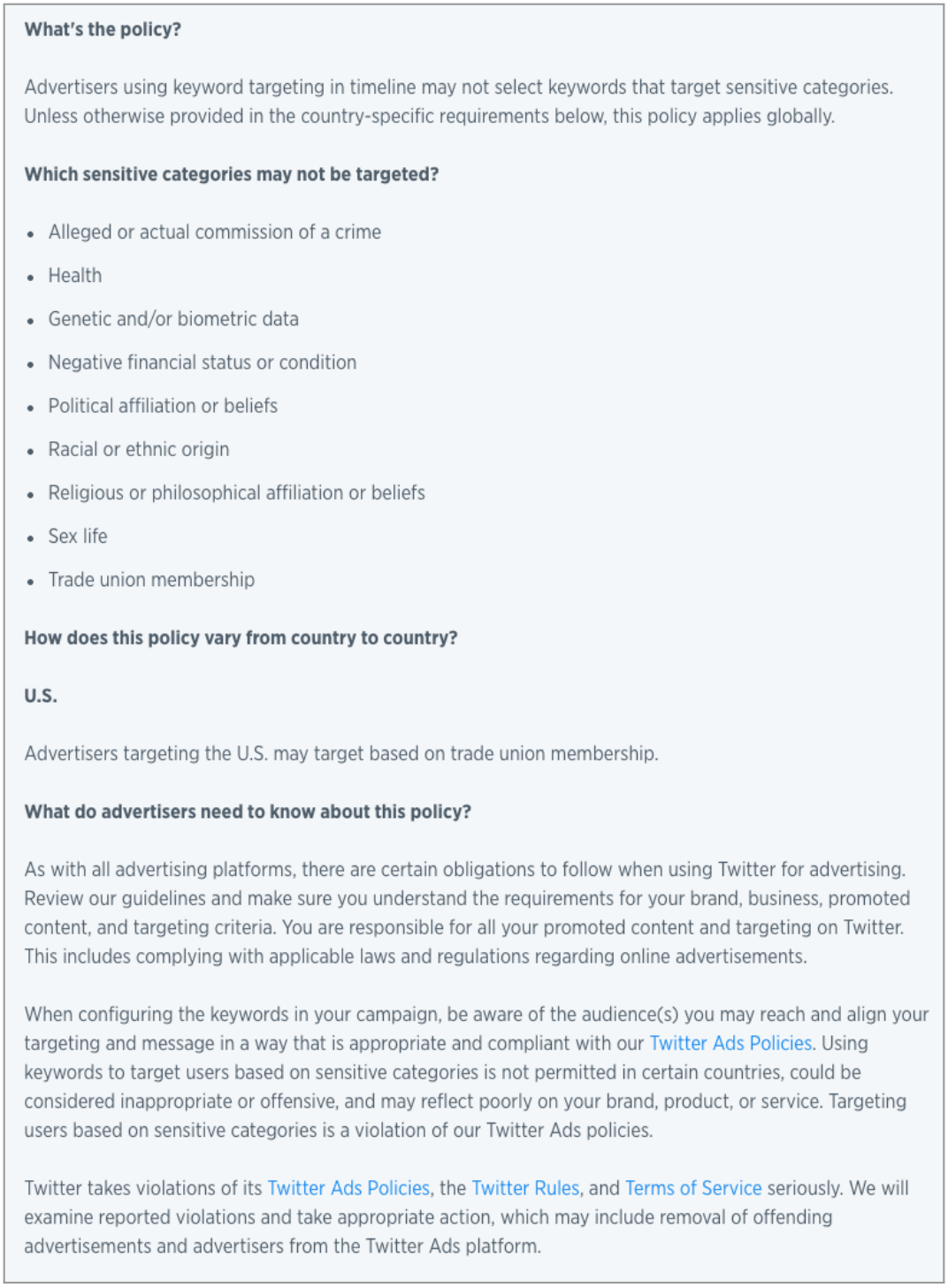
The targeting criteria will be:

[**US** OR **GB** OR **CA**] AND [**Female**] AND [**Tailored Audiences** ∪ **Keyword**]

Additional examples

- Select Gender and Geo but no primary: (**Male**) AND (**US** OR **GB**)
- Select Gender, Geo, Interest: (**Female**) AND (**CA**) AND (**Computers** OR **Technology** OR **Startups**)
- Select Gender, Geo, Interest, Tailored Audiences, Keywords: (**Male**) AND (**GB**) AND (**Cars** ∪ **Tailored Audiences for CRM** ∪ **autocross**)

Figure 6: Twitter's explanation on their developers page about how targeting criteria are combined.



What's the policy?

Advertisers using keyword targeting in timeline may not select keywords that target sensitive categories. Unless otherwise provided in the country-specific requirements below, this policy applies globally.

Which sensitive categories may not be targeted?

- Alleged or actual commission of a crime
- Health
- Genetic and/or biometric data
- Negative financial status or condition
- Political affiliation or beliefs
- Racial or ethnic origin
- Religious or philosophical affiliation or beliefs
- Sex life
- Trade union membership

How does this policy vary from country to country?

U.S.

Advertisers targeting the U.S. may target based on trade union membership.

What do advertisers need to know about this policy?

As with all advertising platforms, there are certain obligations to follow when using Twitter for advertising. Review our guidelines and make sure you understand the requirements for your brand, business, promoted content, and targeting criteria. You are responsible for all your promoted content and targeting on Twitter. This includes complying with applicable laws and regulations regarding online advertisements.

When configuring the keywords in your campaign, be aware of the audience(s) you may reach and align your targeting and message in a way that is appropriate and compliant with our Twitter Ads Policies. Using keywords to target users based on sensitive categories is not permitted in certain countries, could be considered inappropriate or offensive, and may reflect poorly on your brand, product, or service. Targeting users based on sensitive categories is a violation of our Twitter Ads policies.

Twitter takes violations of its Twitter Ads Policies, the Twitter Rules, and Terms of Service seriously. We will examine reported violations and take appropriate action, which may include removal of offending advertisements and advertisers from the Twitter Ads platform.

Figure 7: Twitter's policy for prohibiting targeting based on sensitive categories from `https://business.twitter.com/en/help/ads-policies/other-policy-requirements/policies-for-keyword-targeting.html`.

G Targeting Types Seen by Participants

Table 4: The number of participants who saw each targeting type in the survey.

Targeting Type	# Participants	Targeting Type	# Participants
Location	79	Tailored web	58
Age	77	Tailored lists	56
Lookalikes	67	Conversation	51
Language	66	Mobile	45
Platform	64	Event	40
Gender	63	Movie/TV	40
Keyword	63	Retargeting	37
Interest	59	Behavior	33

H Regression Tables

In this section, we present tables depicting the full results of our mixed-effects ordinal logistic regression models analyzing targeting type data from Part 1 of the user study.

Table 5: Mixed-effect ordinal logistic regression model of how participants' agreement responding to **General: Fair** ("I believe it is fair that Twitter allows advertisers to choose ads for me using *targeting type*") varied by targeting type.

Factor	Baseline	Odds Ratio	β	Std. Error	z	p
Type: Age	Interest	0.640	-0.446	0.389	-1.146	.252
Type: Behavior	Interest	0.155	-1.863	0.475	-3.924	**<.001**
Type: Conversation	Interest	0.362	-1.017	0.431	-2.360	**.018**
Type: Event	Interest	0.401	-0.914	0.464	-1.969	**.049**
Type: Gender	Interest	0.350	-1.050	0.408	-2.576	**.010**
Type: Keyword	Interest	0.310	-1.171	0.404	-2.898	**0.004**
Type: Language	Interest	4.480	1.450	0.429	3.494	**<.001**
Type: Location	Interest	0.262	-1.339	0.388	-3.449	**<.001**
Type: Lookalikes	Interest	0.218	-1.522	0.396	-3.845	**<.001**
Type: Mobile	Interest	0.128	-2.057	0.441	-4.668	**<.001**
Type: Movie/TV	Interest	0.386	-0.952	0.460	-2.068	**.039**
Type: Platform	Interest	0.497	-0.699	0.410	-1.706	.088
Type: Retargeting	Interest	0.409	-0.895	0.458	-1.954	.051
Type: Tailored lists	Interest	0.061	-2.802	0.430	-6.511	**<.001**
Type: Tailored web	Interest	0.120	-2.118	0.418	-5.064	**<.001**

Table 6: Mixed-effect ordinal logistic regression model of how participants' agreement responding to **General: Comfortable** ("I am comfortable with *targeting type* being used to choose ads for me") varied by targeting type.

Factor	Baseline	Odds Ratio	β	Std. Error	z	p
Type: Age	Interest	0.743	-0.297	0.394	-0.755	.451
Type: Behavior	Interest	0.162	-1.817	0.489	-3.714	**<.001**
Type: Conversation	Interest	0.326	-1.121	0.438	-2.556	**.011**
Type: Event	Interest	0.439	-0.823	0.451	-1.823	.068
Type: Gender	Interest	0.525	-0.644	0.411	-1.569	.117
Type: Keyword	Interest	0.290	-1.239	0.404	-3.071	**.002**
Type: Language	Interest	5.411	1.689	0.440	3.838	**<.001**
Type: Location	Interest	0.305	-1.188	0.391	-3.039	**.002**
Type: Lookalikes	Interest	0.251	-1.383	0.396	-3.495	**<.001**
Type: Mobile	Interest	0.110	-2.204	0.450	-4.897	**<.001**
Type: Movie/TV	Interest	0.433	-0.838	0.457	-1.836	.066
Type: Platform	Interest	0.399	-0.918	0.408	-2.250	**.024**
Type: Retargeting	Interest	0.286	-1.253	0.459	-2.730	**.006**
Type: Tailored lists	Interest	0.063	-2.758	0.434	-6.363	**<.001**
Type: Tailored web	Interest	0.158	-1.846	0.422	-4.375	**<.001**

Table 7: Mixed-effect ordinal logistic regression model of how participants' agreement responding to **General: Want** ("Assume the number of ads you see doesn't change. I want some of the ads I see on Twitter to be chosen for me using *targeting type*") varied by targeting type.

Factor	Baseline	Odds Ratio	β	Std. Error	z	p
Type: Age	Interest	0.304	-1.190	0.377	-3.157	**.002**
Type: Behavior	Interest	0.152	-1.885	0.477	-3.955	**<.001**
Type: Conversation	Interest	0.390	-0.941	0.421	-2.235	**.025**
Type: Event	Interest	0.403	-0.910	0.443	-2.055	**.040**
Type: Gender	Interest	0.301	-1.202	0.398	-3.024	**.002**
Type: Keyword	Interest	0.221	-1.511	0.392	-3.852	**<.001**
Type: Language	Interest	3.318	1.200	0.415	2.893	**.004**
Type: Location	Interest	0.254	-1.369	0.377	-3.633	**<.001**
Type: Lookalikes	Interest	0.200	-1.611	0.382	-4.217	**<.001**
Type: Mobile	Interest	0.099	-2.312	0.435	-5.311	**<.001**
Type: Movie/TV	Interest	0.406	-0.900	0.444	-2.028	**.043**
Type: Platform	Interest	0.210	-1.561	0.393	-3.975	**<.001**
Type: Retargeting	Interest	0.273	-1.299	0.445	-2.917	**.004**
Type: Tailored lists	Interest	0.061	-2.803	0.420	-6.680	**<.001**
Type: Tailored web	Interest	0.114	-2.171	0.406	-5.345	**<.001**

Table 8: Mixed-effect ordinal logistic regression model of how participants' agreement responding to **Specific: Comfortable** ("I am comfortable with *specific example of targeting type* being used to choose ads for me") varied by targeting type.

Factor	Baseline	Odds Ratio	β	Std. Error	z	p
Type: Age	Interest	0.745	-0.295	0.354	-0.833	.405
Type: Behavior	Interest	0.266	-1.324	0.447	-2.963	**.003**
Type: Conversation	Interest	0.276	-1.287	0.401	-3.212	**.001**
Type: Event	Interest	0.271	-1.307	0.434	-3.011	**.003**
Type: Gender	Interest	0.686	-0.377	0.381	-0.990	.322
Type: Keyword	Interest	0.306	-1.183	0.372	-3.176	**.001**
Type: Language	Interest	6.017	1.795	0.401	4.477	**<.001**
Type: Location	Interest	0.394	-0.932	0.359	-2.598	**.009**
Type: Lookalikes	Interest	0.356	-1.033	0.359	-2.877	**.004**
Type: Mobile	Interest	0.081	-2.511	0.425	-5.909	**<.001**
Type: Movie/TV	Interest	0.284	-1.258	0.433	-2.903	**.004**
Type: Platform	Interest	0.575	-0.553	0.373	-1.482	.138
Type: Retargeting	Interest	0.188	-1.673	0.434	-3.852	**<.001**
Type: Tailored lists	Interest	0.121	-2.115	0.406	-5.210	**<.001**
Type: Tailored web	Interest	0.105	-2.255	0.399	-5.655	**<.001**

Table 9: Mixed-effect ordinal logistic regression model of how participants' agreement responding to **Specific: Want** ("Assume the number of ads you see doesn't change. I want some of the ads I see on Twitter to be chosen for me using *specific example of targeting type*") varied by targeting type.

Factor	Baseline	Odds Ratio	β	Std. Error	z	p
Type: Age	Interest	1.315	0.274	0.340	0.806	.420
Type: Behavior	Interest	0.746	-0.293	0.429	-0.683	.495
Type: Conversation	Interest	0.348	-1.054	0.387	-2.725	**.006**
Type: Event	Interest	0.237	-1.439	0.428	-3.362	**<.001**
Type: Gender	Interest	1.016	0.016	0.359	0.044	.965
Type: Keyword	Interest	0.401	-0.913	0.364	-2.510	**.012**
Type: Language	Interest	8.434	2.132	0.380	5.619	**<.001**
Type: Location	Interest	0.788	-0.239	0.347	-0.688	.491
Type: Lookalikes	Interest	0.436	-0.831	0.352	-2.358	**.018**
Type: Mobile	Interest	0.064	-2.750	0.433	-6.357	**<.001**
Type: Movie/TV	Interest	0.331	-1.106	0.418	-2.644	**.008**
Type: Platform	Interest	0.965	-0.036	0.355	-0.102	.919
Type: Retargeting	Interest	0.175	-1.742	0.435	-4.002	**<.001**
Type: Tailored lists	Interest	0.237	-1.440	0.379	-3.797	**<.001**
Type: Tailored web	Interest	0.138	-1.984	0.389	-5.100	**<.001**

Table 10: Mixed-effect ordinal logistic regression model of how participants' agreement responding to **Specific: Accurate** ("*Specific example of targeting type* describes me accurately") varied by targeting type.

Factor	Baseline	Odds Ratio	β	Std. Error	z	p
Type: Age	Interest	2.939	1.078	0.339	3.182	**.001**
Type: Behavior	Interest	0.477	-0.740	0.384	-1.926	.054
Type: Conversation	Interest	0.437	-0.827	0.356	-2.326	**.020**
Type: Event	Interest	0.348	-1.056	0.392	-2.696	**.007**
Type: Gender	Interest	5.158	1.641	0.380	4.315	**<.001**
Type: Keyword	Interest	0.428	-0.848	0.333	-2.548	**.011**
Type: Language	Interest	9.691	2.271	0.392	5.789	**<.001**
Type: Location	Interest	1.880	0.631	0.334	1.889	.059
Type: Lookalikes	Interest	0.492	-0.710	0.324	-2.192	**.028**
Type: Mobile	Interest	0.104	-2.266	0.399	-5.684	**<.001**
Type: Movie/TV	Interest	0.501	-0.692	0.401	-1.724	.085
Type: Platform	Interest	2.922	1.072	0.342	3.138	**.002**
Type: Retargeting	Interest	0.165	-1.801	0.406	-4.434	**<.001**
Type: Tailored lists	Interest	0.240	-1.428	0.347	-4.118	**<.001**
Type: Tailored web	Interest	0.146	-1.923	0.364	-5.282	**<.001**

Table 11: Mixed-effect ordinal logistic regression model of how participants' agreement responding to **Specific: Reason** ("I can think of a reason why *[phrase explaining that Twitter would conclude that I am similar to, or described by, specific example of targeting type]*") varied by targeting type.

Factor	Baseline	Odds Ratio	β	Std. Error	z	p
Type: Age	Interest	1.696	0.528	0.329	1.604	.109
Type: Behavior	Interest	0.434	-0.835	0.396	-2.108	**.035**
Type: Conversation	Interest	0.406	-0.902	0.363	-2.481	**.013**
Type: Event	Interest	0.327	-1.117	0.403	-2.772	**.006**
Type: Gender	Interest	2.253	0.812	0.362	2.243	**.025**
Type: Keyword	Interest	0.458	-0.782	0.354	-2.210	**.027**
Type: Language	Interest	8.172	2.101	0.396	5.308	**<.001**
Type: Location	Interest	1.742	0.555	0.333	1.665	.096
Type: Lookalikes	Interest	0.523	-0.649	0.332	-1.956	.050
Type: Mobile	Interest	0.203	-1.595	0.395	-4.040	**<.001**
Type: Movie/TV	Interest	0.582	-0.542	0.401	-1.352	.176
Type: Platform	Interest	3.055	1.117	0.352	3.169	**.002**
Type: Retargeting	Interest	0.306	-1.184	0.410	-2.892	**.004**
Type: Tailored lists	Interest	0.287	-1.248	0.360	-3.462	**<.001**
Type: Tailored web	Interest	0.248	-1.393	0.367	-3.795	**<.001**