Is Contextual Advertising Safe? Analyzing Systemic Risks with Ads on YouTube

Anonymous Author(s)

Abstract

Contextual advertising is seeing a resurgence in popularity as a privacy-preserving alternative to behavioral advertising. While often regarded as a coarse-grained approach, advances in AI-driven content analysis have transformed it into a highly granular form of targeting. This work examines the safety risks of contextual targeting through a two-part empirical study, analyzing its potential to enable targeting of audiences with sensitive attributes and exposing users to harmful or exploitative ads. In controlled ad experiments, we show that advertisers can target audiences defined by sensitive attributes (e.g., religious belief, mental health condition, and political ideology) by strategically selecting contextual placements—circumventing policies that prohibit such targeting through behavioral signals. To understand how this risk manifests in practice, we develop an automated measurement framework to collect contextual ads delivered on high-risk content environments, focusing on conspiracy videos. We find that contextual ads are highly prevalent in these environments, disproportionately deliver sensitive categories (e.g., alternative health, religion, and political), and lack meaningful transparency. We argue that contextual ad systems require deeper empirical scrutiny and robust transparency mechanisms to prevent exploitation and abuse, and regulators should extend behavioral advertising risk principles to the contextual domain.

Keywords

Contextual advertising, online advertising transparency, vulnerable audiences, YouTube

1 Introduction

Behavioral targeting relies on large-scale user tracking to build detailed profiles that inform when, where, and to whom ads are shown. This model has faced growing criticism for its intrusiveness and lack of transparency [8, 10, 11, 34, 39], fueling widespread user distrust and driving adoption of privacy-enhancing tools such as Virtual Private Networks (VPNs), ad blockers, and privacy-first browsers. At the same time, major browsers like Safari and Firefox have phased out third-party cookies by default [32, 48], directly undermining the infrastructure that powers behavioral tracking. Furthermore, regulatory frameworks like the General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), and the EU's Digital Services Act (DSA) impose legal limits on the collection, sharing, and use of personal data for targeting [14, 16, 41].

Within this shifting ecosystem, **contextual advertising**—which matches ads to the content a user is currently viewing rather than their historical behavior profile—has reemerged as a leading alternative [7, 17, 42]. Because it is regarded as an ethical and privacy-preserving approach, and because it does not rely on third-party cookies—now being phased out—its adoption has accelerated in regulation-heavy environments such as the EU, where it now accounts for a growing share of digital advertising [17, 42].

Advertisers can manually define contextual targeting parameters by selecting topics, keywords, or specific placements (e.g., particular channels, blogs, or videos) where their ads should appear. Increasingly, however, these decisions are also shaped by platform-side optimization systems that apply artificial intelligence techniques—such as semantic and sentiment analysis, clustering, and reinforcement learning—to infer relevant contexts automatically [25]. As a result, contextual advertising can enable precise audience targeting without relying on individual profile data. For instance, an advertiser could place predatory offers or pseudoscientific "miracle cures" adjacent to videos or blogs about health anxieties or personal hardship, or be algorithmically matched to such material based on semantic similarity—effectively reaching users within vulnerable segments based on contextual cues alone.

This growing algorithmic complexity raise important safety and accountability concerns. Yet, contextual targeting is often portrayed as a coarse or "safer" alternative to behavioral targeting, and has received little to no scrutiny in the literature. One notable exception is Medjkoune et al. [33], who demonstrate that contextual placements can be exploited to circumvent protections intended for children, exposing a policy loophole enabled by this form of targeting. In this paper, we examine the extent to which contextual advertising can be used to target users based on sensitive traits and whether its deployment in high-risk content environments introduces safety risks. Specifically, we ask:

- **RQ1:** *Can advertisers reach audiences defined by sensitive attributes (such as religion or political ideology) by using content as a proxy for those traits?*
- **RQ2:** *How prevalent is contextual advertising on sensitive or fringe content, and how does the surrounding content influence the nature of the ads served?*
- **RQ3:** *Does contextual advertising provide a level of transparency sufficient for meaningful auditing and user understanding?*

To address these questions, we conduct a two-part study of contextual advertising on YouTube, one of the largest video platforms and a central actor in the digital advertising ecosystem [9].

In the first part of our study (**RQ1**), we conduct ad experiments using curated YouTube placements as proxies for sensitive traits such as religion, mental health condition, personal hardship, and political ideology (§ 3). We find that **advertisers can cheaply and effectively use contextual placements to reach sensitive audiences at scale, despite policies prohibiting such targeting through behavioral advertising**. This exposes a policy loophole with significant potential for abuse, including political manipulation and predatory commercial advertising.

Conference acronym 'XX, UAE
2026. ACM ISBN 978-1-4503-XXXX-X/2026/06
https://doi.org/XXXXXXX.XXXXXXX

The second part of our study (**RQ2**) examines the prevalence of contextual advertising on sensitive content, using fringe and conspiratorial videos on the platform as a case study. To this end, we constructed a curated corpus of conspiracy videos reflecting active misinformation narratives and developed an automated measurement framework to capture ads served on this content, together with the platform-provided explanations of their purported targeting rationale. The framework deploys multiple simulated user personas, systematically exposing them to both conspiracy and control videos (§ 4). We find that **contextual advertising plays a central role in sustaining the monetization of conspiracy content**—30% of ads in our dataset included a contextual targeting rationale—and **disproportionately delivers sensitive ad categories** such as *alternative health* (1.37% vs. 0.38%; $p < 0.001$), *religion* (8.08% vs. 6.01%; $p < 0.01$), and *political content*(6.88% vs. 1.42%; p<0.001) when compared to control videos.

Furthermore, **we identify significant gaps in contextual advertising transparency** on YouTube (§ 5; **RQ3**). The explanations provided to clarify why an ad was shown, including for political messaging, refer to contextual targeting only in vague terms, making it impossible to discern the specific contextual cues driving ad delivery or to determine whether placements are set by advertisers or by opaque platform-level optimization systems. Additionally, users who opt out of behavioral targeting receive even less disclosure and are exposed to contextual ads without any indication that this targeting rationale was applied.

Taken together, our results show that contextual targeting can be just as opaque and susceptible to abuse as behavioral advertising. While it avoids direct profiling, it still allows advertisers—and ad platforms—to shape who sees what, in ways that are difficult to audit or contest. We argue that contextual advertising demands the same degree of empirical scrutiny, regulatory oversight, and transparency expectations that behavioral advertising has attracted.

2 Background

Advertisements on YouTube are served through the Google Ads platform. To display their ads, advertisers can choose from a set of targeting options, such as demographics, locations, interest groups (such as affinity and in-market audiences), customer lists, and content-based targeting, i.e., contextual advertising [19].

Behavioral advertising on YouTube. Behavioral advertising—also referred to as *personalized*, *profiling-based*, or *interest-based advertising*—relies on tracking users' online activity to build detailed consumer segments. On YouTube, this involves aggregating data from watch history, search queries, and activity across Google services (and potentially on 3rd party websites) to infer users' interests, demographics, and intent. Advertisers can then target these inferred segments (e.g., "fitness enthusiasts," or "in-market for financial services") through Google Ads.

Contextual advertising on YouTube. The mechanism functions by delivering ads based solely on the real-time content signal (i.e., the specific YouTube video or channel the user is currently watching), not on inferred user attributes. When opting for contextual targeting, advertisers have three targeting methods:

Content keywords: Advertisers may supply a list of keywords they deem relevant to their campaign. Google uses these keywords to match ads to content containing those terms [19].
Content topics: Ads can be targeted to a broad range of themes. Google automatically analyzes videos to determine the dominant topics and match them to advertiser-selected categories [19].
Placements: Advertisers can directly specify any set of YouTube videos or channels where they want their ads to appear. This method, also referred to as *managed placements*, bypasses the need to analyze and match content with keywords or topics [19].

Algorithmic advertising on YouTube. Advertisers can also delegate complete audience selection to Google's automated ad delivery systems (selecting only the location), allowing the algorithm to dynamically adjust targeting. In this case, Google's algorithms may autonomously decide which contextual signals (topics, keywords, or placements) or behavioral profiles match the ads.

Ad transparency on YouTube. Ad transparency on YouTube is facilitated through both user-facing and public-facing tools:

User-facing tools: YouTube provides users with two main transparency tools: *"Why you're seeing this ad?"* information panel (see Appendix 2) and *"My Ad Center"* [43]. These tools are intended to disclose information about why a specific ad was shown and the advertiser's identity (mainly name and location).
Public-facing tools: Google's *Ads Transparency Center* is a platform designed to provide visibility into ads running across its platforms. It allows users—and auditors—to search for specific advertisers, view the ads they have served, and access information such as the geographic and temporal reach of ads.

3 Mechanisms for reaching vulnerable audiences

Sensitive attributes in personalized advertising refer to categories of personal data that are considered particularly private or potentially exploitative if used for targeting (e.g., religion or political affiliation). To protect users from discrimination and manipulation, both platforms and regulators place strict restrictions on the use of such attributes in behavioral advertising. For instance, the GDPR classifies these as "special categories" of data, which cannot be used for advertising except in rare cases and only with explicit consent[1]. The DSA goes further, outright prohibiting platforms from targeting ads based on such data[2]. Accordingly, platforms such as Google Ads restrict advertisers from directly targeting these categories through behavioral targeting [19].

In this section, we present an empirical investigation into whether these regulatory and platform-level protections can be *circumvented* through the strategic use of contextual advertising. We hypothesize that advertisers can effectively reach audiences associated with "special categories" by selectively curating placements on YouTube

[1]"Article 4 (1) Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited [...]" [15]
[2]Article 26 (3)"Providers of online platforms shall not present advertisements to recipients of the service based on profiling as defined in Article 4, point (4), of Regulation (EU) 2016/679 using special categories of personal data referred to in Article 9(1) of Regulation (EU) 2016/679." [13]

videos or channels whose content serves as a reliable *proxy* for sensitive traits. We then conduct a series of controlled ad delivery experiments to validate this circumvention mechanism.

3.1 Identifying proxies for sensitive attributes

We focus on several categories of sensitive attributes: *political ideology*, *belief systems and worldviews*, *religious belief*, *health condition*, *vulnerable populations*, and *sexual or gender identity*, categories referenced in Article 9 (1) of the GDPR[1]. Within each category, we selected specific traits to serve as proof-of-concept examples of how contextual placements can be curated to proxy for sensitive personal characteristics (Table 1). The selection of these traits is illustrative, not exhaustive.

We curated a list of YouTube videos and channels whose content aligns with each target attribute, employing a structured, two-part placement curation methodology.

(1) **Query-based selection:** We constructed search queries designed to reflect the content-seeking behavior of users exhibiting the specific sensitive attribute (e.g., users with specific political views)–see Table 1 for examples. Using the YouTube Data API [23], we retrieved the top 10 relevant videos per query. To ensure high alignment, two independent researchers manually reviewed and labeled each video based on its title, description, hashtags, and top comments. Only videos with inter-rater agreement were retained for the experiment, establishing a validated set of highly targeted placements. We curated **653 videos** (85.8% of the original corpus). (2) **Channel-based selection:** We identified YouTube channels whose primary content is dedicated to, or whose audience is strongly associated with, the sensitive attribute. This selection was informed by analysis of external sources such as online forums, Subreddit communities, and journalistic articles. Placing an ad on a channel makes the ad eligible to appear on any monetized video published by that channel, providing a broader, sustained audience reach. In total, we curated **54 channels** via this method.

Table 1 presents examples of the placement input (queries or channels) used to operationalize the targeting for each sensitive attribute category, together with the number of placements curated, and the total audience size estimated by Google. The complete set of placement inputs is provided in Appendix A.1.

This placement curation method scales easily, allowing continuous discovery of new content and enabling advertisers to target users engaging with material linked to sensitive attributes. We employ limited manual validation, but even without human oversight, the method likely reaches audiences associated with sensitive traits with reasonable accuracy—though it may also capture unrelated users. Such precision aligns with industry norms, as prior work has documented notable accuracy issues in behavioral audience segments [4, 47].

3.2 Controlled ad delivery experiments

We conducted a series of controlled ad experiments on Google Ads to provide empirical validation of this contextual circumvention mechanism across 12 sensitive characteristics.

3.2.1 *Experiment design.*

Campaign design: For each trait, we launched a distinct campaign targeting its curated placements.
Settings and budget: All campaigns used the skippable in-stream video ad format. The bidding strategy was set to maximize impressions within a small €10 lifetime budget. Critically, only placement-based targeting was enabled; all audience expansion and campaign optimization parameters were explicitly disabled to isolate the effect of the selected placements.
Ad creative: The ad creative was an identical, neutral video (promoting drinking water, see Ethics 6), ensuring that the ad content itself did not trigger policy flags or influence ad delivery based on content sensitivity.
Policy compliance check: Campaigns were submitted through Google's standard ad review process, and we documented approval outcomes, including any rejections or flags related to the targeted placements.
Delivery Validation: Each campaign was monitored for 24 hours to confirm successful ad delivery on the intended placements and to record the number of impressions generated.

3.2.2 *Results.*

Policy Compliance Check: All 12 ad experiments, despite targeting content proxies for sensitive attributes, passed Google's ad review process without restriction [20]. This demonstrates that as long as the ad creative and general campaign settings comply with policy, platform enforcement does not prevent advertisers from exploiting content proxies to reach sensitive audiences.

Delivery Validation: To further confirm the circumvention mechanism, we focused on validating ad delivery on a subset of six audience traits: left-leaning and right-leaning political views, Muslim and Jewish religious affiliations, financial hardship, and conspiracy-oriented worldviews. Using Google Ads reporting tools (see Appendix A.6), we confirmed that ad delivery occurred exclusively on the intended videos and channels for each attribute. We stopped the campaigns once successful delivery was confirmed. Collectively, the experiments generated 2,834 impressions within a few hours with an average cost per view of €0.02. This demonstrates that the circumvention mechanism is possible and is resource-efficient—ads can be delivered to audiences defined by sensitive traits reliably and quickly without direct use of user data.

3.3 Scaling conspiratorial worldview proxies

We assess whether contextual placement curation can be scaled to reach audiences aligned with sensitive attributes. Rather than developing a deployable system for *discovering* such audiences—which could be misused—our aim is to evaluate whether a proxy-based approach can *automatically surface thematically consistent content at scale*. To test this, we apply the method to **conspiratorial worldview-related content**, a well-studied domain with existing labeled YouTube datasets, allowing us to benchmark its accuracy in identifying content associated with sensitive audience traits.

We implement an automatic labeling framework using GPT-4o to review the videos, excluding ones that debunked or critically examined the conspiracies, and retaining only those presenting or endorsing the narratives. Each video is assigned a relevance score on a 5-point scale, ranging from 1 ("very unlikely") to 5 ("highly likely") to contain conspiracy or misinformation content (prompt

Table 1: Sensitive attributes proxies.

Sensitive category	Attribute	Examples of input queries and channels	Total no. placements	Total audience size
Political ideology	Left-leaning	HasanAbi, The Majority Report, David Pakman.	10	84M
	Right-leaning	Tucker Carlson, The Daily Wire, Nuance Bro, Charlie Kirk	12	290M
Belief systems and Worldview	Conspiratorial worldview	"Project Blue Beam ", "The Dead Internet Theory", "PROOF AI Is Sentient"	86	150k
Religious belief	Christian	BibleProject, Elevation Church, Grace Digital Network	11	22M
	Jewish	JewishUncensored, Jewish Voice, AishJewish	8	1.7M
	Muslim	Mufti Menk, One path network, Islamic Guidance, Talk Islam	11	1.1M
Health conditions	Mental health conditions	"living with adhd", "living with depression", "how to cope with depression", "lies depression tells you"	80	90k
	Special medical conditions	"Living with Cerebral Palsy", "how to cope with cerebral palsy", "dealing with diabetes"	92	30k
Vulnerable populations & Personal hardships	Financial hardship	"how to live on a small budget", "how to apply for public housing", "low-income tips for budget"	137	190k
	Disabilities	"How to Apply for Disability Benefits", "Coping with Disability", "How I Stay Positive Despite Chronic Illness"	89	20k
Sexual/Gender identity	LGBT	"LGBTQ+ Relationship Advice", "Coping with Homophobia", "Sexuality crisis", "Lesbian Relationship Struggles"	86	60k
	Transgender identity	"how to start HRT", "Top surgery q&a", "Coping with Gender Dysphoria", "Binder safety Tips", "binder tutorial"	69	20k

structure is provided in Appendix A.2). The model is instructed to assess each video using its title, description, and transcript, based on the following criteria:

(1) Promotion of unfounded claims, unsupported theories, or contradictions of widely accepted facts.
(2) Use of language suggesting secrecy, cover-ups, or manipulation by powerful entities (e.g., governments, secret societies, etc.).
(3) Patterns of fear-mongering, distrust in authoritative sources, or appeals to emotion over logic.
(4) Recurrent themes commonly associated with conspiracy theories or misinformation (e.g., anti-vaccine rhetoric, flat earth, etc.).

Our labeling scheme was evaluated against two datasets: (1) a set of 862 videos labeled as conspiracy-related in prior work [18, 30, 31] And, (2) a set of 382 videos unrelated to conspiracy themes which we curated as control, spanning topics such as music, comedy, and mainstream news (§ 4.1.1). Our framework achieved an **overall accuracy of 90.1%**. It correctly identified 86% of the conspiracy-related videos and flagged only two videos from the control set as conspiratorial. A manual review confirmed that both flagged control videos contained speculative or misleading narratives, aligning with our defined labeling criteria.

3.4 Discussion

Our findings show that contextual advertising on YouTube can be systematically used to reach audiences defined by sensitive attributes, circumventing restrictions on behavioral targeting.

Scalability: Our approach for identifying proxies is highly accessible as placement lists can be retrieved efficiently using the YouTube Data API or simple search queries, and the range of targetable attributes is limited only by the availability of content consumption proxies. More importantly, this method does not rely on sophisticated behavioral profiling or algorithmic inference; advertisers simply reach users while they are actively engaging with content that directly signals their traits.

Risk: This mechanism introduces multiple avenues for abuse that raise societal concerns. One is *exploitation of vulnerability*, where advertisers, for instance, can target users seeking content related to financial hardship or mental and physical health struggles with predatory schemes or unproven products. Another is *ideological manipulation*, in which contextual placements are used to isolate audiences already engaged with fringe material, reinforcing echo chambers and amplifying extreme or polarizing messages.

*Legality:*The current EU legal framework does not explicitly address contextual advertising that relies on combinations of content proxies to reach audiences associated with sensitive attributes. Because such targeting operates without directly processing personal data—i.e., profiling as defined by the GDPR[3]—it may fall outside the narrow prohibitions related to targeting of special categories[1,2]. However, while this mechanism might be compliant, it could still be undermining the spirit and protective aims of EU data-protection law (e.g., GDPR (Article 5(1)(a–b)[4] [16]).

4 Contextual advertising and conspiracy content

This section investigates the prevalence of contextual advertising on high-risk YouTube content, specifically using conspiracy and misinformation videos as a test bed. We select fringe content because it raises important safety and integrity concerns while avoiding the direct ethical risks associated with testing other sensitive traits. Our analysis aims to directly assess the extent to which contextual targeting exposes users to potentially harmful or unsafe ad placements (**RQ2**), reflecting the advertising landscape as of 2025.

4.1 Measurement framework

To capture ad delivery, we built a custom *browser extension* that automatically records ads shown during video playback. We deploy

[3]Article 4 (4)"'profiling' means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements;" [16]
[4]Article 5 (1)(a) "Personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency')";[...]

this extension with *automated personas* that simulate different user types, each configured with distinct browsing and search histories reflecting varying interactions with factual and non-factual content. This design allows us to observe how the availability of behavioral data influences the overall prevalence of contextual advertising. We utilized this setup to monitor ad delivery across two video sets: our *curated corpus of conspiracy and misinformation videos* and a matched *control dataset* of general-interest videos. For every recorded ad, we collected ad and advertiser metadata and the *ad targeting explanation* provided by the platform (§ 2), enabling us to isolate and analyze ads attributed to contextual signals.

4.1.1 Conspiracy videos corpus. Existing conspiracy video datasets [18, 30, 31] are outdated, cover limited time windows, and attract few active ads. To address this, we constructed a new corpus reflecting current misinformation and fringe narratives circulating on the platform. Following the proxy-identification process (§ 3.1), we formulated keyword queries based on trending narratives from online forums, social media discussions, fact-checking organizations, and news sources (e.g., "Project Blue Beam New Jersey Drones"). We retrieved videos via the YouTube Data API, and we filtered them using our automated labeling method (§ 3.3). Only videos rated 4 or 5 were retained, resulting in a **curated set of 536 conspiracy and misinformation videos**. This corpus is recent (modal publication year 2024), highly engaged (median 167k views), and has a median duration of 15 minutes.

Control videos corpus. For comparative analysis, we constructed a control dataset of general-interest YouTube videos unrelated to conspiracy and misinformation. We began by retrieving the top 500 English-language videos via keyword queries across twelve broad categories (e.g., Music, Gaming, News), creating an initial pool of 6,000 videos. To ensure comparability with the conspiracy dataset, we applied a filtering process to match the control videos across three dimensions: video length, engagement (log-transformed views), and publication date. After filtering and excluding two videos containing potential misinformation, we obtained a **final control dataset of 382 videos**.

4.1.2 Persona design. To compare ad delivery across varying levels of behavioral signal availability and isolate the role of contextual targeting, we simulated 13 user personas (Table 2). We designed four personas with distinct *browsing history training* and four with *search history training*, with each set further stratified by gender and content factuality (factual or non-factual). To establish critical reference points and isolate contextual cues, we included two personas with *no behavioral history* and two personas where we explicitly *disabled ad personalization* through Google ad settings. We also included a guest user (non-logged-in) to represent an *anonymous viewing scenario*. We provide a detailed description of all configurations in Appendix A.4.

4.1.3 Ad collection.

Browser extension. We developed a custom browser extension that extends prior methodologies [33] to capture and analyze ad delivery. This tool monitors network traffic to pinpoint the request triggering a YouTube video ad. It then programmatically simulates user interaction by clicking the ad's information ("i") panel and parsing the resulting HTML to scrape the targeting explanations. The extension is designed to collect comprehensive details, including the **ad's metadata** (title, advertiser information, and Transparency Center link [24]), the specific **ad targeting explanations** that reveal why an ad has been shown to a user (see § A.5 for examples), and the ad's predicted **topics** from Google.

Table 2: Constructed personas.

Reference	Persona details	Total Ads
P1(F/M)	Profiles with factual browsing history.	1360
P2(F/M)	Profiles with fact-based search history.	1408
P3(F/M)	Profiles with misinformation-related browsing history.	1397
P4(F/M)	Profiles with non-factual-based search history.	1563
P5(F/M)	Profiles with no history.	872
P6(F/M)	Profiles with personalization on Google turned off.	1549
P7	No Google profile.	1058

'F/M' indicates two profiles were created: a female persona and a male one.

Automation setup. To scale the experiment across multiple profiles, we utilized Selenium WebDriver [37] to automate interactions with YouTube within isolated browsing sessions. Each persona was loaded into a separate Selenium instance, authenticated with its associated Google account (excluding the non-logged-in guest user), and equipped with our custom ad collection extension.

We divided the data collection into fixed, short viewing batches during specific times of the day (Morning, Afternoon, Evening, and Late Night) to mitigate detection and simulate natural user behavior. Each batch (i.e., watch session) involved 50 videos, equally split between the conspiracy and control content. To further simulate natural viewing, the personas were programmed to watch each video for a randomly selected duration (5–10 minutes), with randomized pause intervals inserted between videos.

We collected a total of **9,207 ads across 13 personas** and 35 watch sessions.

4.1.4 Ad labeling. To characterize the ads served on conspiracy content, we label each ad along three dimensions:

Targeting type: We grouped the raw ad targeting explanation strings into several high-level categories (e.g., Time/Location, Demographic, Interests, Contextual, etc.). Table 7 in the Appendix presents all resulting categories and their corresponding explanation strings. We then further consolidated these categories into three core analysis groups: **Contextual Targeting**: Explanations that explicitly reference the current content being viewed (e.g., "The video you're watching"). This has been shown through controlled ad experiments in prior work [33] to be the explanation used when ads are contextually targeted. **Personalized Targeting**: Explanations that encompass categories relying on user data or estimations derived from user data, including interests, lookalike audiences, behavioral signals, and customer and remarketing lists exclusions. **No Targeting**: Explanations that only generically reference time and/or location, indicating a lack of granular user or content-based targeting.

Political relevance: To determine whether an ad is political, we apply the SVC-based classifier from Sosnovik et al. [38], originally developed for detecting political ads on Facebook. The classifier is run on the transcript of each ad video. To reduce false positives,

two human annotators manually reviewed the ads that the model labeled as political. We use the agreement between the annotators as ground truth. Ads where agreement was not established are labeled as non-political.

Advertiser and content types: We adopt the labeling taxonomy from Ballard et al. [3] for both advertiser and content types. We use GPT-4o to assign these labels (see § A.3 for prompt structure). We infer advertiser type from metadata on their YouTube page (name, title, description, keywords)[5]. Content type is derived from the ad video's transcript and the hosting YouTube channel's title.

4.2 Results

4.2.1 Prevalence of contextual ads. Table 3 reports the distribution of collected ads by persona, targeting type (contextual, personalized, or none), and corpus (conspiracy vs. control). We identify 1,595 contextual ads (30.86% of the total) during conspiracy watch sessions. In comparison, 1,545 ads (38.35%) on control content were contextual. Although contextual advertising was more common in the control dataset, its substantial prevalence on high-risk conspiracy content indicates that it is a major driver of monetization in these environments.

When examining the distribution across behavioral personas, we found that contextual ads were delivered to all personas except those with disabled ad personalization (P6F and P6M). This outcome is unexpected, as users with disabled ad personalization should receive more contextual ads [44]. We therefore exclude these profiles from analysis for the remainder of this section and investigate this discrepancy further in § 5.

To assess the relative importance of different targeting mechanisms, we compare the proportion of each targeting type received by a given persona across the two corpora (conspiracy vs. control) using two-proportion z-tests. The differences were not statistically significant, except for profiles P4F, P5F, and P5M, which received a higher share of contextual ads on control content.

We further examined variation in contextual advertising across personas within the conspiracy corpus. The differences were also found not statistically significant, with the exception of the fact-based browsing profiles (P1F and P1M), which consistently received the highest average share of contextual ads.

Overall, while contextual ad exposure was slightly higher on control videos, it remained relatively stable across the behavioral profiles, suggesting that a user's past behavioral signals do not significantly alter the overall rate of contextual ad delivery.

4.2.2 Sensitive ad categories. Table 4 presents the distribution of ad labels (§ 4.1.4) across conspiracy and control videos, grouped by targeting mechanism (contextual, personalized, or none). Profiles P6F and P6M were excluded from analysis. We compare the prevalence of each content category across targeting types between the two datasets (conspiracy vs. control) using two-proportion z-tests, applying Fisher's exact test for sparse categories (e.g., differences in medical ads within contextual targeting).

Political relevance: We identified 235 ad instances as political in our dataset, originating from 54 distinct advertisers. We found that political ads appeared disproportionately more on conspiracy-related content compared to control ($p < 0.001$), which shows disproportionate exposure to political ads in conspiracy environments. Moreover, when political ads do appear, they're more often explained as contextually targeted on conspiracy videos than on control ones (6.88% vs. 1.42%, $p < 0.001$).

Table 3: Distribution of targeting type by persona for conspiracy vs. control videos. "Personalized" refers to ads where explanations include the use of any user data. "No Targeting" refers to ads where the targeting explanations include only time and/or location.

	Conspiracy			Control		
Persona	**Contextual**	**Personalized**	**No Targeting**	**Contextual**	**Personalized**	**No Targeting**
P1F	42.17%	51.15%	6.68%	46.11%	47.31%	6.59%
P1M	43.08%	48.31%	8.61%	47.19%	46.44%	6.37%
P2F	46.71%	51.82%	1.46%	49.54%	49.07%	1.39%
P2M	36.27%	61.72%	2.00%	40.33%	55.37%	4.29%
P3F	36.83%	60.99%	2.18%	47.85%	47.85%	4.30%
P3M	40.60%	53.69%	5.70%	48.19%	44.14%	7.66%
P4F	32.55%	**61.60%****	5.85%	**39.56%***	56.87%	3.57%
P4M	35.68%	59.80%	4.52%	44.80%	52.78%	2.43%
P5F	31.28%	**58.97%****	9.74%	**43.21%***	50.62%	6.17%
P5M	40.58%	**54.11%****	5.31%	**53.74%***	38.32%	7.93%
P6F	0.0%	0.0%	100%	0.0%	0.0%	100%
P6M	0.0%	0.0%	100%	0.0%	0.0%	100%
P7	34.53%	46.15%	**19.32%*****	**43.13%***	48.41%	8.46%

p-values were adjusted using Benjamini-Hochberg FDR
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Advertiser types: We highlight two sensitive advertiser categories: *self-improvement* (dubious "get-rich-quick" schemes, physical transformation programs, and alternative health products) and *special interest groups* (government agencies, NGOs, charities, and lobbying organizations). Both categories appeared disproportionately more on conspiracy content compared to control ($p < 0.001$). We find that ads from both of these categories made up a significantly larger share of contextual ads on conspiracy than on control videos (*self-improvement:* 7.46% vs. 2.13%, $p < 0.001$; *special interests:* 10.21% vs. 4.14%; $p < 0.001$). This demonstrates that the over-representation of these sensitive advertisers on high-risk content is not just a volume effect but is also disproportionately concentrated within contextual targeting mechanisms.

Content types: We identify five categories of sensitive ads: (1) *alternative health* (e.g., dubious health products and unverified medical services), (2) *government*, (3) *medical* (e.g., medicinal products and healthcare services), (4) *third sector* (e.g., non-profit organizations), and (5) religion (e.g., religious services or products). Given the conceptual overlap between government and third sector ads and those from special interest groups (all of which were included in that broader category), we focus our analysis on the remaining three categories. Each appeared disproportionately more frequently on conspiracy than on control videos, consistent with prior findings [3] (*alternative health* and *religion*: $p < 0.001$; *medical*: $p = 0.10$). These categories were overrepresented in contextually targeted ads on conspiracy content compared to control (*alternative health*: 1.37% vs. 0.38%, $p < 0.01$; *medical*: 8.08% vs. 6.01%, $p < 0.05$; *religion*: 2.13% vs. 0.0%, $p < 0.001$). Contextual targeting was also the dominant explanation used for ads with these labels (values in red in Table 4).

[5]We chose not to rely on Wikidata, as many advertisers were either not listed or marketed themselves under names that did not match their formal entries.

Table 4: Distribution of targeting types by advertiser and content category.

	Conspiracy			Control		
	Contextual	Personalized	No Targeting	Contextual	Personalized	No Targeting
Political (manual)	97***	60	5	22	12	3
Advertiser Category						
Aggregator sites	24	**61****	24	15	16	4
Information Media	200	299	33	217	243	11
Merchant	1064	**1730*****	173	**1194*****	1325	162
Miscellaneous	25	19	1	22	**26***	3
Self Improvement	**119*****	**150*****	46	33	31	5
Special Interest Groups	163***	89	13	64	53	1
Content Category						
Alternative Health	**22****	**18***	24	6	3	3
Automotive	12	68	3	91**	**70***	0
Beauty	11	19	2	19	**27***	1
Business and Finance	99	102	37	**155*****	94	7
Education	**133*****	**136*****	9	47	36	2
Entertainment	**100***	**148***	18	70	137	38
Food and Drink	54	77	4	**105*****	**90****	3
Games	**32***	77	2	17	42	1
Gold and Precious Metals	4	0	0	1	2	0
Government	33*	6	6	10	11	0
Home Goods	35	**146****	16	51	66	5
Industrial	49*	11	2	28	**17***	1
Insurance	25	43	11	25	41	13
Legal	30***	22	13	3	2	3
Lifestyle (fitness, fashion)	31	209	17	**70*****	130	10
Major Retailer	**36***	118	4	20	95	9
Marketing and Advertising	51	108	26	33	82	30
Medical	129*	84	16	93	58	2
Miscellaneous	220	304	27	230	244	9
Third sector	92***	**31*****	6	18	5	2
Real Estate	13	**35*****	3	14	5	1
Religion	34***	12	0	0	4	0
Software	93	131	14	83	93	6
Sports	0	2	0	0	0	0
Technology	225	354	30	312***	261	30
Travel and Tourism	32	87	11	44	62	13
Total	1595	2348	290	1545	1694	186

Values in red indicate cases where contextual targeting is the predominant method within the category
p-values were adjusted using Benjamini-Hochberg FDR
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

4.3 Discussion

Prevalence of contextual targeting on fringe content. The high prevalence of contextual ads on conspiracy content highlights the role of contextual advertising in monetizing these environments. **From a user perspective**, this raises serious safety concerns. Because ads are tied directly to the content being consumed, they inevitably appear alongside material that is misleading, fringe, or harmful. If the match between the ads and the placements (i.e., conspiracy content) is algorithmically inferred, the "context" driving ad delivery may reflect distorted or unsafe themes (e.g., pseudoscience or extremist narratives). If the placement is advertiser-selected, the choice to appear in such environments carries risks of manipulation or exploitation. **From an advertiser's perspective**, the reliance on contextual signals introduces brand safety risks. When ad delivery is steered by opaque algorithms, advertisers may be unintentionally funding and legitimizing harmful content, while also exposing their audiences to unsafe associations.

Contextual targeting of sensitive ads on fringe content. Sensitive ad categories appeared disproportionately within conspiracy content and were most often delivered through contextual targeting on conspiracy compared to control. *Political* ads were notably overrepresented, indicating that contextual placements can expose users to targeted political messaging in high-risk environments. We also observed higher shares of *sensitive advertiser types*, particularly *self-improvement* (e.g., dubious financial or health products) and *special interest groups* (e.g., government agencies, NGOs, charities, and lobbying organizations). Contextual ads further clustered in categories with safety implications: *Alternative Health* and *Medical* intersecting with misinformation spaces could amplify pseudoscientific claims, while *Religion* ads may profile users along sensitive attributes, raising privacy concerns. Overall, these patterns reveal a broader systemic risk of contextual advertising—namely, its capacity to reproduce sensitive and unsafe ad placements even without behavioral data.

5 Contextual advertising and transparency

Using the dataset described in § 4 and supplementary experiments, we examine how contextual ad targeting is communicated to users and represented in Google's public transparency resources.

Transparency in user-facing explanations. User-facing ad explanations are intended to clarify the rationale behind ad delivery, a requirement formalized by EU legislation such as the DSA Article 26 [13][6], which mandates that users receive information about why an ad was shown and the main parameters used for targeting.

Across all ad explanations we collected in § 4, the only explanation provided with respect to contextual advertising is the generic phrase: "The video you are watching." This provides no actionable information regarding how the specific placement was determined—whether the result of explicit **advertiser choices** (e.g., keyword or topic selection) or complex **platform-level optimization**. Consequently, the lack of granularity prevents users from meaningfully interpreting the targeting logic or detecting potentially malicious or manipulative ad practices.

Auditing political ad transparency. The auditability of political advertising depends on the accessibility and completeness of transparency repositories mandated by regulation. In the EU, the Digital Services Act (DSA) requires very large online platforms (e.g., Google) to provide publicly searchable ad repositories containing information on each ad's content, display period, targeting parameters, and audience reach.[7] The more recent European Democracy Shield (EDS) regulation extends these obligations for political advertising, requiring disclosure of micro-targeting and ad-delivery mechanisms.[8] Although our dataset was collected in the United

[6] Article 26 (1) "[..](d) meaningful information directly and easily accessible from the advertisement about the main parameters used to determine the recipient to whom the advertisement is presented and, where applicable, about how to change those parameters."
[7] Article 39(1): "Providers of very large online platforms or of very large online search engines that present advertisements on their online interfaces shall compile and make publicly available in a specific section of their online interface, through a searchable and reliable tool [...]" [12].
[8] Article 19(1): "When using targeting techniques or ad-delivery techniques in the context of online political advertising involving the processing of personal data, controllers shall [...] adopt, implement and make publicly available an internal policy [...] and keep records on the relevant mechanisms and parameters used."

States, comparable transparency mechanisms for political advertising still apply. Google's Ads Transparency Center serves as the main instrument for meeting these requirements across its advertising platforms, offering public access to data intended to support external scrutiny of political ads.

We cross-referenced political ads from our dataset with entries in Google's Ads Transparency Center for Political Advertising [21] and identified 62 matches. Several advertiser entries were missing or inaccessible, as the Center does not support straightforward matching (scraping was required). In addition, a number of ads we labeled as political were not classified as such by Google. This is due to the inherent complexity of detecting political advertising, as definitions can vary across platforms, jurisdictions, and even annotators [38]. Out of the matched entries, 33 cases included contextual targeting explanations in our data, yet none of their matched entries in the repository contained information about the exact contextual targeting parameters that were used for targeting; disclosures were limited to demographic (age, gender) and geographic targeting. We identify this as a gap in transparency that does not meet the requirements put forward by regulation, and makes independent auditing of political advertising on the platform difficult.

Moreover, despite Google's policy restricting political ads to only demographic, geographic and contextual targeting [22], 24 matched ads featured interest-based, behavioral, or lookalike audience explanations in user-facing disclosures, yet no violations were reported in the Transparency Center.

Transparency for users with disabled personalization. Profiles with disabled ad personalization (P6F and P6M; see Table 2) were designed to capture ads served solely through contextual cues, as ad personalization is disabled. However, as shown in Table 3, these profiles only received explanations referencing "time and location", with no mention of contextual signals. This contradicts Google's documentation [44], which specifies that ads for such users should rely on contextual information. Two interpretations are possible for this: either contextual targeting was genuinely not applied, or it was applied but not disclosed.

To test the reliability of ad explanations shown on disabled personalization profiles, we conducted a targeted experiment using two YouTube profiles: one with personalization enabled and one disabled. Both profiles were used to search for the query "Travel to Tokyo," a behavior typically known to trigger keyword-based targeted ads on the search results page. Both profiles were subsequently served the identical travel ad from the same advertiser, with identical URL parameters. However, the ad explanations provided to the users differed significantly: the personalized profile correctly cited search terms as the targeting mechanism, while the disabled personalized profile listed only time and location. This result strongly suggests that contextual cues are being used for ad delivery even when personalization is disabled, without being adequately disclosed to the user.

A similar pattern appears in our main dataset. We identified 50 instances of ads that were shown on the exact same videos to both personalized and non-personalized profile types. For users with personalization enabled, these ads were labeled as being contextually targeted, while, when the same ads were delivered to the non-personalized users, the ad explanations provided no targeting rationale besides time and location. This pattern of inconsistent disclosure suggests that contextual signals were likely applied in both contexts but were selectively disclosed to users. This finding aligns with the work of Medjkoune et al. [33], who demonstrated through controlled ad campaigns that explanations for ads delivered to videos labeled as "Made for kids" similarly failed to disclose the actual targeting reasons specified by advertisers.

6 Ethics

This research was conducted under an approved Institutional Review Board (IRB) protocol. The ad experiments in § 3 used neutral campaign material (a generic video ad about drinking water) and fully complied with Google Ads' terms of service. The measurement study on conspiracy content (§ 4) relied solely on automated personas and controlled browsing environments, without involving human subjects or affecting real user experiences. Our analysis critiques the opacity and potential misuse of contextual advertising systems to promote transparency and accountability in online advertising. No proprietary data was accessed, and no platform integrity was compromised. The findings expose a **systemic risk** rather than a technical vulnerability.

7 Related Work

A substantial body of work has examined the privacy and ethical risks of behavioral advertising, highlighting its dependence on large-scale user tracking and opaque algorithmic optimization [1, 2, 8, 10, 11, 36, 39]. Such practices enable the inference of sensitive attributes and discriminatory targeting, prompting regulatory interventions such as the GDPR, CCPA, and DSA [14, 16, 41]. These developments have spurred renewed interest in contextual advertising as a privacy-preserving alternative. Yet empirical work on its privacy and safety implications remains limited. Medjkoune et al. [33], for example, demonstrate that contextual systems can be exploited to reach protected groups such as children.

Parallel research has analyzed how YouTube's recommendation and search algorithms propagate fringe and conspiratorial content [18, 26, 27, 35], producing echo chambers and filter bubbles driven by personalization and algorithmic amplification [6, 28, 29, 31, 40, 45, 46]. However, the role of advertising in these environments remains underexplored; the only prior study by Ballard et al. [3] focuses on monetization. We extend this line of inquiry by examining how contextual signals within fringe narratives shape ad delivery and targeting dynamics.

8 Conclusion

This study examined how contextual advertising on YouTube can be leveraged to reach vulnerable audiences. Our findings show that contextual advertising warrants the same level of scrutiny as behavioral targeting, as it can produce comparable risks when deployed in sensitive or ideologically charged contexts. We recommend that platforms disclose advertiser-selected contextual criteria—especially for political ads—in their Transparency Center, and clarify which content signals (e.g., keywords, topics, metadata) inform ad delivery. Restricting behavioral targeting alone does not ensure safety, fairness, or compliance with public-interest goals.

References

[1] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

[2] Julia Angwin and Terry Parris Jr. 2016. Facebook lets advertisers exclude users by race. *ProPublica blog* 28 (2016).

[3] Cameron Ballard, Ian Goldstein, Pulak Mehta, Genesis Smothers, Kejsi Take, Victoria Zhong, Rachel Greenstadt, Tobias Lauinger, and Damon McCoy. 2022. Conspiracy brokers: Understanding the monetization of youtube conspiracy theories. In *Proceedings of the ACM Web Conference 2022*. Association for Computing Machinery, 2707–2718.

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

[5] Salim Chouaki, Abhijnan Chakraborty, Oana Goga, and Savvas Zannettou. 2024. What News Do People Get on Social Media? Analyzing Exposure and Consumption of News through Data Donations. In *Proceedings of the ACM Web Conference 2024*. Association for Computing Machinery, 2371–2382.

[6] Sam Clark and Anna Zaitsev. 2020. Understanding YouTube communities via subscription-based channel embeddings. *arXiv preprint arXiv:2010.09892* (2020).

[7] Corvidae. 2024. *Programmatic Advertising Vs Contextual Advertising | Corvidae.* https://corvidae.ai/blog/programmatic-advertising-vs-contextual-advertising-exploring-the-path-to-advertising-excellence/

[8] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2014. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on Privacy Enhancing Technologies 2015* (2014).

[9] Stacey J Dixon. 2025. *Most popular social networks worldwide as of February 2025, by number of monthly active users.* https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[10] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 1388–1401.

[11] Steven Englehardt, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W Felten. 2015. Cookies that give you away: The surveillance implications of web tracking. In *Proceedings of the 24th International Conference on World Wide Web*. 289–299.

[12] European Parliament and Council of the European Union. 2022. *Article 39, the Digital Services Act (DSA).* https://www.eu-digital-services-act.com/Digital_Services_Act_Article_39.html

[13] European Parliament and of the Council. 2022. *Article 26, the Digital Services Act (DSA).* https://www.eu-digital-services-act.com/Digital_Services_Act_Article_26.html

[14] European Parliament and of the Council. 2022. *Regulation - 2022/2065 - EN - DSA - EUR-Lex.* https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng

[15] European Union. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* http://data.europa.eu/eli/reg/2016/679/oj

[16] European Union. 2018. *General Data Protection Regulation (GDPR) – Legal Text.* https://gdpr-info.eu/

[17] eurostat. 2024. *Internet advertising of businesses - statistics on usage of ads.* https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/54450.pdf

[18] Marc Faddoul, Guillaume Chaslot, and Hany Farid. 2020. A longitudinal analysis of YouTube's promotion of conspiracy videos. *arXiv preprint arXiv:2003.03318* (2020).

[19] Google Ads Help. [n. d.]. *Targeting your ads.* https://support.google.com/google-ads/answer/1704368?hl=en

[20] Google Ads Help. 2025. *About the ad review process - Google Ads Help.* https://support.google.com/google-ads/answer/1722120?hl=en

[21] Google Ads Transparency Center. [n. d.]. *Political advertising on Google.* https://adstransparency.google.com/political?region=US&format=VIDEO

[22] Google As Help. 2025. *Political content - Advertising Policies Help.* https://support.google.com/adspolicy/answer/6014595#:~:text=Restricted%20targeting%20for%20election%20ads5

[23] Google for Developers. 2025. *API Reference | YouTube Data API | Google for Developers.* https://developers.google.com/youtube/v3/docs

[24] Google Transparency Center. 2025. *Political Advertising.* https://adstransparency.google.com/political?region=US&format=VIDEO&topic=political

[25] Emil Häglund and Johanna Björklund. 2024. AI-driven contextual advertising: Toward relevant messaging without personal data. *Journal of Current Issues & Research in Advertising* 45, 3 (2024), 301–319.

[26] Muhammad Haroon, Magdalena Wojcieszak, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, and Zubair Shafiq. 2023. Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the national academy of sciences* 120, 50 (2023), e2213020120.

[27] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring misinformation in video search platforms: An audit study on YouTube. *Proceedings of the ACM on human-computer interaction* 4, CSCW1 (2020), 1–27.

[28] Raj Jagtap, Abhinav Kumar, Rahul Goel, Shakshi Sharma, Rajesh Sharma, and Clint P George. 2021. Misinformation detection on YouTube using video captions. *arXiv preprint arXiv:2107.00941* (2021).

[29] Prerna Juneja and Tanushree Mitra. 2021. Auditing e-commerce platforms for algorithmically curated vaccine misinformation. In *Proceedings of the 2021 chi conference on human factors in computing systems*. 1–27.

[30] Mark Ledwich and Anna Zaitsev. 2019. Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *arXiv preprint arXiv:1912.11211* (2019).

[31] Shao Yi Liaw, Fan Huang, Fabricio Benevenuto, Haewoon Kwak, and Jisun An. 2023. Younicon: Youtube's community of conspiracy videos. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 17. 1102–1111.

[32] MDN. 2025. *Third-party cookies - Privacy on the web | MDN.* https://developer.mozilla.org/en-US/docs/Web/Privacy/Guides/Third-party_cookies

[33] Tinhinane Medjkoune, Oana Goga, and Juliette Senechal. 2023. Marketing to children through online targeted advertising: Targeting mechanisms and legal aspects. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. 180–194.

[34] Lukasz Olejnik, Tran Minh-Dung, and Claude Castelluccia. 2013. *Selling off privacy at auction.*

[35] Kostantinos Papadamou, Savvas Zannettou, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Michael Sirivianos. 2022. "It is just a flu": assessing the effect of watch history on YouTube's pseudoscientific video recommendations. In *Proceedings of the international AAAI conference on web and social media*, Vol. 16. 723–734.

[36] Franziska Roesner, Tadayoshi Kohno, and David Wetherall. 2012. Detecting and defending against {Third-Party} tracking on the web. In *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*. 155–168.

[37] Selenium. 2025. *WebDriver.* https://www.selenium.dev/documentation/webdriver/

[38] Vera Sosnovik and Oana Goga. 2021. Understanding the complexity of detecting political ads. In *Proceedings of the Web Conference 2021*. 2002–2013.

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

[40] Ivan Srba, Robert Moro, Matus Tomlein, Branislav Pecher, Jakub Simko, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, Adrian Gavornik, et al. 2023. Auditing YouTube's recommendation algorithm for misinformation filter bubbles. *ACM Transactions on Recommender Systems* 1, 1 (2023), 1–33.

[41] State of California Department of Justice. 2018. *California Consumer Privacy Act (CCPA).* https://oag.ca.gov/privacy/ccpa

[42] Statista. [n. d.]. *Social media marketing's top global benefits 2024.* https://www.statista.com/statistics/188447/influence-of-global-social-media-marketing-usage-on-businesses/

[43] Google Support. [n. d.]. *Get started with My Ad Center - My Ad Center Help.* https://support.google.com/My-Ad-Center-Help/answer/12155154?hl=en

[44] Support Google. 2025. *How personalized ads work - Android - My Ad Center Help.* https://support.google.com/My-Ad-Center-Help/answer/12155656?hl=en&co=GENIE.Platform%3DAndroid

[45] Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, and Maria Bielikova. 2021. An audit of misinformation filter bubbles on YouTube: Bubble bursting and recent behavior changes. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 1–11.

[46] Matus Tomlein, Branislav Pecher, Jakub Simko, Ivan Srba, Robert Moro, Elena Stefancova, Michal Kompan, Andrea Hrckova, Juraj Podrouzek, and Mária Bieliková. 2022. Black-box Audit of YouTube's Video Recommendation: Investigation of Misinformation Filter Bubble Dynamics. In *IJCAI*. 5349–5353.

[47] Giridhari Venkatadri, Piotr Sapiezynski, Elissa M Redmiles, Alan Mislove, Oana Goga, Michelle Mazurek, and Krishna P Gummadi. 2019. Auditing offline data brokers via facebook's advertising platform. In *The World Wide Web Conference*. 1920–1930.

[48] John Wilander. 2020. Full third-party cookie blocking and more. *WebKit. Preuzeto s: https://webkit. org/blog/10218/full-third-party-cookie-blocking-and-more* (2020).

A Appendices

A.1 Placement selection in ad experiments

We provide the extended list of queries and channels we used for conducting ad experiments (§ 3). See Table 5.

Table 5: Extended list of queries and channels for selecting placements to target users in sensitive attribute categories.

Attribute category	Sensitive attribute	Placements examples ("query" or channels)
Political ideology	Left-leaning	The young turks, The majority report, HasanAbi, David Pakman, The Bulwark, Democracy Now, Some More News, Philosophy Tube, Luke Beasley, Second thought.
	Right-leaning	Tucker Carlson, The Daily Wire, Fox News, Steven Crowder, PragerU, Actual Justice Warrior, Paul Harrell, Tim Pool, Nuance Bro, Mark Dice, Hodge Twins, Charlie Kirk
Belief systems and Worldview	Conspiracy	"Project Blue Beam New Jersey drones", "Luigi Mangione psyop", "AI Will Cause the Apocalypse", "Proof AI Will Soon Control the World", "Proof AI Is Controlling the Internet", "PROOF AI Is Sentient", "The Great Reset", "5G radiation make you sick", "the simulation theory", "Social Media Manipulation", "The Dead Internet Theory", "fake youtube views dead internet", "The plot to steal the elections"
Religious belief	Christian	Joel Olsteen, T.D. Jakes Ministries, BibleProject, CBN News Channel, Sid Roth's It's Supernatural!, Elevation Church, Living Waters, Off The Kirb Ministries, Grace Digital Network, Paul Begley, Treasure Christ.
	Jewish	Discovering the Jewish Jesus with Rabbi Schneider, JewishUncensored, J-TV: Jewish Ideas. Global Relevance, Jewish Voice, AishJewish, Hidabroot - Torah & Judaism, The Jewish Chronicle, The Jewish Convert, Jewish Marriage, Discovering the Jewish Jesus Podcast
	Muslim	merfulservant, Mufti Menk, One path network, I'm a Muslim, Islamic Guidance, Islam The Ultimate Peace, Talk Islam, Islam Populer, Hi-Tech Islamic Naat, Islamic Teacher Official, Madani Channel
Health conditions	Mental health conditions	"living with adhd", "living with depression", "solutions for depression", "solutions for adhd", "how to deal with depression and anxiety", "how to cope with depression", "lies depression tells you", "what helps during depression", "depression in workplace"
	Special medical conditions	"Living with Cerebral Palsy", "how to cope with cerebral palsy", "dealing with diabetes", "diabetes quick solutions", "cooking recipes for diabetes", "dealing with chronic illness", "Parkinson's medication side effects", "Deep brain stimulation for Parkinson's", "Meditation for tremors", "Parkinson's clinical trials"
Vulnerable populations/Personal Hardships	Financial hardship	"how to live on a small budget", "how to apply for public housing", "low-income tips for budget", "tips to budget with small income", "how to pay dept with low-income", "dollar tree shoping tips for low budget", "$1 meals idea", "food stamps tutorial", "food stamps tips", "cooking on a tight budget", "how to apply for SNAP", "how to apply for medicaid", "how to apply for EBT", "how to apply for free internet programs for low-income families", "how to get housing assistance", "meal planning with SNAP"
	Disabilities	"How to Apply for Disability Benefits", "Wheelchair Dance Class", "Coping with Disability", "Disability Burnout", "How I Stay Positive Despite Chronic Illness", "Best Wheelchairs for Travel", "Adaptive Kitchen Tools for One-Handed Cooking", "Using Voice Assistants as a Visually Impaired Person", "How I Drive with Hand Controls", "Navigating college with a Disability"
Sexual/gender identity	LGBT	"How I Knew I Was Gay", "LGBTQ+ Relationship Advice", "Coping with Homophobia", "Sexuality crisis", "Lesbian Relationship Struggles", "am I a lesbian", "am i gay", "how to know you are queer", "how to deal with homophobia in workplace", "gay relationship advice"
	Transgender identity	"Starting Hormone Replacement Therapy", "how to start hormone replacement therapy", "Top surgery q&a", "changing your gender legally", "Coping with Gender Dysphoria", "Makeup for trans woman", "Binder safety Tips", "binder tutorial"

A.2 Conspiracy video labeling

We provide the labeling prompt used to identify conspiracy content. The prompt follows a Chain of Thought (CoT) structure, integrating a Likert scale assessment. We also requested the model to output its reasoning, which allows us to verify the quality of the labeling.

```
**Task**
   You are an annotator trained to evaluate text for themes,
   language patterns, and claims associated with conspiracy
   theories or mis/disinformation. To do this, you are
   provided with a YouTube video transcript, the video's
   title, and its description (when it is available).

   **Instructions**
   1. Identify any claims that lack credible evidence,
   promote unfounded theories, or contradict widely
   accepted facts.
   2. Look for language that suggests secrecy, cover-ups,
   or manipulation by powerful entities (e.g., governments,
   corporations, or secret societies).
   3. Check for patterns of fear-mongering, distrust of
   authoritative sources or appeals to emotion over logic.
   4. Note any recurring themes commonly associated with
   conspiracy theories or mis/disinformation (e.g., anti-
   vaccine rhetoric, flat Earth theories, election fraud
   claims, etc.).
   5. Provide a final assessment, using a Likert scale:
       1: Very Unlikely
       2: Unlikely
       3: Neutral
       4: Likely
       5: Very Likely
   That the video supports conspiracy theories or
   mis/disinformation.
   Video Title: {title}
   Video Description: {description}
   Video Transcript:
   {transcript}

   **Output Format**:
   Format the output as a JSON object:
   {{
       "Final Assessment": The Score you gave for the video
       using the description
       "Reasoning": In a couple of sentences, explain your
       conclusion based on evidence from your assessment
   }}
```

A.3 Content Labeling Prompt

We experimented with multiple prompt strategies to improve labeling accuracy of content and advertiser type labeling, including concise prompts, detailed prompts, Chain of Thought (CoT), and varying the number of classes. We provide the best-performing prompt in terms of accuracy, which was selected for labeling the ads. All prompts were evaluated on a manually labeled dataset of 50 ads (we used agreement between two annotators to set the label).

The selected prompt reached an accuracy of 86% for content type and 83% for advertiser type.

```
You are provided with the title of a YouTube channel, its
description, keywords added by the advertiser to the channel,
and a transcript of a video advertisement they posted.

Here is the input:
    * Advertiser Channel Name: {advertiser_channel}
    * Transcript: {transcript}
    * Description: {description}
    * Keywords: {keywords}

Your task is to analyze these inputs to determine:
    1. **Business category**: assign a high-level category
    that describes the business of the advertiser based on
    the Advertiser Channel Name, Description, and Keywords.

    2. **Content category**: assign a high-level category
    that describes the content/topic of the ad, and what
    it is promoting.

Choose one value from each of the following lists:

Business Category:
    - Aggregator sites : Link aggregators, review sites,
    and other aggregated information advertising for products
    not owned by them
    - Information Media: News sites, blogs, podcasts and other
    media that provide information not advertising a specific
    product
    - Merchant: Sites that sell products, or provide some sort
    of services, such as e-commerce sites, retailers, service
    providers, lawyers, real estate agents, etc.
    - Self Improvement: Sites that provide self-improvement
    products or services, such as courses, coaching, or personal
    development tools
    - Special Interest Groups: Charities, Governmental or
    Political organizations, non-profits, lobbies

Content Category:
    - Alternative Health
    - Automotive
    - Beauty
    - Business and Finance
    - Education
    - Entertainment
    - Food and Drink
    - Games
    - Gold and Precious Metals
    - Government
    - Home Goods
    - Industrial
    - Insurance
    - Lifestyle (fitness, fashion, etc.)
    - Legal
    - Marketing and Advertising
    - Real Estate
    - Religion
    - Retail
    - Medical
    - Miscellaneous
    - Political (outside of government)
    - Software
    - Technology
    - Travel and Tourism

Return your output strictly in the following JSON format:

{{
    "business_category": "<single best-matching category
    from Business Category list>",
    "content_category": "<single best-matching category
    from Content Category list>",
    "ad_id": "{ad_id}"
}}

Do not include any additional text or explanations, just
return the JSON.
```

A.4 Persona design

A.4.1 Behavior-based personas. We construct eight personas in total along two behavioral dimensions: **browsing history** and **search history**. For each dimension, we create two behavioral conditions—factual and non-factual—and instantiate one male and one female persona per condition.

Browsing history training. We construct four personas (two factual and two non-factual) with distinct browsing patterns. **Non-factual browsing personas** visit a curated list of misinformation websites identified in prior work [5], which propagate false or misleading narratives in domains such as health and politics. **Factual browsing personas** visit mainstream, evidence-based news sources with established reputations for accuracy [5]. Each persona is programmed to visit 50 URLs in total, distributed across multiple sessions. The browsing behavior is automated to resemble natural interaction patterns, including page scrolling and dwell time.

Search history training. We construct four personas based on search activity (again split by factuality and gender). **Non-factual search personas** issue queries based on conspiracy-related phrases (e.g., "AI controlling elections", "vaccine microchip theory"). **Factual search personas** search for evidence-based alternatives (e.g., "how vaccines work", "certified election results"). Each persona performs a fixed sequence of queries over multiple sessions to simulate natural user behavior. The profiles search 36 queries.

A.4.2 Non-behavioral personas. To isolate the influence of behavioral signals from contextual cues, we design personas with **no behavioral history**. These profiles allow us to assess whether, in the absence of behavioral data, users are more likely to receive ads based on context cues. We implement two experimental conditions:

No history profiles. Two personas (one male and one female) are not exposed to any browsing or search activity prior to ad collection. These profiles serve as *neutral baselines* that reflect the ad experience of users who have not yet engaged in behavior that might inform targeting.

No personalization profiles. Two personas (one male and one female) are configured to explicitly opt out of ad personalization through Google ad settings. These users do not generate or retain any browsing, search, or cookie-based data. According to Google [44], ads shown to such users should rely exclusively on broad demographics, location data, and contextual relevance.

A.4.3 No profile persona. Finally, we simulate an anonymous user by accessing YouTube in *Guest* mode, without log-in.

A.5 User-facing ad explanations

Figure 2 shows a screenshot of an ad explanation panel (*"Why are you seeing this ad?"*). This panel appears when users click on the information icon ("i") of an ad on YouTube. Table 7 shows example explanation strings and their corresponding targeting category.

A.6 Google reports screenshot

Figure 3 is a screenshot of a Google Ads reporting tool showing the ad impressions we received while running the Political right-leaning proxies experiment. The impressions are shown per placement, in this case, the video channels we included for targeting.

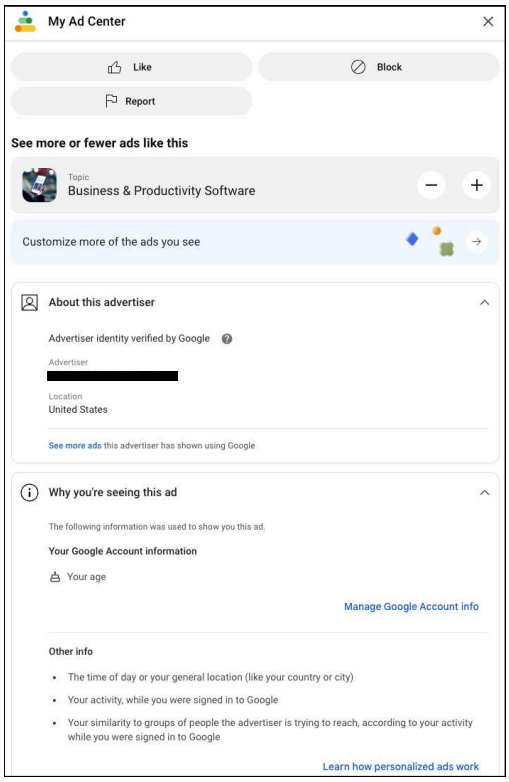

Figure 2: Example screenshot of user-facing ad explanations on YouTube.

Table 7: Ad targeting categories with examples of explanation strings from the "Why you're seeing this ad" notice.

Category	Representative Explanation Strings
Time/Location	"The time of day or your general location (such as your country or city)", "The time of day", "Your general location"
Demographic	"Your age", "Your gender", "Household income range", "Parental status"
Detailed Demographic	"Education status", "Job industry", "Relationship status", "Employer size range"
Interests	"Google's estimation of your interests"
Contextual	"The video you're watching"
Lookalike Audiences	"Your similarity to groups of people the advertiser is trying to reach"
Behavioral	"Websites you've visited", "Your activity while signed in to Google"
Device Type	"The type of device (such as phone, tablet, or desktop computer) where you're currently seeing this ad"
Customer Lists exclusions	"The advertiser's interest in reaching new customers who haven't bought something from them before"
Remarketing exclusions	"The advertiser's interest in reaching people who may not have interacted with them recently"

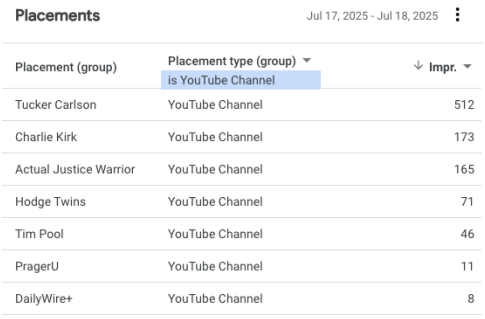

Figure 3: Example screenshot for ad campaign placement monitoring on Google Ads.