Online misinformation pays. Why? Taking stock of a broad range of evidence

March 2025

Report authors

WHO TARGETS ME?

SCIENCE FEEDBACK

Executive summary

This report examines the financial infrastructure that enables misinformation in Europe through a broad analysis of advertising and monetization practices across major **social media platforms and services offering ads on the open web** (Facebook, Twitter/X, YouTube and Google Display Ads).

Monetization of Low-Credibility Content:

Approximately half of low-credibility accounts and websites are actively monetized across platforms:

- 55.6% of eligible[1] low-credibility YouTube channels are monetized (vs. 75.0% of high-credibility channels),
- 48.1% of low-credibility websites carry Google Display Ads (vs. 79.2% of high-credibility sites),
- 41.7% of eligible low-credibility Facebook Pages are monetized (vs. 68.3% of high-credibility Pages).

Brand Safety and Ad Placement:

- Interviews with ad placement professionals suggest that advertisers largely rely on platforms' default brand safety options. Tools and incentives to ensure strong scrutiny of the actors supported by advertising are overall lacking.
- Low-quality advertisers (promoting potential scams, questionable health products, get-rich-quick schemes) appeared in similar proportions on both high and low-credibility content on Twitter/X and YouTube, suggesting that brand safety choices, if any, did not have a material impact on ad placement.

[1] Channel or account meeting the geographic and audience criteria necessary to be considered for ad-revenue sharing on the platform (see Section IV. Research).

By collecting a large dataset of ads, we were also able to audit whether the platforms' ad repositories and political ad transparency satisfactorily met European Union requirements.

Political Ad Transparency:

- Facebook: **23.7%** of political/social issue ads from non-political entities were **not properly labeled**,
- Twitter/X: At least 68 unique political ads were shown in Germany in the run-up to the February 2025 elections despite platform policies prohibiting them,
- YouTube: All political campaign ads were properly labeled, but issue-based ads that could influence elections are not covered by current policies.

Platform ad repositories:

- **Twitter/X's ad repository** exhibits systematic technical issues making it largely **unusable** for research,
- One ad (out of 200) sampled from Facebook was not findable in the platform's ad library.
- On YouTube, all ads sampled were found on the service's ad repository.

Regulatory Considerations

These findings suggest that the systems used by Twitter/X, Meta and Google to ensure their platforms do not reward misinformation are insufficient. Meta and Google have withdrawn from key measures under the EU Code of Conduct on Disinformation that address monetization by repeat spreaders of misinformation, while Twitter/X has left the Code entirely.

The significant monetization of low-credibility content coupled with inadequate brand safety tools and inconsistent ad transparency practices warrants further attention from regulators assessing

compliance with the Digital Services Act, particularly regarding systemic risk mitigation for civic discourse and compliance with relevant sections of the Code of Conduct on Disinformation.

Report Authors

Who Targets Me

Who Targets Me is an organization focused on increasing transparency in online political advertising. Founded in 2017, it aims to expose how political campaigns use targeted ads, particularly on social media. By creating a crowdsourced browser extension, the organization has gathered data from over 115,000 users, driving public conversations on the need for ad transparency.

Science Feedback

Science Feedback is a publication verifying the credibility of influential information and media coverage that claims to be scientific in fields that are particularly prone to misunderstandings and misinformation such as climate change and health. It is operated by a not-for-profit organization.

Acknowledgements

European | MEDIA AND INFORMATION | Fund

Managed by
Calouste Gulbenkian Foundation

The sole responsibility for any content supported by the European Media and Information Fund lies with the author(s) and it may not necessarily reflect the positions of the EMIF and the Fund Partners, the Calouste Gulbenkian Foundation and the European University Institute.

the bright initiative | by Bright Data

Some of the data used in this report was collected using tools provided under The Bright Initiative by Bright Data.

I. Overall research design and objectives

The report aims to look at both sides of the for-profit disinformation ecosystem in Europe:

- **Advertisers**, that, inadvertently or willingly, fund disinformation,
- **Repeat spreaders of misinformation** themselves and, more specifically, their (in)ability to monetize using various mechanisms provided by the largest technology companies.

The main research questions this report aims to address are:

Q1- Are social media platforms and the largest ad placement services effectively preventing repeat sources of misinformation from benefiting from their ad revenue-sharing programs?

Q2- Are advertisers showing up next to low-credibility content different from those showing up next to high-credibility content?

Q3- Are current brand safety tools adequate in allowing advertisers to shun repeat sources of misinformation?

The online advertising ecosystem is highly complex. Taking this complexity as its starting point, this research takes a breadth-first approach, using a range of methods to study various aspects of the phenomenon, which are detailed in each of the relevant sections below. It takes a Europe-first lens, as the author organizations are based in the UK and France.

The specific approaches taken to answer the research questions, as well as the services covered in this report, are shaped by a key constraint: **access to data**. Due in part to commercial sensitivity concerns, adtech providers and social media platforms rarely disclose official data about their relationship with specific publishers, leaving researchers reliant on finding workarounds, proxies, or unofficial data to study the online advertising ecosystem.

As a result, this data availability constraint shapes the perimeter of the research (which specific phenomenon can be studied on which services and how). In

particular:

- Because most of our data harvesting relied on data donation collected through Who Targets Me's browser extension (see below), reaching sufficient volumes on mobile app-first platforms such as TikTok or Instagram was impossible,
- The specificities of how each platform's monetization system works meant that not all phenomena could be reliably estimated on each. For instance, Twitter/X does not split ad revenue directly with a given account on whose content ads appear, but ties payments to broader account activity and influence. Likewise, since ads appear on all accounts, inferring whether a given account is monetized is not possible (unlike, for instance, YouTube).

A significant part of the resources allocated to this project therefore went into data gathering, both to develop new technical tools and to conduct manual data collection.

Despite these efforts however, some cases remained in which some of the data required to run full cross-platforms comparisons for all questions of interest was simply unavailable, and no satisfactory replacement could be found. In those cases, we ran the analysis on the platforms that could be covered using data at our disposal.

Note that this analysis focuses only on monetization via ad revenue sharing systems, as it is the primary driver of revenue to online content creators, and does not look at other monetization mechanisms (affiliate marketing, subscriptions, tipping…), although those can be substantial.

II. The applicable legal framework in the EU

Digital Services Act

The Digital Services Act (DSA) creates a direct obligation for the largest adtech providers to:

- create a repository of ads shown to users in the European Union,
- ensure that any ad displayed is clearly marked as such (and displays information such as who paid for it and the parameters used to decide to target that user).

The DSA also calls for the drawing up of an Advertising Code of Conduct in 2025, which as of February 2025, has not yet been negotiated and adopted by the industry.

In addition, the DSA forces Very Large Online Platforms (which include Facebook, Instagram, YouTube, Google and Twitter/X) to identify and mitigate any systemic risks that their services pose to European societies.

Some systemic risks, such as those to civic discourse or public health, can be directly impacted by these platforms' advertising services (e.g. the creation of an attention economy thriving on polarized political discourse, paid-for political influence operations, promotion of hazardous health products...).

Code of Conduct on Disinformation

The Code of Conduct on Disinformation is a voluntary instrument drawn up by a broad array of technology companies, adtech providers, civil society organizations and fact-checkers. The Code lists a number of Commitments that actors active in the online information space can take to mitigate the spread of harmful misinformation.

In particular, the Code's Commitment 1 states that "Relevant Signatories participating in ad placements, commit to defund the dissemination of disinformation, and improve the policies and systems which determine the

eligibility of content to be monetised, the controls for monetisation and ad placement, and the data to report on the accuracy and effectiveness of controls and services around ad placements".
The relationship of the services studied in this report with the Code varies:

- Google (and hence YouTube) and Meta have withdrawn from most of the Measures under Commitment 1.
- Twitter/X left the Code in May 2023 (although it was a founding signatory).

As a Code of Conduct under the DSA, the document creates a set of best practices that signatories are expected to adhere to and which can help them mitigate the systemic risks under the DSA.

Because the European Commission considers that the Code will be a "relevant benchmark" against which platforms' risk-mitigations obligations are judged, services that decide to not sign up to the Code (or some portions thereof) could still have to prove that they have systems in place to effectively address the Commitments set out in the Code.

Regulation on the transparency and targeting of political advertising (TTPA)

In March 2024, the European Council adopted rules around political advertising, which:

- forbid non-EU entities from sponsoring political ads 3 months before an election,
- enhance the transparency obligations regarding political advertising (disclosing the sponsor, ad budget…).

Most of the rules will take effect in autumn 2025.

III. Overview of key data sources

The Who Targets Me browser extension

Since 2017, Who Targets Me's browser extension collects anonymized data on ads seen by volunteers who choose to opt-in. Originally, the extension focused on political advertising on Facebook but its focus was expanded for this project to capture all types of ads on Facebook, Instagram, Twitter/X and YouTube.

The extension collected information on more than 30,000 ads per day over a period of 5 months.

For this project, two data collection streams were used:

- Passive collection of ads seen by regular browser users, subsequently filtered to keep only ads that appeared on surfaces of interest (e.g. accounts with known credibility),
- Volunteers with the extension activated signing up to visit specific URLs of interest given by the project team, in order to go and collect the ads that appeared on specific surfaces of interest.

Automated web browsing tools

To complement the ads seen by users, automated web browsing tools simulating human users were developed to go and periodically visit surfaces of interest (Open Web URLs, posts by a given account of known credibility on a social media platform).

Account credibility datasets

A first core focus of the project was the study of the capacity of low-credibility actors to monetize across platforms (using high-credibility actors as a comparison point). A second focus was to observe whether the types of brands

whose ads appeared on high-credibility surfaces were qualitatively different from those that appeared on low-credibility surfaces.

As such, having a robust, widely-agreed, pan-European dataset of website and social media account credibility was a central methodological requirement.

Science Feedback's Consensus Credibility Scores bring together publicly-accessible credibility ratings from a broad range of raters to generate a single credibility score at the web-domain level. By drawing on dozens of sources, the Consensus Credibility Scores:

- Minimize any possible bias that would come with using just a single source,
- Maximize the coverage of websites evaluated, which, in a European context, is important as any single source tends to cover only one country or subregion.

A subset of popular domains was extracted from this list of credibility scores. In order to strike a balance between covering a large part of the European population and regional representativeness, domains from the following countries were selected: Bulgaria, France, Germany, Hungary, Poland, Spain, and the UK.

The official social media accounts belonging to these domains were extracted, resulting in a total (across 3 platforms) of 185 high-credibility accounts and 171 low-credibility accounts addressing primarily European audiences. These accounts were linked to 72 high-credibility and 79 low-credibility web domains.

Examples of high-credibility domains include: zeit.de, dw.com, wyborcza.pl, polityka.pl, lcp.fr, france24.fr, bbc.co.uk, ft.com, hvg.hu, telex.hu, elpais.com, lavanguardia.com, capital.bg, dnevnik.bg

Examples of low-credibility domains include: epochtimes.de, compact-online.de, gazetapolska.pl, dorzeczy.pl, breizh-info.fr, francesoir.fr, thepoke.co.uk, ukcolumn.org, oroszhirek.hu, orientalista.hu, buscandolaverdad.es, periodismo--alternativo.com, pik.bg, rikoshet.org.

The full list is available upon request to the study's authors.

Official publisher lists of monetized accounts

Under its brand safety suite of tools, Meta offers official "partner-publisher lists" of accounts that are signed up for monetization on Facebook and are not in violation of Meta's Partner Monetization policies. These lists are meant for advertisers that want to review the accounts on which their ads can appear and that their ad spending supports.

Ad-hoc manual research

To partially alleviate the significant data gaps left by the absence of public datasets, some manual research was conducted to collect:

- User testimonies of the revenue they were able to generate on a given platform via its ad-revenue sharing system,
- Industry, academic or civil society estimates of per-country, per-platform breakdown of the cost of advertising impressions (usually expressed as a CPM or RPM, which is the cost of revenue per thousand impressions of an ad), from which estimates of revenue to given accounts could be derived.

IV. Research

A- Do social media platforms and ad service providers limit access to ad revenue sharing to repeat spreaders of disinformation?

Some large social media platforms offer ad-revenue sharing features to some accounts, directly tying the creator's payouts to the number and price of the ads displayed on their content (while others such as TikTok under its Creator Rewards programme tie payments to broader account-level influence and activity).

Accounts must typically meet two criteria to benefit from ad-revenue sharing:

- reach some minimum size or activity threshold (such as number of followers, watch time, or number of pieces of content posted). These can usually be directly observed or reasonably inferred using publicly-available data.
- be in good standing vis-a-vis the platform's community standards, including their misinformation policy. This can usually not be directly observed using publicly-available data.

Assuming similar levels of willingness-to-monetize across both high- and low-credibility groups, comparing the **proportion of likely-eligible high-credibility accounts that are able to monetize** to that **of likely-eligible low-credibility accounts that are able to monetize** allows us to get some insights into whether platforms' systems work to prevent repeat misinformers from benefiting from ad-revenue sharing.

We find that, across platforms and ad services:

- Approximately half of low-credibility accounts or domains are benefitting from ad-revenue sharing,
- Low-credibility accounts or domains are somewhat less able to monetize their content than high-credibility accounts, suggesting that content credibility could factor in their inability, albeit to a limited extent.

Importantly, two main limitations should be addressed:

- Should low-credibility content creators know that a platform has strict policies in place regarding the monetization of misinformation, they might decide to invest less into growing their presence on this platform and, consequently, not meet the size requirements. This seems however unlikely given the low thresholds used by platforms (see Methodological details).

- Similarly, should low-credibility accounts be largely banned from the platform, this data would not capture this as it only looks at active accounts.

Platform / service	High-cred accounts / domains meeting observable criteria for monetization	High-cred accounts / domains monetized (% of eligible accounts)	Low-cred accounts / domains meeting observable criteria for monetization	Low-cred accounts / domains monetized (% of eligible accounts)
Facebook	41	28 (68.3%)	12	5 (41.7%)
YouTube	44	33 (75.0%)	18	10 (55.6%)
Google Display Ads	72	57 (79.2%)	79	38 (48.1%)

Table 1 *– Summary table of the proportion of high- and low-credibility accounts with an active monetization relationship with each service.*

Methodological details

Facebook

135 Facebook Pages with a European focus were included in the credibility dataset (72 high-credibility, 61 low-credibility). Each was evaluated for the publicly-observable audience- and activity-related eligibility criteria (1, 2, 3) or reasonable proxies for monetization on Facebook's in-stream videos or Reels:

- 5,000 followers,

- 60,000 watched minutes over the previous 60 days. Since this is not directly observable, the length (in minutes) of videos posted by the Page was multiplied by their number of views, and divided by 2 to account for mid-view drops.
- Page owner is based in a country where monetization is available (all countries in the sample except Hungary and Bulgaria).

This left as the main non-observed criteria the Page's standing vis-a-vis Facebook's community standards and content monetization guidelines, which include provisions related to sensitive topics and misinformation.

Each account was then searched to see if it appeared in the official Partner-Publisher lists of accounts that are eligible for monetization.

Examples of low-credibility accounts monetized on Facebook as of February 2025: OKDiario, the Sun, Deutschland Kurier.

YouTube

92 YouTube channels with a European focus were included in the credibility dataset (49 high-credibility, 43 low-credibility). Unlike Facebook, YouTube does not publish official lists of which channels are benefiting from ad revenue sharing.

As a result, we adapted the two-step process used for Facebook. First, we screened the channels to see which were eligible for monetization on the basis of publicly-observable criteria (ie, criteria not related to the channel's or content's standing vis-a-vis YouTube's community guidelines). These criteria are:

- Whether the channel has more than 1,000 subscribers,
- Whether the channel has more than 4,000 hours of watch time over the last 12 months on its videos (excl. Shorts). As this is not directly observable, we made the same assumption as for Facebook (sum of number of views on videos posted in the last 12 months multiplied by the video length, divided by 2 to account for view drops).

This first step resulted in 62 channels (44 high-credibility, 18 low-credibility) that met the audience criteria for monetization. We then checked the effective monetization status of each, to test the hypothesis that low-credibility YouTube channels would be faced with some barriers to monetization under the YouTube channel monetization policies, using higher-credibility channels as a control group.

Because the monetization status of the channel is not available from official databases, we collected the last ten videos published by the account and observed how many of these videos had ads (either before or during the video playing, or in the top-right-hand corner of the video). Any channel with three or more videos displaying ads was considered monetized.

Examples of low-credibility channels monetized on YouTube as of February 2025: CompactTV, Omerta, Blitz BG.

Google Display Ads (open web domains)

Unlike Facebook and YouTube, Google Display Ads do not set a traffic threshold for web domains to be eligible to have ads served by Google displayed on their website, leaving the partner website's standing vis-a-vis the Google Publisher Policies as the main explanatory variable.

Each of the 151 websites in the credibility dataset (72 high-credibility, 79 low-credibility) were visited by a human analyst. Up to five pages (randomly accessed from the homepage) were visited by the analyst. If the analyst saw ads served by one of Google's services (doubleclick.net or googlesyndication.com), the website was marked as being monetized by Google. If no such ads were found on the 5 pages, the website was marked as not being monetized by Google.

Examples of low-credibility websites monetizing with Google Display ads as of February 2025: oroszhirek.hu, epochtimes.fr, gazetapolska.pl.

B- Estimating ad revenue flowing to low-credibility accounts

We explored above the extent to which a binary 'monetized'/'not monetized' status for each account of interest in its relationship with a given adtech provider (incl. social media platforms) could be observed.

Going one level deeper, we aim to estimate the extent to which ad placement services are incentivizing disinformation actors by attempting to put a monetary value on the ad-revenue income of some accounts.

Our research did not identify any official database or other way of estimating the ad revenue of a given social media account or website, for any of the ad tech providers surveyed (Facebook, Twitter/X, YouTube, Google Display Ads).

Consequently, we attempted to collect data on how much revenue an account could hope to generate through ad revenue data, for each platform in each country of interest. Our research covered three types of sources:

- User testimonies posting proof of their ad revenue intake on social media (e.g. on Reddit, some subreddits are dedicated to exchange of tips on monetizing YouTube channels, with users frequently sharing screenshots of their income for the previous month). The main country of origin of the account's audience and the main topic covered by the account were recorded.
- Existing aggregated reports or other industry estimates offering EU-country level data,
- Simulating ad campaigns using the platforms' official advertising tools to collect campaign reach estimates as a function of budget and country targeted, which can then be turned into ad revenue flows.

Sources of data

By combining these independent data sources, we aim to provide realistic estimates of the lower and upper-bound of ad revenue flows to a given account on the basis of its audience size and country of origin.

1. User testimonies

For each of the 7 countries of interest (Bulgaria, France, Germany, Hungary, Poland, Spain, and the UK) and each platform of interest (YouTube and Facebook, as these were the platforms were the ad-revenue sharing system makes a direct connection between an ad running and a payment to the content creator), searches were conducted on Reddit, Google and YouTube. Three queries were run on each:

- "How much money did I make on [platform of interest]", translated into the local language, and, where applicable, with a geographic filter or mention of the country to ensure results were geographically-relevant.
- "[country] RPM [platform of interest]"
- "[country] CPM [platform of interest]"

Results of the searches were manually reviewed (including with the help of translation tools when necessary) to identify posts that were discussing the advertising revenues that the accounts controlled by the user were able to generate over a given period. The account's main topic (e.g. news commentary, cars, video games...) was also collected.

When directly provided by the user, the number of views corresponding to the payment was recorded.

In some cases, no proof of the user's earnings was shared. In those cases, the data collected, although broadly indicative, cannot be considered hard evidence.

2. Aggregated 3rd-party estimates

Some actors specialized in digital marketing have published research or aggregated data that can be helpful to estimate the value of an audience in various countries for a given platform.

- *Ebiquity*

Ebiquity is a consultancy specialized in helping brands maximize the value of their media investments, including online advertising. As a signatory of the EU's Code of Conduct on Disinformation, Ebiquity provides country-level estimates of CPM for various ad types, which can provide a basis to estimate the revenue going to an advertiser (the underlying data is not public, enquiries can be made with the study's authors regarding methodology to arrive at the account-level revenue estimation).

- *Metricool*

Metricool operates a platform used to manage brands' presence across social media platforms as well as advertisers' campaigns. With more than two million users, Metricool is able to collect vast amounts of data on social media content across a number of countries, which are then aggregated into research reports (e.g. here).

- *Is This Channel Monetized*

Is This Channel Monetized is a blog dedicated to content strategies and other monetization-related tools on YouTube. The owner of the blog operates a number of YouTube channels, which are used to publish data, including on CPM per country.

The CPM per country data is limited, as it only stems from observations of the country-level CPM on only one of the channels operated by the website owner. In addition, this data is not externally auditable.

3. Collecting reach estimates using simulated advertising campaigns

We also explored the simulation of ad campaigns using the official adtech provider tool. By setting a budget and choosing to display the ad only on specific accounts of interest on the platform, we aimed to obtain useful estimates of campaign reach that could be used to calculate a CPM for the campaign and an RPM for the account. Unfortunately, none of the services offered sufficiently targeted account-level estimates.

For Facebook and YouTube, we were able to derive estimated content creator revenue by running campaigns with set budgets targeting entire countries (and not just individual accounts). We then derived estimated revenue going to the creator by using ad revenue sharing splits:

- For Facebook, a mock campaign was created on a €50 daily budget for each country with a goal of maximizing brand awareness via a video ad (to reflect the fact that ad-revenue sharing happens mostly on video surfaces). In line with (now discontinued) ad revenue sharing agreements on video on Facebook, 55% of the CPM was estimated to accrue to the account owner. As Facebook provides both a lower-end and higher-end reach estimate for the campaign, both were recorded.

- For YouTube, a mock campaign was created with a €200 budget for two weeks for each country. The objective of the campaign was set to "Awareness and consideration", using "Video" ad type, and "Video reach" subtype and "Efficient reach" as the preferred way of operating. For each country, YouTube offers a "suggested CPM bid", which we took as the baseline scenario. Because YouTube offers a 55-45 ad revenue sharing agreement, we took as a revenue per mille figure 55% of the suggested CPM for each country.

Results

Unsurprisingly, the data collected displays great in-country variability for each platform (see below), to the point that no reliable estimate of account-specific ad-sharing revenue flows can be generated using only publicly-available data.

We therefore conclude that any meaningful study of this phenomenon can only be achieved by using platforms' internal data, which academic researchers should be able to request once DSA Article 40 is fully operational.

Figure 1 *– Revenue Per Mille (€ per thousand views on monetization-eligible content) data points by country on YouTube*

Figure 2 *– Revenue Per Mille (€ per thousand views on monetization-eligible content) data points by country on Facebook*

C- Auditing brand safety tools: how do publishers, platforms and advertisers think about online advertising brand safety?

The size of the global market for online advertising (estimated at $535bn in 2023) brings about a wide range of problems. Ad fraud is notably rampant: the World Federation of Advertisers previously estimated that in 2025, $50 bn - around 7% of predicted global spend - is likely to be lost to fake websites, fake traffic, and other fraudulent activities. Per the WFA, ad fraud was already 'second only to the drug trade as a source of income for organised crime' in 2016.

Over the past couple of years, ad fraud has been given a boost by the advent of so-called "Generative AI", which has accelerated the creation of new ad fraud schemes, according to DoubleVerify, an ad verification and data firm. The company claims Gen AI enables malicious actors to rapidly generate user agents to mimic human behaviour, making bot traffic patterns more difficult to detect.

These issues are symptoms of the deeper problem at the heart of online advertising: opacity. The UK lobby vehicle for advertisers, ISBA, produces a regular audit of the industry's supply chain, known as the 'Programmatic Supply Chain Transparency Study'. Produced in partnership with professional services firm PwC, the first study, published in 2020, found just 12% of claimed impressions could be matched between advertisers and publishers, and 15% of total programmatic spending was 'unattributable', a figure that became known as the 'unknown delta'.

The opacity and consequent lack of trust in advertising supply chains has generated an industry of ad verification solutions.

In the platforms' shoes

The Very Large Platforms and Search Engines have contributed to the internet developing towards an attention economy where engagement is the key metric.

Mark Zuckerberg recognised the danger of this in a key 2018 note on how Facebook was trying to address content governance:

"One of the biggest issues social networks face is that, when left unchecked, people will engage disproportionately with more sensationalist and provocative content."

This idea was vividly illustrated by the following famous graph, showing that as content became more 'borderline', it generated more engagement. While Zuckerberg pointed out that this was not a new phenomenon, noting the existence of tabloid newspapers and cable TV news, he did recognise that the scale of social media meant that social polarisation was a more likely potential consequence.

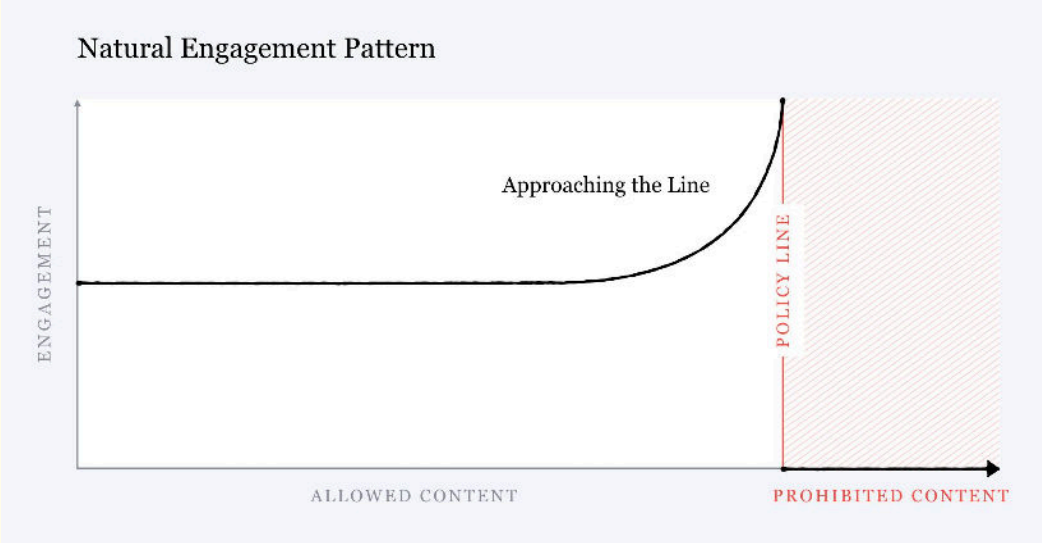

Figure 3 *– "Natural Engagement Pattern" from Facebook's Blueprint for Content Governance and Enforcement*

The platforms are now in the position of being power-brokers between publishers and advertisers as they own the mechanisms and channels through which ads can be seen by large audiences.

They have successfully inserted themselves into a buying process in a way that has destroyed newspaper revenue and is now increasingly taking market share from traditional audiovisual channels (chiefly TV buys). The chart below, from

WARC Media, a leading marketing and strategy agency, says it all, as does this quote:

"Content media owners (including all forms of TV, publishing, audio and cinema) are forecast to receive only 27.2% of all global ad investment in 2024, down from 71.0% a decade earlier."

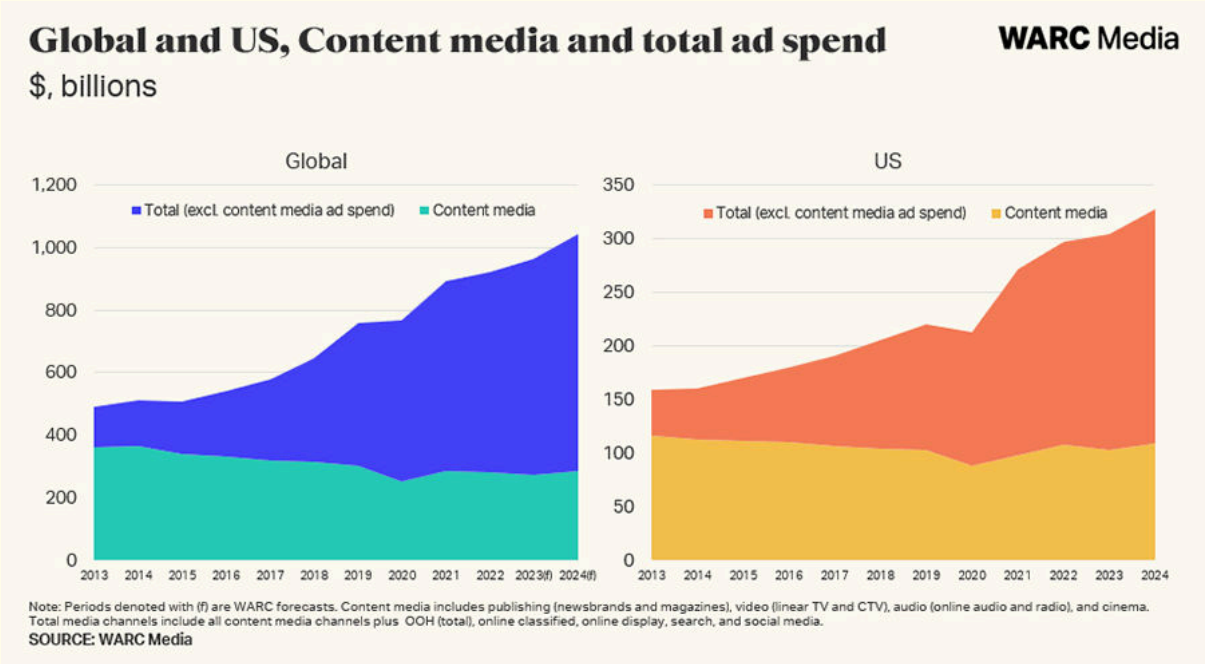

Figure 4 *– Global and US ad spend on 'traditional' content (print, TV, radio) and in total, including online ad spend (source: WARC Media)*

As the advertising landscape changes, so do brand safety measures. Traditionally, advertisers could ensure their content was placed correctly by favouring high-trust channels. This is how the partnership between advertising and journalism originally arose: advertisers could trust that their brand would be associated with high-quality information read by a wide audience. Meanwhile, newspaper publishers could be discerning about the brands they were willing to carry.

Governments and regulators around the world have moved slowly to address some of this change. Because ads increasingly run in environments that do not require professionally-produced content, there are long-term questions about the viability not just of written journalism but TV news journalism. The knock-on

effects for civil society organisations, political parties, and other important democratic actors are troubling, as citizens cannot ultimately make positive and accurate decisions in a confused information landscape where reality and ground-truth themselves are increasingly contested. No wonder, then, that we've seen in recent years the rise of several civil society pressure groups dedicated to embarrassing the full programmatic supply chain by pointing out instances of unsuitable content.

These groups include the likes of Check My Ads, the Conscious Ad Network, Stop Funding Hate, and the Center for Countering Digital Hate. Some have added significant complexity to the concept of 'brand safety' by loudly and publicly questioning whether advertisers *really* want their brand represented in the pages of tabloid newspapers, or next to video footage of war, violence or misogyny. This has placed further pressure on publishers, too, to consider how the automation of the ad buying process creates a plethora of new risks for their own credibility - like in the very topical example below.

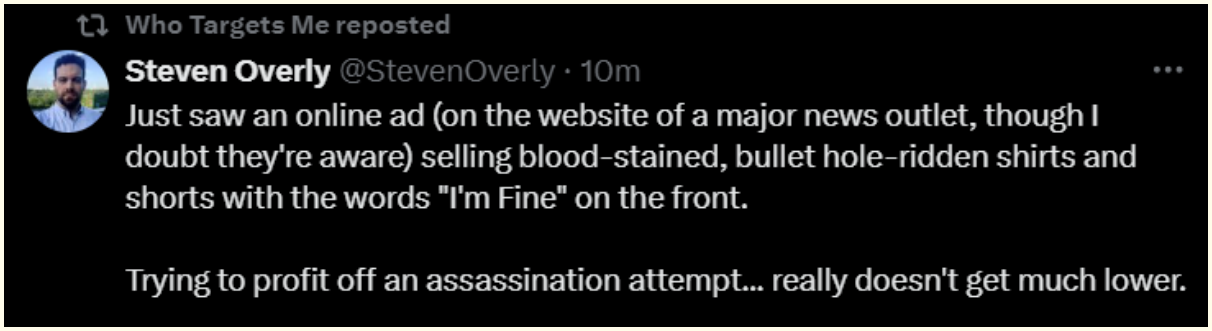

***Image 1** – Commentary on inadequacies in brand safety measures - (source: Steven Overly)*

A crucial aspect of that example is the point that 'I doubt they're aware'. Because programmatic advertising happens pretty much instantly, it requires all parties in the process to do their own audits. But that in turn requires a level of resourcing and social responsibility beyond most marketing teams' budgets or boards' calculation of brand risk. This is how we end up with car companies and pharma giants sponsoring political streamers calling for the beheading of public figures, or reproductive healthcare nonprofits finding their ads next to articles peddling dangerous herbal abortion recipes.

Solutions are only partial (and it's hard to be impartial)

The platforms that created the advertising supply chain ultimately have the most responsibility for its end-products and its impact. They have an obvious financial incentive to get it right, regardless of where their owners lie on the political spectrum - an untrustworthy ecosystem, one that creates an ongoing stream of controversies for platforms and their advertiser clients - is not in their long term interest.

This is mirrored by regular research showing that most advertisers (75% in this 2022 study) want more oversight and control over where their ads are placed, and who profits from those placements. Those advertisers are also under pressure from journalism publishers, who are concerned about an unequal playing field for news vs. social media placements.

As such, platforms have responded to this pressure from governments, regulators and citizens - as well as the publishers they have relied upon for high quality content - by seeking to create a new set of brand safety tools for advertisers.

These tools have become increasingly sophisticated over time. This is a reflection of the sheer quantity of "unadvertisable content" on the internet. But that in turn requires both platforms and advertisers to commit resources to ensure 'brand suitability' and 'brand safety'.

Typical brand safety and suitability tools

To give advertisers control, the major online advertising platforms all currently offer a variety of mechanisms to give their clients confidence that ads will not run against or next to content they might find problematic, or that detracts from their perceived "value" as brands.

These tools include:

- **Publisher lists:** visibility into a regularly updated list of partners and publishers where ads can appear (usually millions of websites).
- **Block lists:** which allow advertisers to exclude specific URLs and partners and publishers from running their ads

- **Allow lists:** the opposite of a block list, this involves uploading a list of approved publishers which ads will run against
- **Content exclusions:** allowing advertisers to exclude specific types of content - for example, live video (which has been a huge issue during live shooting or terrorism incidents), specific keywords, or to apply broad topic exclusions like 'gaming', 'politics' or 'religion'
- **Inventory filtering:** a way to exclude certain types of sensitive content from ad placements. It relies both on a constantly updated list of banned and excluded content, and on a rapid review process for sensitive or misleading content that often draws on third-party fact-checking
- **Delivery reports:** designed to give the advertiser transparency on where ads are placed, enhancing their confidence in the process

Ad funded platforms where organic content is crucial for engagement tend to offer a combination of the above. They might also seek industry recognition as a way to build further trust, for example by obtaining accreditation from the Media Rating Council or, until recently, a coalition such as the Global Alliance for Responsible Media (GARM)[2]. These partnerships and coalitions tend to involve independent auditing of processes and progress against brand safety guidelines, though some argue that these sorts of arrangements are more of a PR fiction than actually protective of brands and users.

Brand safety functionalities are often powered by companies behind the big platform brands, which are unlikely to be recognised by end users. These include services such as DoubleVerify, Zefr and Integral Ad Science. All offer a variety of tools, levels of configurability, integration with other services and platforms. They aren't pitched as direct-to-advertiser, or even agency services, instead operating as a layer underpinning the brand safety offerings of advertising services.

While all toolsets, whether platform or brand-safety-as-a-service, promise a high degree of customisation for advertisers, they all lean heavily on defaults, which essentially fall into three buckets:

[2] GARM's parent organization, the World Federation of Advertisers announced in August 2024 that it would be shutting down GARM's activities, in response to a lawsuit from Elon Musk's X.

- **Low** (or “maximum inventory”): a setting that allows ads to run against most content, including topics such as politics or religion.
- **Medium** (or “standard inventory”): where ads are “less likely” (note the intentional lack of guarantee) to run against content that contains strong swearing, sexual or violent content.
- **High** (or “limited inventory”): which is similar to medium (again, no warranty), with a filter on “moderate” swearing, sexual or violent content.

Google and TikTok’s self-service ad platforms default to a “medium” setting, while Meta’s is closer to the “low” option, requiring advertisers to tweak settings to find the right combination for them.

If advertisers wish to use the tools provided by platforms to exclude their ads appearing on certain sites or accounts, they are able to create exclusion lists, but the practicalities of this do seem formidable. For example, Meta has over 2.5 million “publisher-partners” and a simple keyword search will often return thousands of results, making blocking ads at the per-website/page level daunting to the point of being impossible. Essentially, these allow advertisers or agencies to block a handful of sites that they might not want ads to run against, but they aren’t likely to be at all comprehensive in keeping ads away from particular types of content.

It is also worth noting that platform brand safety tools are not available everywhere, with smaller countries and languages (including many EU languages) un- or inconsistently supported. This means that while English, Spanish, French or Arabic- speaking users and brands do get some level of protection, Hungarian, Bulgarian, Romanian (and so on) users get a much lower level of protection, or none at all.

Once ads have run, all major services offer reports showing where ads were placed. These typically include the reach/impressions of the ads, placements (both sites and positions, such as “Facebook Feed”), type of device and so on. These act as the effective “receipt” for the advertising the client has bought. As we will see below though, these reports are more usually used to manage advertising performance rather than monitor and improve brand safety.

What do advertising professionals think about platforms' brand safety offerings?

To understand how effective brand safety tools are perceived to be, we interviewed (off the record, to protect their clients and brands) three digital advertising professionals (two from agencies, one "client side") on their views on brand safety for advertisers.

The agencies recognised that their clients don't want their brands to be associated with or be accused of funding problematic, violent or deceptive content. At the same time, they acknowledged they have relatively little ability to directly control this, beyond the tools offered by advertising services.

One of the agencies we spoke with said they felt confident they could report to clients that their ads were "99%" brand safe, and they would never fall below this level, though found it difficult to explain how this figure was arrived at, and knew they had no capacity to truly audit where ads were actually appearing and revise their blocklists and policies accordingly. There are simply too many data points to evaluate themselves. They also recognised that they had even less control or knowledge when it came to the content ads appearing in social media feeds would be placed next to. Finally, they felt they simply had to accept that the advertising platforms currently offer no standards or guarantees about the quality of the filters and tools they offer, nor any warranty should they fail.

In terms of the practicalities of placing "brand safe" ads, the agencies reported that they would generally err on the side of safety over reach. This means that clients' ads might be seen by slightly fewer people, and that the cost of advertising would be slightly higher, but with the comfort blanket that the risk of ads showing next to controversial content would be lower. Asked if clients ever chose "more reach" over safety, the representatives we spoke with couldn't think of an example. As such, agencies simply accept the default set of brand safety options, and apply these generally across all of their clients, rather than tweaking them on a case by case basis.

The challenge with erring on the side of safety is the risk of demonetising legitimate content. Controversial topics are worthy of monetisation (for

example, providing sexual health information), but might be flagged as inappropriate for advertising. No one we spoke with raised this trade off.

One of the agencies, part of a global advertising network, said there was someone in their parent company who maintained a block list of websites, and that this was used, but there wasn't a lot of interaction between the day-to-day management of a client's ads, any specific wishes about brand safety and the content of this list.

After ads have run, once a campaign is over, both agencies said they include "brand safety" scores and benchmarks in the standard campaign performance reports presented to their clients. When we asked if these figures were ever then raised or questioned by the client, we were told that they were not.

The only conclusion we can draw from these conversations was that brand safety was seen as a requirement, but low in terms of the day to day priorities of advertisers and agencies. "Some" was good enough.

The challenges of "brand safety" laid bare

This review of brand safety finds three clear problems.

First, the complexity of brand safety tooling and the power of defaults means that most advertisers simply accept the options available to them, barely exploring the options available, let alone doing their own research and finding the right level of brand safety for them. This makes it less likely that advertisers will take specific steps to dissociate their content from known disinformation.

Second, clients' interest in brand safety is shallow, not deep. They want to be "brand safe", but most likely to avoid controversy, negative associations and potential blowback. This is not the same as wanting brand safety to reduce the quantity of low quality monetised content online. Again, this serves as another reason to avoid deeper investment of time, money and thought in brand safety.

Third, at no point are potential trade offs with "brand safety" considered by any of the providers. The assumption is that, by keeping your content away from

certain types of content, you'll avoid controversy. However, false positives and over-definition can also have democratic consequences.

The outcome is that while digital advertising is an industry worth hundreds of billions of dollars, and brand safety initiatives are very much a tacked on afterthought, with no actor in the supply chain truly incentivised or motivated to take on the full cost of demonetising problematic content.

D- Which ads appear next to disinformation content ?

Another aspect of our study was to compare the type of ads that showed on low-credibility vs high-credibility surfaces, to test the hypothesis that major brands with a well-established reputation might be less likely to appear next to low-credibility content (either because of a possible higher sensitivity to reputational damage or simply because of higher proficiency in using brand safety tools).

Because of a mix of technical limitations in the ads captured and platform design choices relative to how ads are shown, we were able to collect ads on surfaces with an associated credibility rating on YouTube and Twitter/X. Ads captured on Facebook appeared in users' feeds and therefore not on specific accounts to which we could attach a credibility.

YouTube

13,307 ad impressions (4,241 unique ads) were collected on videos posted by channels for which we had a credibility rating. Each ad was inspected and coded to identify the types of ads promoting services that were considered as presenting a high risk of being a scam or otherwise objectionable content. The types of ads identified as such belong to the following categories:

- Ads for trading services (most frequently, cryptocurrency) promising a guaranteed implausibly high return on investment, passive income, or training packages guaranteeing such results,
- Ads for health supplements or get-healthy-effortlessly products and services,
- Ads for investments in foreign real estate,
- Ads for dating or mail-order bride services,

- Ads for clairvoyance services,
- Ads for services promising visa or migration regulatory assistance,
- Clickbait ('If you are a senior citizen living in France, this could change your life!')
- Ads promising tax rebates.

Although some of these ads might run against Google Ads' content guidelines, looking into this aspect was outside the scope of this study.

Looking at the distribution of such ads across credibility categories, we hypothesized that low-quality advertisers would show up less often on high-credibility channels, as 'standard' advertisers would opt for stronger brand safety protection and partially drive lower-quality ads to the low-credibility channels.

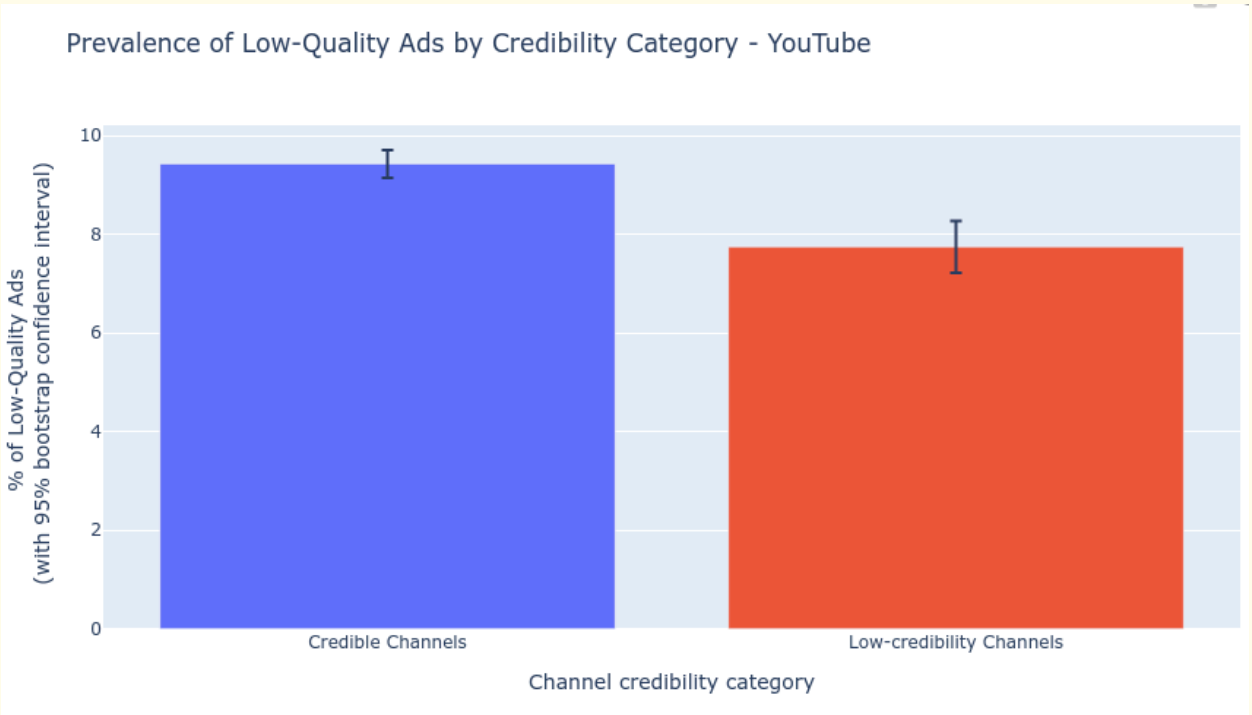

Figure 5 *– Proportion of low-quality ads that appear on channels of high and low credibility on YouTube*

However, the data does not support this hypothesis. To the contrary, low-quality ads make up a larger proportion of ad impressions on high-credibility channels than on low-credibility ones, suggesting that even

advertisers from reputable brands are either unwilling or unable to prevent their ads from showing on and funding low-credibility content.

Twitter/X

46,607 ad impressions (4,814 unique ads) were collected in the discussion threads under tweets for which we knew the posting account's credibility. As for YouTube, each ad was coded as to whether it was promoting a service considered as presenting a high risk of being a scam or otherwise objectionable content. Those marked as such were:

- Cryptocurrency-related and other get-rich-quick schemes (including trading courses and pyramid schemes),
- Dating services,
- Free giveaways if one enters a lottery,
- Products baselessly pretending to offer implausibly high energy savings,
- Clairvoyance,
- Lockpicking products.

We then also looked at whether ads that appeared on the response threads to tweets posted by low-credibility accounts were more likely to be low-quality. As for YouTube, we do not observe such a phenomenon.

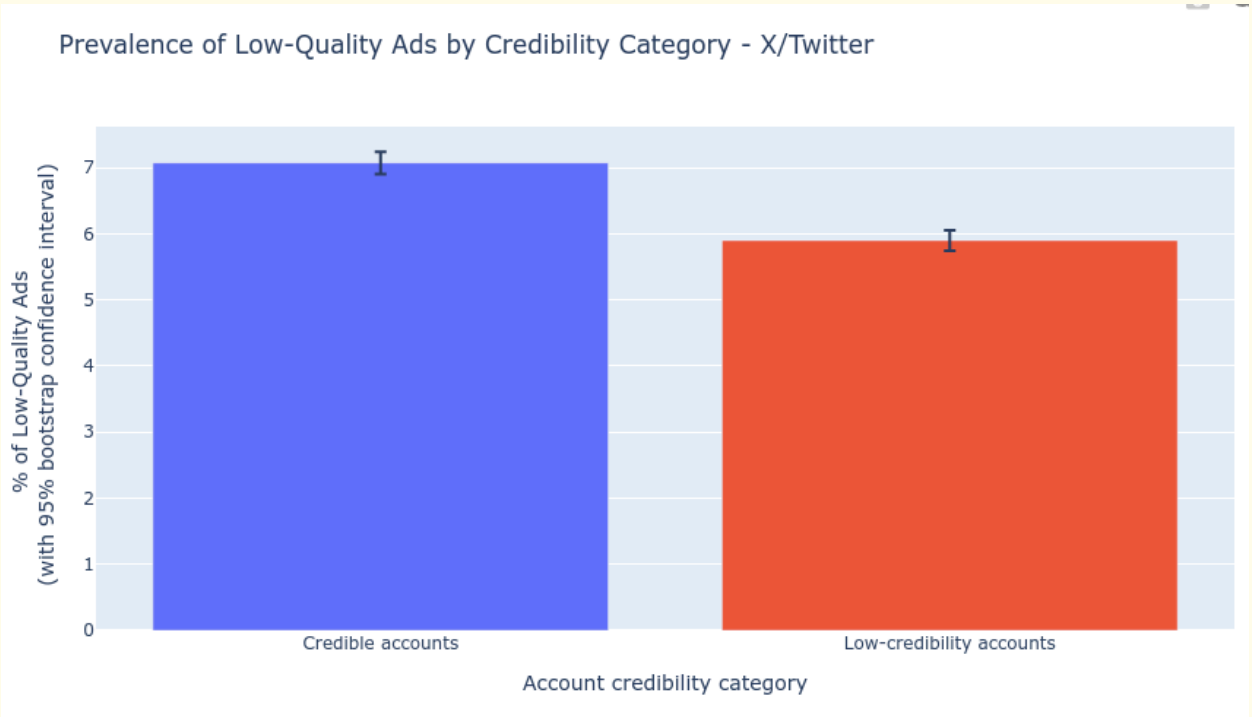

Figure 6 *– Proportion of low-quality ads that appear on accounts of high and low credibility on Twitter/X*

Ad repositories and transparency obligations

Testing ad repositories

Under the EU's Digital Services Act, each Very Large Online Platform has an obligation to create a public, searchable repository of ads shown to users, in order to increase transparency in the often-opaque adtech market. As such, we audited whether these repositories worked well, expecting that each of the ads collected throughout the course of this project would end up on the corresponding platforms' ads repository.

Twitter/X

Twitter/X theoretically offers two ways of accessing ads data (the Twitter Ads API only gives access to an account's own ads campaigns, not to those of third parties):

- A csv list of all ads that appeared on the platform, accessible as a file download from the official ads repository user interface. Although the file contains over 15 million ads, all are marked as having been posted in the month of December 2023 and none of them matched the ad IDs and URLs collected using the WTM browser extension in Q4 2024-Q12025, suggesting that the data in the file is indeed outdated.

- Its ads repository search user interface. Since the search interface only allows for a username-based search, attempts at searching for the usernames of advertisers in our database of ads collected with the WTM extension were conducted.

  A first attempt to cross-reference our data with that accessible through the search interface was conducted on Feb. 26, 2024. Regardless of the parameter input, we received an error message ("Unable to create your report - Please try again"), suggesting a technical issue on the Twitter/X side. Twitter/X was made aware of the issue on Feb 26.

  On Feb. 27, an analyst was able to create one search report. The system took over three minutes to generate a 13-row csv file of the ads created

by a single advertiser over a 3-month period. The ad collected by the WTM extension did appear among the ads returned by the ads repository.

Subsequent requests (on Feb. 27 and 28 and March 5) ran into the same bug as described above.

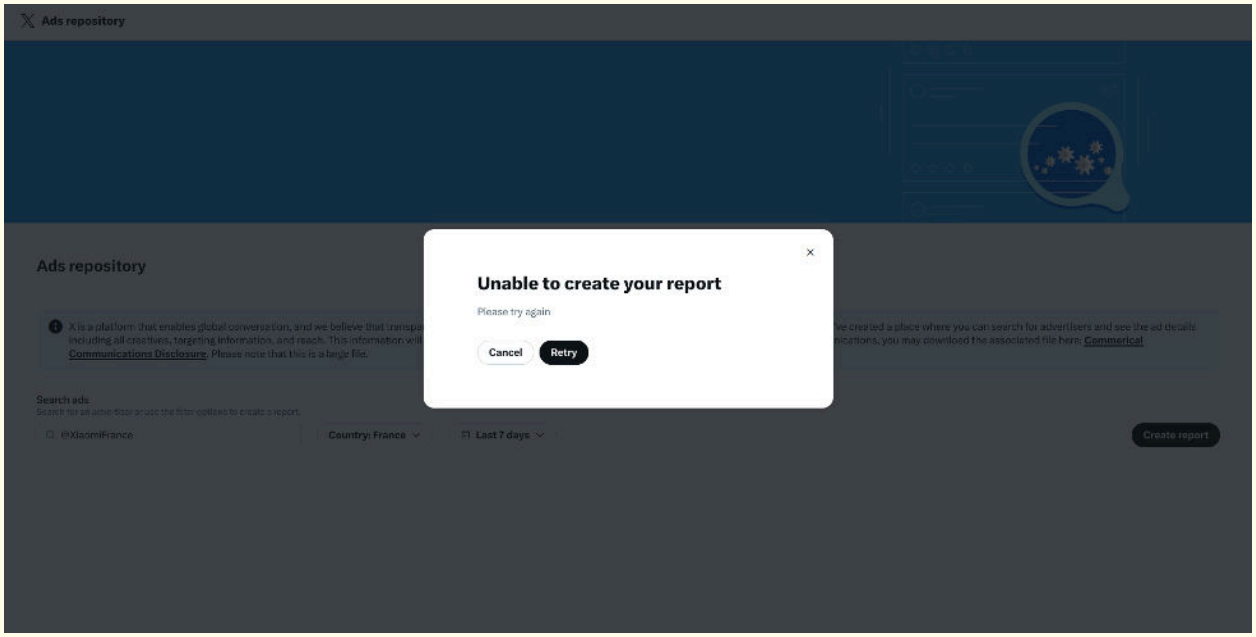

***Image 2** – Error message when trying to use Twitter/X's ad library.*

Overall, practical constraints on the availability of the data therefore made it impossible to audit at any significant scale the quality of the data contained by the Twitter/X ads repository. Our experience echoes the European Commission's findings on Twitter/X being in breach of its DSA requirements when it comes to the usability of its ads repository.

Facebook

A random sample of 200 ads shown to Facebook users was drawn. Each ad was then looked up in the Facebook ad library to ensure that all ads actually shown to users would end up in the publicly-accessible database. One of the 200 did not (1), showing an error message stating "the most common reason is that the ad hasn't received any impressions yet". Because the ad was picked up by the WTM extension, it had unequivocally received impressions, making it unclear why it did not appear in the ad library.

The remaining 199 ads could be found in the ad library.

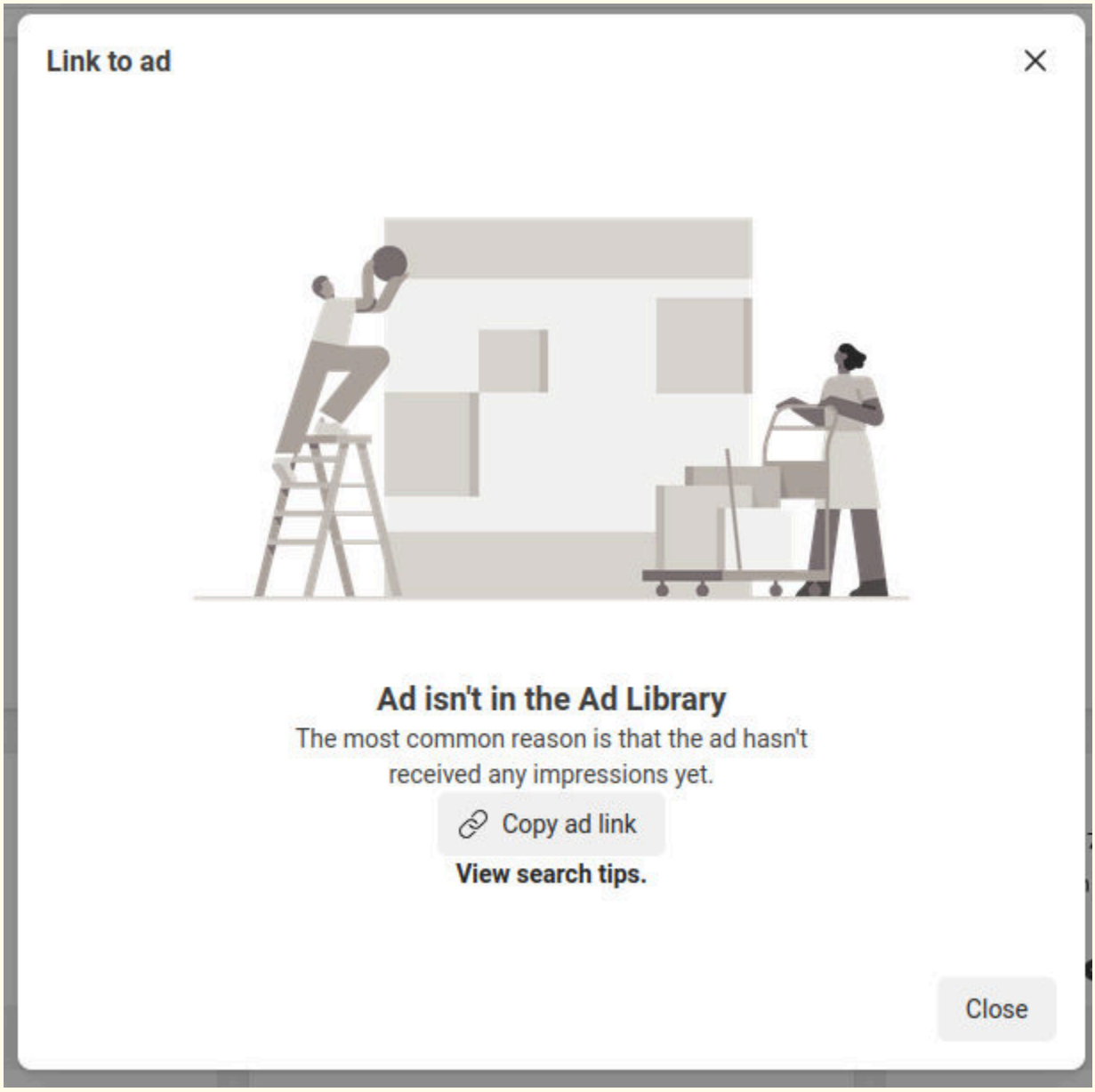

Image 3 *– Error message when trying to find this ad in the Facebook ad library.*

YouTube

Likewise, a random sample of 200 ads captured by the WTM extension on YouTube were inspected to ensure they could be found on the ad repository. All were found.

Are political ads adequately labeled ?

Facebook

Meta requires labeling for ads that are about an election or "any social issue" (as of February 2025, the list of social issues for the EU includes "Civil and social rights, Crime, Economy, Environmental politics, Health, Immigration, Political values and governance, Security and foreign policy").

30,000 ads were extracted at random from the sample of ads shown in Germany during the month of January 2025 (close to the federal elections which took place on Feb. 23, 2025 and hence presumably offering a higher density of political advertising than other periods and countries).

To provide a first filter, each ad text was then submitted to Mistral's ministral-8b Large Language Model with the following prompt: "I am going to give you the text of an advertisement, most likely in German. I need your help to assess whether the ad talks about a political subject. Return a political score between 0 to 10 - 0 is when the ad has nothing to do with politics, 10 is when the ad is definitely about a political topic."

Ads with a score of 8 or greater were shortlisted for further scrutiny, yielding 320 unique ads.

182 (56.9%) ads out of those were published by Pages belonging to political parties or candidates. All were properly marked as being about a 'political or social issue', suggesting that no major political advertiser on Facebook was able or willing to avoid being labeled.

The remaining 138 ads were marked by the LLM as being about a political topic, but were not posted by an account belonging to a politician or political party. Each was subject to analyst scrutiny to assess whether the content was indeed about a 'political or social' issue and properly labeled as such. In our assessment, 93 of the 138 ads were indeed about a political, electoral or social issue topic (the remainder being false positives).

Out of those 93, 70 (75.3%) were labeled by Facebook as 'political or social issue' ads and 22 (23.7 %) were unlabeled, despite discussing the election or one of the social issue topics listed by Meta. One ad is unaccounted for as the ad library has no record of it being served to users, showing the same error message as described in the section above.

Examples of unlabeled ads include ads calling for support to Ukraine, ads selling keychains supposedly made from the remains of Russian tanks that were "heroically destroyed" by Ukraine, or a whitepaper on electric mobility.

Interestingly, one ad was seen by 3.2 million accounts but was not accessible from the Facebook ad library anymore because its sponsor had been suspended for breaching Meta's advertising standards. Because we were able to save the ad's text content before it was deleted, we know that it was political in nature ("Sahra Wagenknecht wusste nicht, dass die Kamera noch aufnahm... ist das das Ende ihrer Karriere?" / "Sahra Wagenknecht didn't know that the camera was still recording... is this the end of her career?").

These findings suggest that political ads from non-official accounts are able to circumvent Meta's labeling systems to a significant extent, aligning with previous studies that found major lacks on political advertising labeling on Meta services (1, 2).

Twitter/X

Germany is not among the countries for which Twitter/X allows political content or political campaign ads. In Twitter/X's definition, political content refers to "ads that reference a candidate, political party, elected or appointed government official, election, referendum, ballot measure, legislation, regulation, directive, or judicial outcome" while political campaigns ads "advocate for or against, appeal directly for votes, or solicit financial support for" a given candidate, party or election.

We extracted all ads seen by WTM users from Germany on or after January 1st, 2025, yielding 10,300 unique ads (accounting for a total of over 276,000 impressions in our data).

Each unique ad was sent to the same ministerial-8b prompt described in the Facebook section above.

442 ads that received a score of 8 or greater were isolated for further human analysis, to assess the effectiveness of the political content and political campaign bans in Germany. Of those, 68 were unambiguously political as they addressed the upcoming election (e.g. calling for voting for a given candidate, ads by political parties, attacks on a political opponent…). The remainder were largely about important societal issues that played a role in the election (e.g. migration, Germany's economic conditions, the war in Ukraine, …) but, out of an abundance of caution, were left out of the analysis so as to focus only on the most clear-cut cases.

Those 68 political campaigning and political content ads appear to evidently run contrary to Twitter/X's policy to not allow them in Germany. These included ads run by official accounts of political parties (e.g. local sections of the Free Voters and AfD parties). No ready explanation was found as to why they were shown on the service.

In addition, a few political ads appeared to come from foreign-based actors, raising questions around foreign interference in German political processes. For instance, one tweet that was boosted through ads and viewed a total of 351.6K times was published by a blue-checkmarked account which appears to belong to a Moscow-based RT DE editor. Another advertiser, which pretended to belong to a Texas-based bookshop, published ads criticizing then-chancellor and leading SPD candidate Olaf Scholz for his support of Ukraine. The account has since been deactivated and cannot be found in the Twitter ad repository's search bar.

YouTube

Google's policy as of February 2025 (due to change, see below) is to allow election-related ads on its services, including YouTube. Specifically, in the EU, Google defines election ads as those that "feature any of the following:

- A political party, current elected officeholder, or candidate for the EU Parliament;

- A political party, current officeholder, or candidate for an elected national office within an EU member state. Examples include members of a national parliament and presidents that are directly elected; or
- A referendum question up for vote, a referendum campaign group, or a call to vote related to a national referendum or a state or provincial referendum on sovereignty."

Unlike Meta, Google has a narrow definition of political advertising, leaving out broader social and political issues that likely directly influence voters' perceptions (e.g. immigration, foreign policy, the economic situation…).

We extracted all ads seen by WTM users from Germany on or after January 1st, 2025, yielding 33,700 unique ads (accounting for a total of over 343,000 impressions in our data).

Each unique ad's text was sent to the same ministerial-8b prompt described in the Facebook section above. Note that, due to resource constraints, the ad's image or video was not stored by the extension (which collected millions of ads over the observation period), leaving the one or two sentences of the ad's text as the primary signal to judge whether an ad was political. A multimodal analysis would likely have yielded many more examples of political ads, although we see no evident reason to assume it would have significantly changed the distribution of our findings.

351 ads that received a score of 8 or greater were isolated for further human analysis, and coded to determine a- whether the ad fit Google's definition of a political advertising and b- whether the ad was about a political or social topic of relevance in the German electoral campaign context, regardless of Google's definition.

Of those, 68 unambiguously fit Google's definition of a "political campaigning ad" as they addressed the upcoming election (e.g. calling for voting for a given candidate, ads by political parties, attacks on a political opponent…). All could be found on Google's political ads library.

A further 98 were about societal issues that were prominent in the election campaign (e.g. the state of Germany's railways, support for Ukraine, taxation

policy…). While Google currently does not consider them to be political, this narrow definition might be at odds with the TTPA, which considers that messages that are “liable and designed to influence the outcome” of a vote constitute political advertisement.

It is possible that this discrepancy motivated Google to announce in November 2024 that it would withdraw from political advertising in the EU sometime in 2025, ahead of the TTPA going into effect.

Our findings therefore suggest that, while Google Ads is currently able to identify and label political campaigning ads that run on YouTube, its capacity to identify at scale broader social and political issue ads that are “liable and designed to influence the outcome” of a vote is untested.

V. Per-service summary of key results

Facebook

- **A significant portion of low-credibility accounts on Facebook are eligible for monetization**. Meta (and by extension, Facebook) decided to unsubscribe from Measures 1.1. and 1.2. of the Code of Conduct on Disinformation, which aim to address monetization by repeat spreaders of disinformation. This raises questions as to whether the service is adequately mitigating the risks to civic discourse stemming from its financial support of low-credibility actors.

- All political ads in our sample (from Germany, posted in the run-up to the February 2025 general election) posted by a politician or political party were properly labeled as 'Political'. However, **one quarter of political ads posted by other accounts were not labeled as 'political'** although they met Meta's definition of a social or political issue ad.

- One ad (out of 200) picked up by our data donation tool **did not show up in Facebook's ad library**, suggesting that some technical improvements should be made to ensure greater reliability.

Twitter/X

- **Twitter/X's ad repository** displays systematic technical issues that make it unusable for basic research activities.

- **Despite policies that disallow political ads in the EU, at least 68 such ads were found in Germany in the run-up to the February 2025 election**, suggesting that Twitter/X's enforcement of its rules is insufficient. Some ads were sponsored by political parties, which should have made them easy to identify, while others were sponsored by foreign-based actors.

YouTube & Google Ads

- **A significant portion of low-credibility channels on YouTube are actively monetized. Likewise, a significant portion of low-credibility websites are actively monetized using Google Display Ads.**

  Google decided to unsubscribe from Measures 1.1. and 1.2. of the Code of Conduct on Disinformation, which aim to address monetization by repeat spreaders of disinformation. This raises questions as to whether its services are adequately mitigating the risks to civic discourse stemming from its financial support of low-credibility actors.

- **Political campaign ads on YouTube are properly labeled as such.** As Google prepares to exit the political advertising market in the EU, its capacity to properly identify ads about broader issues that are "liable and designed to influence the outcome" of a vote, as required under the Transparency and Targeting of Political Advertising legislation, remains untested.